Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Deep Learning for Multi-Class Gastrointestinal Endoscopy: A Survey of Recent Advances and Reliability Challenges

Asaad Kadhim Abd[a]*, Zena H.Khalil[a], Ali Mohsin Aljuboori[a]

Department of Computer Science, College of Computer Science and Information Technology, University of Al-Qadisiyah.Iraq.

E-mail: asdaltmymy762@gmail.com , zena.khalil@qu.edu.iq, ali.mohsin@qu.edu.iq

ARTICLE INFO

Article history:
Received: 21 /02/2026
Rrevised form: 10 /03/2026
Accepted : 12 /03/2026
Available online: 30 /06/2026

Keywords:

Gastrointestinal endoscopy
Deep learning
Multi-class classification
Reliability

ABSTRACT

Deep learning has become a cornerstone of computer-aided diagnosis (CAD) for gastrointestinal (GI) diseases from endoscopic imagery, enabling automated recognition of complex and subtle visual patterns that often challenge clinical interpretation. While recent years have witnessed a rapid growth of deep learning-based approaches for multi-class GI disease classification, the existing literature remains highly fragmented across heterogeneous datasets, architectural choices, and evaluation protocols. This fragmentation is particularly problematic in realistic multi-class settings, where severe class imbalance, fine-grained inter-class similarity, and distributional shifts substantially undermine clinical reliability. Unlike prior surveys that primarily emphasize architectural performance or accuracy-centric comparisons, this work provides a reliability-aware analytical review of recent deep learning methods for multi-class GI disease classification from endoscopic images. We critically examine how contemporary studies address- or overlook-key reliability dimensions, including class imbalance handling, patient-level data separation, macro-level evaluation metrics, and robustness under distribution shifts. Furthermore, we identify recurring methodological limitations that may lead to hopeful performance reporting while offering limited translational value in real clinical environments. By organizing recent advances within a unified reliability-focused taxonomy, this survey highlights unresolved challenges and emerging research directions toward dependable and clinically deployable GI CAD systems. The analysis aims to support researchers and practitioners in designing evaluation pipelines and modeling strategies that move beyond optimizing accuracy toward trustworthy decision support for endoscopic diagnosis.

https://doi.org/ 10.29304/jqcsm.2026.18.22677

1. Introduction

Gastrointestinal (GI) diseases constitute a major global health burden, not only because of their prevalence but also due to diagnostic complexity and the clinical consequences of delayed detection. Endoscopy remains the clinical cornerstone for GI assessment; however, interpretation remains vulnerable to inter-observer variability, operator dependence, visual fatigue, and subtle visual overlap between clinically distinct entities-limitations that become more pronounced as diagnostic settings shift from binary detection to multi-class differentiation [1]. These constraints have sustained interest in computer-aided diagnosis (CAD) systems that can provide more objective and reproducible image-based decision support. In particular, deep learning (DL) has reshaped endoscopic image analysis by learning hierarchical representations directly from raw pixels, enabling strong performance across multiple GI tasks and motivating rapid clinical-facing experimentation [2][1]. Recently, research emphasis has increasingly shifted toward multi-class GI disease classification, aligning more closely with the heterogeneity of real clinical workflows. Research has primarily focused on the use of deep convolutional neural network (CNN) for the

∗Corresponding author: *Asaad Kadhim Abd*

Email addresses::*asdaltmymy762@gmail.com*

Communicated by ‘sub etitor’

classification of gastrointestinal (GI) diseases. Many methodologies were introduced for enhancing the classification performance of endoscopic images, such as exploring different pre-trained CNN models, using (CNN)-based spatial attention mechanism and employing CNN-transformer hybrids to complement convolutional inductive biases with global context modeling [2]. A recurring methodological limitation in contemporary GI multi-class studies is the selection and evaluation of metrics. In imbalanced multi-class settings, overall accuracy can be misleading because it may remain high even when minority classes often clinically important are systematically misclassified. Reliability, in an operational sense, therefore requires evaluation beyond accuracy, emphasizing class-sensitive summaries (e.g., macro-F1), confusion-aware measures such as MCC, and statistically sound comparison practices [4]. Beyond metric choice, generalizability remains a central barrier to deployable reliability. Models that perform well on in-distribution test sets may degrade substantially when confronted with realistic variations in acquisition devices, preparation quality, imaging artifacts, or institutional domain shifts. Evidence from GI endoscopy benchmarks/challenges highlights that robustness under distribution shift is nontrivial even for strong-performing methods, underscoring the gap between experimental success and clinical reliability [5]. In parallel, there has been renewed interest in learning paradigms that reduce dependence on exhaustive annotation. Self-supervised and weakly supervised strategies exploit large volumes of unlabeled endoscopic imagery to improve representation learning, showing measurable downstream gains and offering a pragmatic path for data-scarce institutions [3]. However, their integration into reliability-oriented evaluation pipelines (calibration, shift stress tests, and clinically meaningful error analysis) is still inconsistently addressed across the literature. Although multiple reviews have surveyed DL applications in GI endoscopy, recent review-style papers often emphasize architectures and task performance, with comparatively limited structured synthesis of reliability-aware validation in multi-class diagnostic settings-particularly regarding metric adequacy, robustness under distribution shift, and evidence quality across datasets. Accordingly, this survey reviews deep learning-based approaches for multi-class gastrointestinal disease classification from endoscopic imagery published between 2023 and 2026, and analyzes them along three complementary dimensions: (I) architectural design (CNNs, transformers, and hybrid/ensemble variants); (ii) data handling, with particular focus on imbalance and dataset realism; and (iii) evaluation methodology, emphasizing reliability-oriented validation and generalizability. By synthesizing trends and recurring methodological weaknesses, the survey aims to clarify what current "performance" claims do and do not imply about clinically meaningful reliability. The remainder of this paper is organized as illustrated in Figure 1(a). Section 2 details the methodology adopted for the literature search and study selection. Section 3 provides an overview of gastrointestinal endoscopy modalities and examines the characteristics of widely used multi-class datasets. Section 4 presents a comprehensive review of deep learning techniques, categorizing them into CNN-based, Transformer-based, hybrid, and ensemble architectures. Section 5 critiques current evaluation practices and highlight's reliability gaps. Section 6 discusses strategies for data augmentation and handling class imbalance. Section 7 provides a discussion of the findings and suggests directions for future research. Finally, Section 8 presents the concluding remarks.

2. Methodology

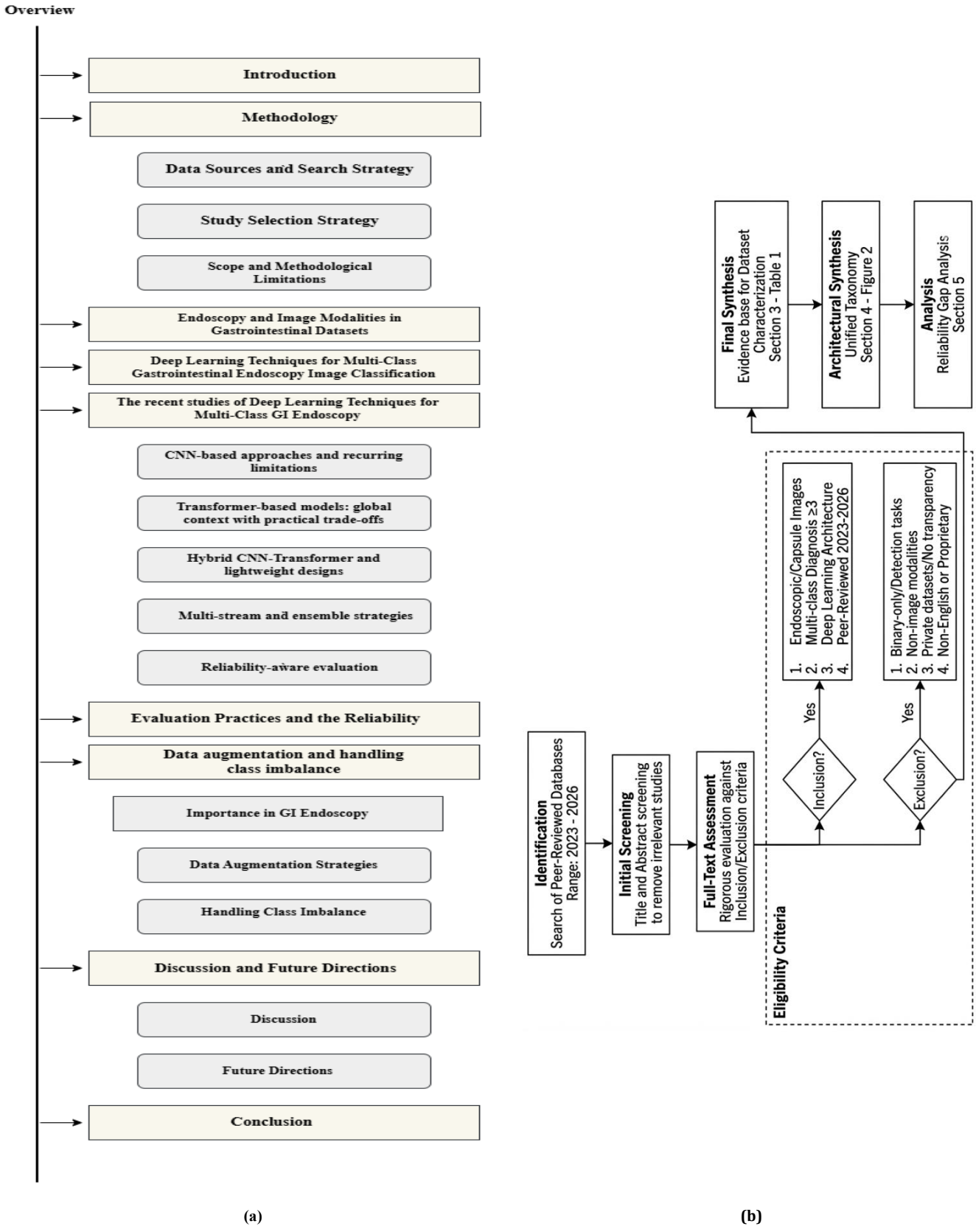
To ensure transparency, reproducibility, and methodological rigor, this survey adopts a structured study selection strategy consistent with contemporary reporting and quality-assessment guidance for medical AI and diagnostic model research, as represented in Figure 1(b).

2.1. Data Sources and Search Strategy

A comprehensive search was conducted in PubMed, IEEE Xplore, Scopus, web of science, and Google Scholar to cover both clinical and engineering-oriented literature. The search window was restricted to January 2023-early 2026 to reflect recent architectural and evaluation shifts in GI endoscopic AI. Queries combined domain and method terms using Boolean operators, including: "gastrointestinal endoscopy" AND "deep learning" AND ("multi-class" OR "multiclass") AND (CNN OR "Vision Transformer" OR ensemble). Additionally, the reference lists of recent endoscopic imaging review papers were screened to mitigate the risk of omission.

2.2. *Study Selection Strategy*

Studies were selected based on strict eligibility criteria to ensure relevance and quality, consist of:



(a) (b)

Fig. 1 - a) the organization of the survey, b) PRISMA flow diagram of the methodology of survey.

2.2.1. *Inclusion Criteria*

Studies were included if they: (1) used endoscopic or capsule endoscopy images for GI disease classification; (2) performed multi-class diagnosis (≥3 classes), reflecting more realistic clinical differentiation; (3) adopted deep learning (CNNs, transformers, hybrid, or ensembles) with a reportable experimental design; and (4) were peer-reviewed and published between 2023 and 2026.

2.2.2. Exclusion Criteria

Studies were excluded if they: (1) targeted binary-only tasks or focused exclusively on detection/segmentation without multi-class diagnostic intent; (2) used non-image modalities as primary input; (3) used private datasets without adequate transparency, limiting assessment of bias/applicability; or (4) were not peer-reviewed or fell outside the specified time window.

2.2.3. Screening and Selection Procedure

Records were screened in a two-stage process: (i) initial title and abstract screening to remove irrelevant studies, followed by (ii) a rigorous full-text assessment for eligibility against the inclusion and exclusion criteria defined above. The extracted set constitutes the evidence base for the dataset characterization in Section 3 (Table 1) and the comprehensive architectural synthesis in Section 4. Based on this selection, the literature was organized into a unified taxonomy (Figure 2) to facilitate the comparative analysis of modeling choices in Section 4 and the critical reliability gap analysis in Section 5

2.2.4. Scope and Methodological Limitations

This review is limited to English-language, peer-reviewed publications and to studies reporting publicly accessible experimental results. Such constraints may exclude proprietary clinical systems; however, they improve auditability and reproducibility.

3. Endoscopy and Image Modalities in Gastrointestinal Datasets

In the field of medical imaging, multi-class classification of gastrointestinal (GI) tract findings has been significantly propelled by the release of several large-scale, high-quality benchmark datasets. These datasets primarily utilize endoscopy images, often stored in the RGB color space to preserve vital clinical features such as tissue vascularization, mucosal texture, and pigmentation, which are critical for identifying lesions like polyps or inflammation and also to differentiate between anatomical landmarks and various disease types simultaneously. Unlike specialized modalities such as chromoendoscopy, which are typically used for binary lesion vs. no lesion tasks. Endoscopy is a minimally invasive clinical procedure in which a flexible optical instrument (endoscope) equipped with a camera and illumination source is inserted into the gastrointestinal (GI) tract to visualize internal mucosal surfaces for diagnostic and therapeutic purposes. Endoscopic imaging captures color, high-resolution frames of anatomical regions such esophagus, stomach, colon; and pathological finding such as polyps, ulcers; enabling clinicians to inspect tissue morphology, detect aberrations, and guide interventions. Such images form the basis for computer-aided diagnosis (CAD) systems in GI disease classification through machine learning and deep learning models. Endoscopic modalities include standard white-light endoscopy (WLE) used in routine gastroscopy and colonoscopy, and wireless capsule endoscopy (WCE) where a swallowable camera captures images throughout the small bowel, generating large sequences of frames for analysis. These modalities produce multi-class datasets where each image is associated with a clinical or anatomical label used for supervised classification tasks [4] [5] . These datasets collectively support supervised deep learning models for automatic GI disease classification, enabling comparative evaluation and algorithmic advancement in medical imaging. The Kvasir datasets (v1 and v2) were one of the first publicly available benchmarks for multi-class GI classification. It comprises eight distinct classes representing three key categories: anatomical landmarks (cecum, pylorus, z-line), pathological findings (esophagitis, polyps, ulcerative colitis), and endoscopic procedures (dyed and lifted polyps, resection margins) [4]. The images are captured using standard endoscopic equipment with varying resolutions, emphasizing the need for models that are robust to equipment-specific artifacts and the need higher quality images. Consequently, an evolution of the original Kvasir, Hyper-Kvasir introduced as more comprehensive dataset for GI endoscopy. It contains 110,079 images in total, of which 10,662 are manually labeled into 23 different classes. This dataset is particularly challenging for multi-class classification due to its high class imbalance, reflecting real-world clinical frequency where normal findings significantly outnumber rare pathologies like Barrett's esophagus or specific

grades of ulcerative colitis [5]. Furthermore, Kvasir-Capsule is released as a video dataset, but it is widely used for frame-level image classification by extracting and labeling individual frames from capsule endoscopy videos. It primarily covering the small bowel and not focused on the esophagus, stomach, or colon, though occasional frames from these regions may appear[6]. The more a recent dataset is GastroVision, which is multi-center dataset designed to address the limitations of single-center data. It comprises 8,000 labeled endoscopic images acquired from upper and lower GI tract regions, covering significant anatomical landmarks, pathological abnormalities, and normal findings. Its primary value lies in its diversity, as the data is sourced from multiple hospitals (e.g., Bærum Hospital in Norway and Karolinska University Hospital in Sweden), making it an ideal candidate for evaluating the generalization capabilities of deep learning models across different clinical environments [1]. In addition to widely used datasets like Kvasir v2, HyperKvasir, Kvasir-Capsule, and GastroVision, newer datasets such as GastroEndoNet and GastroHUN expand multi-class classification opportunities. GastroEndoNet provides four clinically relevant classes for gastroesophageal reflux and polyps, while GastroHUN offers multiple stomach anatomical landmark classes. Comprehensive collections like the ERS dataset include a broad range of MST-standard findings that can support rich multi-class learning. Across commonly used gastrointestinal image datasets, the predominant image modalities employed for multi-class classification include: (I) upper gastrointestinal endoscopy (gastroscopy) images depicting normal and pathological appearances of the esophagus, stomach, and duodenum; (ii) lower gastrointestinal endoscopy (colonoscopy) images covering the colon and rectum with lesions such as polyps and inflammatory changes; (iii) therapeutic and procedural views, such as polypectomy, dye spraying, and resection margins, which are present in large-scale datasets but are not always included in classification tasks; and (iv) wireless capsule endoscopy frames capturing the entire small bowel, introducing significant temporal and spatial variability [4] [1]. Below is a summary table of relevant widely used, benchmark and publicly available GI image datasets that can use for multi-class GI learning and classification represented their image type and source modality table 1..

Table 1 - Widely used multi-class gastrointestinal image datasets

Dataset Name	Image Type	Source Modality	Number of Classes	Reference
Kvasir v1	Upper + Lower GI	Endoscopy	8	[4]
Kvasir v2	Upper + Lower GI	Endoscopy	8	[4]
HyperKvasir	Upper + Lower GI + procedural/interventional frames	Endoscopy	23	[5]
GastroVision	Upper + Lower GI	Endoscopy	27	[6]
Kvasir-Capsule	Small bowel imaging	Wireless Capsule Endoscopy	Multiple (subset-dependent)	[1]

4. Deep Learning Techniques for Multi-Class Gastrointestinal Endoscopy Image Classification

The application of deep learning to multi-class gastrointestinal (GI) image classification has progressed through distinct periods, reflecting both technological advances and the availability of annotated datasets. Before 2015, studies were largely exploratory, with shallow CNNs applied to small, often private datasets; multi-class classification was limited, and models were demonstrating feasibility rather than robust performance. Between 2015 and 2019, the release of publicly available datasets such as Kvasir (v1 and v2) enabled systematic investigation of CNN-based approaches for multi-class classification, with architectures like VGG, ResNet, and Inception used alongside data augmentation and transfer learning to improve generalization [4]. From 2020 to 2022, research expanded to larger and more diverse datasets, including HyperKvasir and Kvasir-Capsule, with deeper CNNs, ensemble learning, and early attention mechanisms facilitating classification across multiple anatomical landmarks, pathological findings, and procedural frames [5] [6]. The period 2023-2026 marks a

pronounced surge in multi-class GI deep learning research, driven by datasets such as GastroVision with 27 classes and by the adoption of modern architectures including ,Vision Transformer, hybrid CNN-transformer models, and explainable AI methods, which together improve classification accuracy, interpretability, and robustness for clinically relevant multi-class tasks [1]. Still, it has exposed persistent weaknesses in dataset realism, class-imbalance handling, and evaluation design. In particular, many studies still report improvements under inconsistent splits and metrics, limiting cross-paper comparability and making "state-of-the-art" claims fragile [7]. This survey primarily concentrates on studies published between 2023 and 2026, as this period represents a phase of rapid methodological advancement in deep learning for multi-class gastrointestinal endoscopy image classification.

5. Recent Studies of Deep Learning Techniques for Multi-Class GI Endoscopy

During recent years (2023-2026), research has shifted from conventional CNN-based pipelines toward transformer-based, hybrid CNN-Transformer, ensemble, and reliability-aware architectures, driven by increased dataset scale, computational efficiency, and clinical deployment requirements as represented in Figure2. This taxonomy summarizes representative deep learning paradigms explored for multi-class GI endoscopy classification during 2023-2026, categorized by architectural paradigm and reliability focus. It is motivating that architecture choice alone rarely addresses real-world reliability constraints, which is an evaluation property as much as a modeling choice, requiring metrics and stress tests aligned with clinical risk.

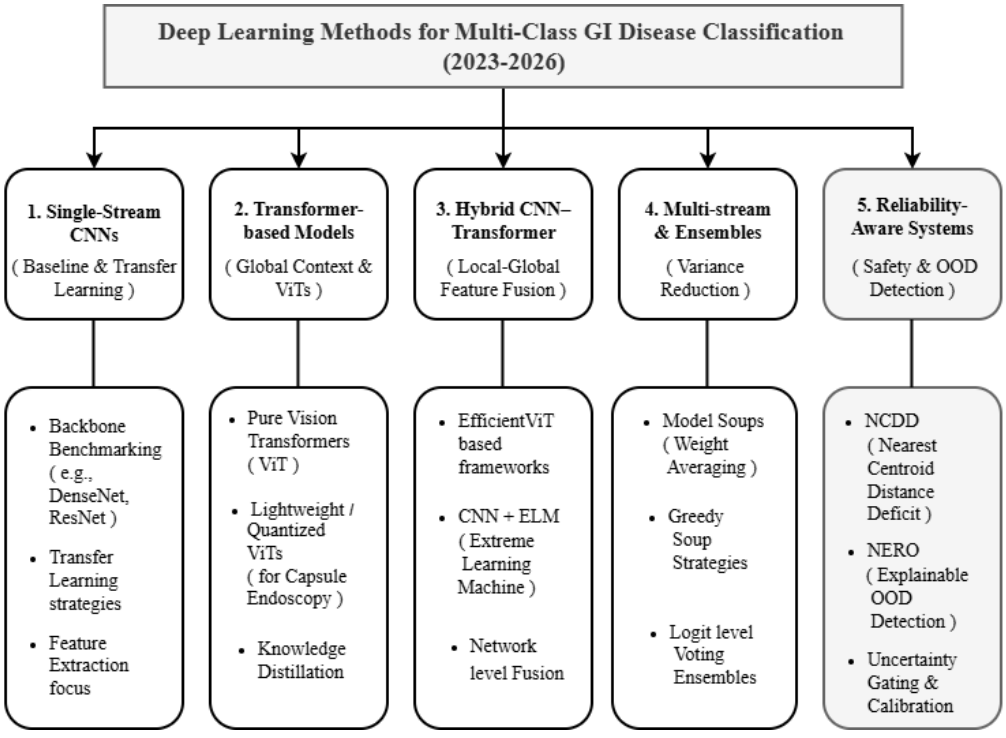


Fig. 2 - Learning Paradigms for Multi-Class Gastrointestinal Endoscopy Image Classifica-tion (2023-2026).

5.1. CNN-Based Approaches and Recurring Limitations

CNNs remain the dominant backbone for GI image classification due to strong local feature extraction and efficient inference. In this period, researchers explored both standard pretrained CNN backbones and customized CNN architectures to address the challenges of diverse anatomical features and pathological classes. These models typically employ pretrained CNN backbones such as ResNet, DenseNet, EfficientNet, and MobileNet, fine-tuned on large public datasets including KvasirV2, HyperKvasir, and GastroVision. Several recent studies have explored CNN-centric methods on Kvasir and Kvasir-2 for multi-class GI classification. Üzen and Fırat (2024) proposed a DenseNet201-based feature integration model on KvasirV2, combining deep layer information to strengthen class

separation [7]. Transfer learning frameworks that fine-tune common CNN backbones (e.g., Xception, InceptionResNetV2, VGG16) on Kvasir and HyperKvasir further confirm the utility of CNNs as baseline models in multi-class GI tasks [8]. Rubab et al. (2026) introduced lightweight CNN variants such as Parallel Depthwise Separable CNN (PD-CNN) and Parallel Squeeze-and-Excitation CNN (PSE-CNN), demonstrating efficient extraction of discriminative features across all 27 GastroVision classes while maintaining reduced computational overhead [9]. Malik et al. (2024) evaluated classical CNN architectures (e.g., VGG-19, ResNet152V2) for multi-class detection of ulcerative colitis, polyps, and dyed-lifted polyps in wireless capsule endoscopy frames, further establishing the efficacy of baseline CNNs in capsule modalities [10]. However, despite these advances, several recurring limitations are consistently reported. First, standard CNNs primarily capture local spatial patterns and often struggle to model long-range contextual relationships between anatomical regions, which are crucial for distinguishing visually similar GI conditions. Second, CNN-based classifiers exhibit sensitivity to domain shift, including variations in endoscope type, illumination, and acquisition protocol, leading to reduced generalization across datasets. Third, performance degradation is frequently observed for minority or visually ambiguous classes, even when class-reweighting or oversampling techniques are applied. Finally, most CNN-based approaches rely on deterministic predictions and lack built-in mechanisms for uncertainty estimation.

5.2. Transformer-Based Models: Global Context with Practical Trade-Offs

Transformer Vision backbones (ViT) and hierarchical variants (e.g., Swin) were introduced to capture longer-range dependencies and global context that CNNs may miss in fine-grained recognition. These models replace or augment convolutional layers with self-attention mechanisms, enabling feature interactions across entire images [11]. In GI classification, comparative studies report that ViT-style models can be competitive, especially with transfer learning, though outcomes remain sensitive to dataset scale, augmentation, and training stability [2]. Authors of [12] proposed a novel Vision Transformer model based on hybrid shifted windows for digestive tract image classification, which can obtain both short-range and long-range dependency concurrently, overcoming the inability of the Swin Transformer cannot capture the long-range dependency well in complex gastrointestinal endoscopy images. Experiments demonstrate the superiority on the Kvasir v2 dataset and HyperKvasir dataset. While pure transformer models have revolutionized the field, they have simultaneously introduced critical bottlenecks that became critical in the clinical environment. This is attributed to the inherent lack of the inductive biases inherent in CNNs, specifically translation invariance and locality. In GI endoscopy, datasets (like HyperKvasir) are relatively small compared to natural image datasets (ImageNet). Also, transformers are masters of the global context, but they often struggle with the "fine-grained" details required for multi-class classification. Therefore, researchers moved to hybrid designs where a CNN stem (e.g., ResNet or DenseNet) acts as a feature extractor first. This learns the model local textures (like the fine vascular patterns of Barrett's esophagus) before the transformer layer analyzes the global relationship between those features. CNN also preserve high-resolution spatial maps that ensures the specific texture of a lesion, for example, do not become blurred or deformed.

5.3. Hybrid CNN-Transformer and Lightweight Designs

Hybrid architectures attempt to combine CNN locality priors with transformer context modeling, while lightweight variants target deployment constraints (latency, memory). Recent work includes EfficientViT-based hybrid frameworks for GI disease classification, reflecting a trend toward architectures that balance accuracy and efficiency [13] .For example, Tabassum et al. (2026) employ lightweight Vision Transformer ensembles and XAI techniques like Grad-CAM for high-class granularity ,and Tang et al. (2023) and Wu et al. (2023) focus on multi-task learning and local feature attention to optimize the trade-off between computational efficiency and segmentation accuracy. By leveraging self-attention mechanisms, these models achieve superior performance in identifying subtle pathological variations across large-scale datasets like HyperKvasir, facilitating more robust real-time clinical decision support [14][16]][16]. Nevertheless, many hybrid studies still primarily evaluate in-distribution test splits; without robustness and calibration analyses, reported gains can overstate deploy ability. Especially that, CNNs produce dense, local feature maps, while transformers produce sparse, global token embeddings. This concatenation often results in a feature mismatch due to the disparity between dense CNN maps and sparse Transformer tokens.

5.4. Multi-stream and ensemble strategies

Ensemble is widely used to reduce variance and exploit complementary error patterns across backbones; survey evidence suggests ensembles are often most effective when constituent models are diverse and when aggregation is tuned to the target failure modes [13]. During 2023-2026, this paradigm has been increasingly adopted to mitigate the limitations of single-model approaches and to enhance discriminatory performance across complex class sets. Aslan (2026) proposed a stacking ensemble framework that combines outputs from three pretrained CNN backbones (ResNet50, DenseNet201, MobileNetV3Large) with classical meta-classifiers to classify eight GI endoscopy classes in the KvasirV2 dataset, achieving higher accuracy compared to individual CNNs and demonstrating that diverse feature extractors can reduce classification variance and improve performance consistency [17]. Similarly, Gunasekaran et al. (2023) introduced GIT-NET, a weighted average ensemble of DenseNet201, InceptionV3, and ResNet50, which outperformed model averaging and individual base learners on the KvasirV2 dataset, illustrating the benefit of weighted fusion for class complementarity and overall multi-class accuracy [18] . Diagnostic studies have further shown that stacking ensemble models with multiple CNN predictors trained via cross-validation significantly surpass single predictors on both KvasirV2 and HyperKvasir datasets, with performance gains supported by McNamar's statistical comparison, highlighting the statistical reliability of ensemble gains over individual architectures [19]. Other work has explored weighted and dynamic ensemble frameworks, such as combinations of DenseNet201, InceptionV3, and VGG19 with optimized weights to balance contributions from each model, yielding improved accuracy and demonstrating that careful weighting schemes can effectively extract complementary class-specific information [20]. The weight-space averaging (e.g., Model Soups) offers a computationally attractive alternative that can improve performance and, in some cases, robustness without the inference-time overhead of ensembles[21]. However, ensemble gains should be interpreted cautiously, improvements may reflect better optimization rather than better clinical reliability unless paired with reliability aware validation.

5.5. Reliability-aware evaluation

A central weakness across GI multi-class literature is the continued reliance on accuracy-centric reporting. In imbalanced multi-class settings, accuracy alone is insufficient; more informative practice includes macro-F1, MCC, and statistically grounded comparisons [22]. Uncertainty-awareness represent the ability of model to quantify how "sure" or "unsure" a specific prediction. In medical imaging, this is important because of a situation with a blurry image or a rare pathology which hasn't seen before. Reliability also requires uncertainty and calibration assessment, because overconfident wrong predictions pose disproportionate clinical risk; recent medical imaging work underscores that modern networks can be mis calibrated and that calibration should be measured explicitly rather than assumed [23]. Equally important is generalizability under realistic conditions. Evidence from GI endoscopy challenges demonstrates that strong in-distribution performance does not guarantee robustness in dynamic clinical colonoscopy scenarios, reinforcing the need for distribution-shift stress testing as a first-class evaluation component [24]. Finally, learning with fewer labels is increasingly explored via self-supervised/curriculum strategies that leverage unlabeled GI images. At the same time, these methods can improve representation quality; their contribution to reliability should still be validated through calibration and shift testing rather than relying on single-split accuracy gains [3]. In the context of GI and capsule endoscopy, recent work has directly targeted the OOD problem. The EndoOOD framework was proposed to improve the reliability of wireless capsule endoscopy diagnosis by integrating uncertainty-aware mixup training, long-tailed in-distribution calibration, and virtual-logit matching to distinguish in-distribution from OOD inputs, addressing challenges posed by anatomical variations and unseen categories in capsule imagery; evaluations with multiple state-of-the-art comparators demonstrated the framework's potential to enhance robustness in diagnostic classification tasks [28]. Complementary to learning-based uncertainty methods, feature-space distance methods such as Nearest Centroid Distance Deficit (NCDD) have been developed specifically for gastrointestinal vision, leveraging the intuition that OOD samples will maintain greater feature distance from known class centroids, thereby enabling reliable identification of novel conditions while preserving performance on in-distribution classes; experimental results across benchmarks like KvasirV2 and GastroVision show such approaches can effectively flag OOD instances relative to baseline methods [29]. Also, a notable contribution in this direction is NERO (Neuron-level Relevance for OOD detection) by Chhetri et al. (2026), which introduces an explainable OOD detection framework tailored to GI imaging. Instead of relying solely on output confidence scores, NERO analyzes neuron-level relevance patterns within deep networks to identify samples deviating from the training distribution, enabling both improved OOD detection and interpretability [30]. Collectively, these studies indicate a methodological shift from accuracy-centric evaluation toward integrated

reliability assessment, where calibration metrics, uncertainty estimation, and OOD detection are treated as first-class components of multi-class GI endoscopy systems to support clinically meaningful claims.

6. Evaluation Practices and the Reliability

Reliability refers to the degree to which a model's predictions are consistent, trustworthy, and clinically safe across different classes, patients, and acquisition conditions. Across architectural families (CNNs, ViTs, and hybrids), the evaluation methodology remains a persistent weakness. Accuracy-centric reporting remains common despite being structurally insufficient for imbalanced multi-class settings, where high accuracy can coexist with clinically consequential minority-class failures and systematic confusions between visually similar categories. To visualize the structural weakness in these prevailing approaches, Figure 3 illustrates how the exclusion of reliability controls directly translates intrinsic data challenges into clinical risks, highlighting the 'missing layer' required for safe deployment.

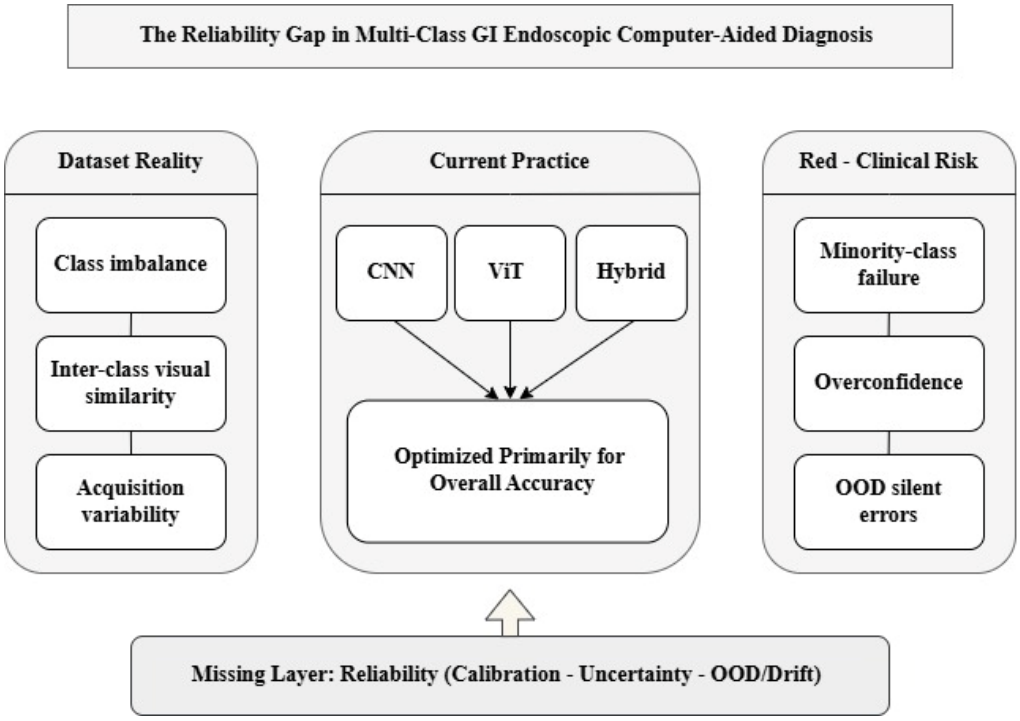


Fig. 3 - The reliability gap in multi-class GI endoscopic computer-aided diagnosis

The Figure showed the contrasts the current accuracy-centric practice (center) with the clinical reality. Intrinsic data challenges such as class imbalance (left) are often propagated by standard models into clinical risks like overconfidence and silent failures (right). The bottom bar highlights the critical "missing layer" of calibration and uncertainty estimation needed to bridge this gap. In multi-class gastrointestinal endoscopic image classification, model reliability is mainly characterized by three complementary components included prediction calibration, uncertainty estimation, and out-of-distribution (OOD) and data drift detection. Calibration ensures alignment between predicted confidence and true correctness, uncertainty estimation enables the model to recognize ambiguous or low-confidence cases, and OOD/drift detection allows the system to identify inputs that deviate from the training distribution. Despite substantial architectural progress in deep learning–based gastrointestinal (GI) disease classification, recent studies continue to report gaps that constrain clinical reliability and real-world applicability. These gaps are especially pronounced in multi-class diagnostic settings, where class imbalance, inter-class visual similarity, and acquisition variability frequently dominate practical scenarios [25]. Overall, studies from 2023-2026 (Table 2) show clear architectural progress in CNN, ViT, and hybrids techniques in addition to increasing attention to data efficiency. However, the literature remains fragmented across datasets, splits, and reporting

practices. The most consequential gap is that "reliability" is often inferred from improvements in accuracy rather than demonstrated using imbalance-aware metrics, statistically sound comparisons, and explicit tests of generalization and robustness [22]. This gap motivates a structured synthesis that prioritizes reliability aware evaluation for multi-class GI classification rather than accuracy alone. Recent works have highlighted out-of-distribution (OOD) detection as a key mechanism for improving the reliability of gastrointestinal imaging models. Tan et al. [28] enhance reliability by incorporating uncertainty-aware predictions, allowing the system to flag low-confidence capsule endoscopy cases that may not conform to training distributions. Pokharel et al. [29] further address reliability through a distance-based formulation, where deviations from class centroids signal unreliable predictions on unseen GI data. More recently, Chhetri et al. [30] extend reliability beyond detection by introducing explainable OOD decisions at the neuron level, enabling interpretation of why a sample is considered out-of-distribution. Collectively, these approaches demonstrate that reliable GI vision systems must not only achieve high classification accuracy but also identify, explain, and appropriately handle unfamiliar or ambiguous inputs.

The evaluating of OOD and model calibration were founded in only 5 studies as 29.4%. The works of [30], [37], and [39] included OOD detection, while the Calibration/Uncertainty supported only in [26] and [32]. However, Cross-dataset validation which mean testing the proposed model on a different dataset than the training one was utilizing in [9] and [31] as 11.8%.

Despite the fact that standard Accuracy is a common benchmark metric, but some studies shift toward more rigorous statistical validation in recent 2025–2026 publications. For example, a comprehensive analysis of the literature expose an emphasis on multi-dimensional evaluation, where 68.75% of the studies reporting the macro-F1 score to asses class imbalancing in gastrointestinal datasets. Other studies as 37.5% incorporated the Matthews Correlation Coefficient (MCC) to provide a more reliable measure. Notably, both F1 and MCC was observed in 25% of the papers (e.g., [26], [28], [33], [40]), signaling to develop in transparency of deep learning-based endoscopic analysis, as reported in Table 2.

Table 2 - Comparative Analysis of Methodologies and Reliability Validation Strategies in Recent Multi-Class GI Disease Classification Studies (2023-2026).

Ref. (Year)	*Methodology*	*Dataset (Classes)*	*Best Result*	*Reliability Strategy*
[26]. (2026)	BUCAN: Bayesian Network + SwirlAttention	Kvasir v2 (8)	Acc: 0.8746% Prec:0.8776% Rec:0.8746% F1:0.8743% MCC:0.8572%	Uncertainty Quantification
[27]. (2026)	Context-CNN: Feature fusion ResNet.	GastroVision (27)	Acc: 98.80%	-
[28]. (2025)	EndNote: Multiscale DL + Feature Selection	HyperKvasir (23)	Acc: 98.40% Sen: 0.980% Spe: 0.998% Prec: 0.981% F1: 0.980% MCC: 0.978%	Feature Optimization
[29]. (2025)	XAI-Inception: InceptionV3 + Boosting + Grad-CAM	HyperKvasir (23)	F1: 0.90% Rec: 0.81%	Explainable AI (XAI)
[9]. (2025)	Deep Info Fusion: ResNet + Optimization (NRMPO)	HyperKvasir (23)	Acc: 84.51% Prec: 84.14% Sen: 82.41% F1: 81.74%	Robustness across datasets
[30]. (2025)	NERO: Neuron-level relevance clustering for	GastroVision (27)	AUC: 82.03% FPR: 76.74%	OOD Detection

Ref. (Year)	*Methodology*	*Dataset (Classes)*	*Best Result*	*Reliability Strategy*
	anomaly detection.			
[31]. (2025)	Multi-Task CViT: Convolutional Vision Transformer for classification & severity.	4 GT Datasets	Acc: 0.965% Prec: 0.960% Rec: 0.952% F1: 0.958%	-
[32]. (2025)	MEGAN: Mixture of Experts + Evidential Deep Learning.	UC-Score (4)	ECE: 0.107% F1: 0.680%	Calibration & Uncertainty
[33]. (2025)	XAI-CNN: Model Soups + Grad-CAM analysis.	GastroVision (27)	Prec: 0.832% Rec: 0.830% F1: 0.830% MCC: 0.809%	-
[34]. (2025)	Deep Ensemble: EfficientNet + NasNet voting.	Kvasir v2 (8)	Acc: 0.978% Prec: 0.98% Rec: 0.97% F1: 0.96% AUC: 0.97%	-
[35]. (2025)	DivGI: Decoupled training + Mixup.	GastroVision (27)	MCC: 82.88%	-
[13]. (2025)	EfficientViT: Hybrid CNN-Transformer architecture.	Private GI (8)	Acc: 99.82% Prec: 99.62% Rec: 99.50%	-
[36]. (2025)	Hybrid Seg-Class: Deep segmentation + classification.	HyperKvasir (23)	Acc: 95.20%	-
[37]. (2024)	NCDD: Nearest Centroid Distance Deficit.	GastroVision (27)	AUC: 85.37% FPR: 52.48%	OOD Detection
[38]. (2024)	Ensemble ELM: CNN + Extreme Learning Machine + SHAP.	GastroVision (27)	Acc: 87.75%	-
[39]. (2023)	OOD-GI: Open-set recognition framework.	Kvasir-Capsule	AUC: 0.91	OOD Detection
[40]. (2023)	Baseline: ResNet-50 / DenseNet-121 benchmarks.	GastroVision (27)	MCC: 80.62% Acc: 93.46%	-

7. Data Augmentation and Handling Class Imbalance

The reliability of deep learning models in gastrointestinal (GI) endoscopy is frequently undermined by the intrinsic properties of medical datasets, particularly severe class imbalance and acquisition variability [26,28]. In realistic multi-class settings, normal findings or common pathologies significantly outnumber rare conditions, leading to models that optimize global accuracy at the expense of minority-class sensitivity [29]. To address these challenges, recent studies have moved beyond standard geometric transformations toward reliability-aware augmentation and algorithmic handling of imbalance. Standard training paradigms on imbalanced GI datasets often result in high global accuracy but poor performance on clinically critical yet rare lesions [2] [41]. This bias is exacerbated by intra-class variability caused by differing lighting conditions, camera angles, and device manufacturers [24]. Consequently, robust augmentation and imbalance handling are not merely optimization tricks but essential prerequisites for clinical safety and generalizability. The literature reveals a three-tiered taxonomy for managing class imbalance in gastrointestinal (GI) endoscopy classification: data-level, feature-level, and algorithm-level strategies. Data-level interventions remain foundational for datasets with high class-density, such as GastroVision, which contains 27 visually similar categories. In this context, researchers utilize offline geometric transformations including rotation, scaling, and color jittering to physically equalize the training distribution before model initialization [1]. Furthermore, the field has transitioned toward domain-specific transformations to enhance feature distinctiveness; for instance, Huang et al. proposed a software-based transformation converting standard White Light Endoscopy (WLE) images into hyperspectral-like representations to improve the visibility of vascular patterns [35]. Advanced information fusion and optimization algorithms have also been integrated at this stage to expand training diversity and regularize models against noise and artifacts [9,13, 40]. Feature-level optimization has emerged as a dominant trend (2024-2026), shifting focus from raw image manipulation to the latent space. Researchers utilize methods such as mRMR (Minimum Redundancy Maximum Relevance) [36]. to ensure that feature vectors are not biased toward majority class signatures. Notably, Ali et al. demonstrated the efficacy of a two-phase transfer learning framework, which stabilizes feature extraction on general datasets before fine-tuning specifically on imbalanced medical classes [41]. Additionally, feature refinement modules such as MECNET, proposed by Kumar et al., have been developed to enhance class separability in the feature space by emphasizing discriminative regions [37]. Finally, algorithm-level interventions modify the learning process itself through cost-sensitive learning and weighted loss functions [34]. These methods allow the model to penalize misclassifications of rare, high-stakes pathologies more heavily, thereby prioritizing clinical sensitivity Fahad et al. [33] introduced of "Model Soups," which averages the weights of diverse fine-tuned models to improve performance on the "long-tail" distribution of GastroVision without increasing inference costs. Comparative studies by Shafi et al. highlight that architectural choice plays a critical role, as specific CNN backbones exhibit varying degrees of innate resilience to minority classes [29]. Consequently, ensemble frameworks are widely adopted as a robust strategy to mitigate imbalance effects without altering the underlying data distribution, ensuring that misclassifications made by one model are compensated by the collective decision of the ensemble [19,15,20].

Table 3- Evolution of Data Augmentation and Imbalance Handling Strategies in Multi-Class GI Datasets (2023-2026).

Ref. (Year)	Dataset	Handling Strategy	Implementation Mechanism
[27]. (2026)	GastroVision	Context-Aware	Feature fusion block to maximize extraction from minority classes.
[33]. (2025)	GastroVision	Generative AI	GANs for synthetic sample generation + Model Soups ensemble.
[42]. (2025)	HyperKvasir	Deep Ensemble	Multi-Model Voting: Aggregating predictions from disparate architectures (EfficientNet + NasNet) to mitigate bias towards dominant classes.
[34].	HyperKvasir	Curriculum Learning	Quality-Aware Scoring: Training on "easy" samples first, then gradually introducing "hard" samples to prevent

(2025)			overfitting on majority classes.
[43]. (2025)	HyperKvasir	Semi-Supervised	Label Propagation: Leveraging unlabeled data (pseudo-labeling) to reinforce decision boundaries for rare/minority classes.
[30]. (2025)	GastroVision	OOD Filtering	NERO: Neuron-level clustering to treat imbalance as anomaly detection.
[35]. (2025)	GastroVision	Decoupled Training	Two-stage training: Instance sampling (Features) right arrow Class sampling (Classifier).
[32]. (2025)	UC-Score	Uncertainty Weighting	Mixture of Experts weighted by Evidential Uncertainty scores.
[36]. (2025)	HyperKvasir	Hard Attention	Segmentation-guided U-Net to isolate ROI and remove background noise.
[38]. (2024)	GastroVision	Dim. Reduction	Depth wise CNNs + PCA to prevent overfitting on majority classes.
[39]. (2023)	HyperKvasir	Cost-Sensitive	Self-supervised Barlow Twins + Focal Loss for hard samples.
[42]. (2023)	GastroVision	Geometric Aug.	Baseline standard augmentation (Flip, Rotate, Resize).

8. Discussion and Future Directions

This section synthesizes the findings of the reviewed studies and critically examines emerging trends, methodological strengths, and persistent challenges in deep learning-based multi-class gastrointestinal endoscopy image classification. Building on the comparative analysis of CNN-based, transformer-based, hybrid, ensemble, and reliability-aware paradigms, the discussion highlights key methodological limitations that continue to impede robust clinical deployment.

8.1. *Discussion*

The literature published between 2023 and 2026 demonstrates clear architectural maturation in gastrointestinal endoscopy image classification, marked by a transition from narrow binary detection tasks toward more challenging multi-class recognition problems. Despite this progress, a dominant trend remains architecture-centric optimization, while reliability-oriented validation-encompassing class imbalance sensitivity, calibration quality, and robustness under distribution shift-is frequently treated as secondary. This imbalance between modeling ambition and evaluation rigor constitutes a central barrier to meaningful clinical translation. Convolutional neural network (CNN) backbones continue to be widely adopted due to their computational efficiency and strong local feature extraction capabilities. However, their limitations become increasingly apparent in realistic GI endoscopy settings characterized by fine-grained inter-class similarity and severe class imbalance. Under such conditions, global accuracy metrics may remain deceptively high, while clinically significant errors-particularly those affecting minority or visually overlapping classes-are insufficiently analyzed or underreported. Vision Transformers (ViTs) and hybrid CNN-Transformer architectures represent a genuine methodological shift toward modeling long-range dependencies and global contextual information. Nevertheless, comparative evidence indicates that performance gains from transformer-based designs are not uniform across datasets and are highly sensitive to factors such as dataset scale, pretraining strategy, and training protocol. As a result, claims of transformer superiority cannot be generalized and must be substantiated through carefully controlled experimental designs rather than isolated

benchmark improvements. A recurring weakness across the surveyed studies is the persistent evaluation-reliability gap. While model complexity and architectural novelty continue to increase, reported evaluation practices remain largely accuracy-centric. In imbalanced multi-class classification scenarios, accuracy is well known to be potentially misleading, underscoring the necessity of confusion-aware and class-sensitive metrics such as macro-averaged F1-score and Matthew's correlation coefficient (MCC), accompanied by statistically defensible comparisons. Furthermore, existing evidence from GI endoscopy tasks consistently demonstrates that performance can degrade substantially under realistic clinical variability, including differences in imaging devices, acquisition protocols, and clinical centers, reinforcing that strong in-distribution results alone are insufficient indicators of real-world deployability.

8.2. Future Directions

Based on the synthesized evidence, several research directions emerge as critical for advancing reliable multi-class GI endoscopy image classification, included:

8.2.1. *Reliability-oriented evaluation*

Future studies should systematically report macro-averaged and class-sensitive metrics (e.g., macro-F1, MCC) and employ appropriate statistical testing to substantiate performance claims, rather than relying on single-run or accuracy-only evaluations.

8.2.2. *Explicit assessment of generalization*

Cross-dataset, multi-center, and device-level validation protocols should become standard practice. Evidence from GI endoscopy consistently shows that model performance may degrade under changing acquisition conditions, emphasizing the need for evaluation beyond single-dataset benchmarks.

8.2.3. *Uncertainty modeling and safe abstention*

Reliable clinical deployment requires not only correct predictions but also well-calibrated confidence estimates. Uncertainty-aware modeling, explicit calibration assessment, and safe "refer-or-abstain" strategies are increasingly recognized as essential components of trustworthy medical imaging systems, particularly for mitigating high-confidence errors under ambiguous or shifted inputs.

9**. Conclusion**

This survey presented a reliability-oriented synthesis of recent deep learning approaches for multi-class gastrointestinal (GI) disease classification using endoscopic images, focusing on studies published between 2023 and early 2026. The analysis revealed a clear methodological shift toward more expressive architectures, including transformer-based and hybrid CNN-transformer models. However, consistent with recent reviews, architectural advances alone do not guarantee clinically meaningful performance in realistic GI diagnostic settings.

A key conclusion of this survey is the persistent gap between reported benchmark performance and clinical reliability. Despite promising accuracy results, many studies continue to rely on accuracy-centric evaluation, underreport class-wise behavior, and omit robustness testing under domain shift conditions. Furthermore, while emerging directions such as self-supervised learning and hybrid ensemble strategies show potential to improve predictive stability, their impact on reliability-oriented criteria-including robustness to class imbalance and cross-domain consistency-has not yet been systematically validated.

While this review focused on recent image-based studies, the identified limitations highlight broader challenges in translating deep learning models into dependable clinical tools. Addressing these reliability gaps requires standardized evaluation protocols, clinically meaningful performance metrics, and rigorous validation across heterogeneous datasets. Only through such reliability-aware research practices can future GI diagnostic models progress from high-performing benchmarks to trustworthy systems suitable for real-world clinical deployment.

References

[1] D. Jha et al., "GastroVision: A Multi-class Endoscopy Image Dataset for Computer Aided Gastrointestinal Disease Detection," in Machine Learning for Multimodal Healthcare Data, vol. 14315, A. K. Maier, J. A. Schnabel, P. Tiwari, and O. Stegle, Eds., in Lecture Notes in Computer Science, vol. 14315., Cham: Springer Nature Switzerland, 2024, pp. 125–140. doi: 10.1007/978-3-031-47679-2_10.

[2] E. Ayan, "Classification of Gastrointestinal Diseases in Endoscopic Images: Comparative Analysis of Convolutional Neural Networks and Vision Transformers," Iğdır Üniversitesi Fen Bilim. Enstitüsü Derg., vol. 14, no. 3, pp. 988–999, Sep. 2024, doi: 10.21597/jist.1501787.

[3] H. Guo, S. A. Somayajula, R. Hosseini, and P. Xie, "Improving image classification of gastrointestinal endoscopy using curriculum self-supervised learning," Sci. Rep., vol. 14, no. 1, p. 6100, Mar. 2024, doi: 10.1038/s41598-024-53955-8.

[4] K. Pogorelov et al., "KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection." Jun. 20, 2017. doi: 10.1145/3193289.

[5] H. Borgli et al., "HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy," Sci. Data, vol. 7, no. 1, p. 283, Aug. 2020, doi: 10.1038/s41597-020-00622-y.

[6] P. H. Smedsrud et al., "Kvasir-Capsule, a video capsule endoscopy dataset," Sci. Data, vol. 8, no. 1, p. 142, May 2021, doi: 10.1038/s41597-021-00920-z.

[7] H. Üzen and H. Firat, "ÖZNİTELİK ENTEGRASYONUNA DAYALI ESA MİMARİSİ KULLANILARAK ENDOSKOPİK GÖRÜNTÜLERİN SINIFLANDIRILMASI," Kahramanmaraş Sütçü İmam Üniversitesi Mühendis. Bilim. Derg., vol. 27, no. 1, pp. 121–132, Mar. 2024, doi: 10.17780/ksujes.1362792.

[8] A. Ali, A. Iqbal, S. Khan, N. Ahmad, and S. Shah, "A two-phase transfer learning framework for gastrointestinal diseases classification," PeerJ Comput. Sci., vol. 10, p. e2587, Dec. 2024, doi: 10.7717/peerj-cs.2587.

[9] S. Rubab et al., "Gastrointestinal tract disease classification from wireless capsule endoscopy images based on deep learning information fusion and Newton Raphson controlled marine predator algorithm," Sci. Rep., vol. 15, no. 1, p. 32180, Sep. 2025, doi: 10.1038/s41598-025-17204-w.

[10] W. Wang, X. Yang, and J. Tang, "Vision Transformer with Hybrid Shifted Windows for Gastrointestinal Endoscopy Image Classification," IEEE Trans. Circuits Syst. Video Technol., vol. 33, no. 9, pp. 4452–4461, Sep. 2023, doi: 10.1109/TCSVT.2023.3277462.

[11] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2020, arXiv. doi: 10.48550/ARXIV.2010.11929.

[12] H. Malik, A. Naeem, A. Sadeghi-Niaraki, R. A. Naqvi, and S.-W. Lee, "Multi-classification deep learning models for detection of ulcerative colitis, polyps, and dyed-lifted polyps using wireless capsule endoscopy images," Complex Intell. Syst., vol. 10, no. 2, pp. 2477–2497, Apr. 2024, doi: 10.1007/s40747-023-01271-5.

[13] V. Tanwar, B. Sharma, D. P. Yadav, and A. Mehbodniya, "Hybrid deep learning framework based on EfficientViT for classification of gastrointestinal diseases," Sci. Rep., vol. 15, no. 1, p. 26982, Jul. 2025, doi: 10.1038/s41598-025-12128-x.

[14] S. Tabassum et al., "GastroViT: A Vision Transformer Based Ensemble Learning Approach for Gastrointestinal Disease Classification with Grad CAM & SHAP Visualization," 2025, arXiv. doi: 10.48550/ARXIV.2509.26502.

[15] S. Tang et al., "Transformer-based multi-task learning for classification and segmentation of gastrointestinal tract endoscopic images," Comput. Biol. Med., vol. 157, p. 106723, May 2023, doi: 10.1016/j.compbiomed.2023.106723.

[16] S. Wu et al., "High-Speed and Accurate Diagnosis of Gastrointestinal Disease: Learning on Endoscopy Images Using Lightweight Transformer with Local Feature Attention," Bioengineering, vol. 10, no. 12, p. 1416, Dec. 2023, doi: 10.3390/bioengineering10121416.

[17] Ş. Aslan, "Ensemble-Based Deep Transfer Learning for Robust Gastrointestinal Endoscopy Image Classification," Balk. J. Electr. Comput. Eng., vol. 13, no. 1, pp. 1–10, Mar. 2025, doi: 10.17694/bajece.1630294.

[18] H. Gunasekaran, K. Ramalakshmi, D. K. Swaminathan, A. J, and M. Mazzara, "GIT-Net: An Ensemble Deep Learning-Based GI Tract Classification of Endoscopic Images," Bioengineering, vol. 10, no. 7, p. 809, Jul. 2023, doi: 10.3390/bioengineering10070809.

[19] E. Sivari, E. Bostanci, M. S. Guzel, K. Acici, T. Asuroglu, and T. Ercelebi Ayyildiz, "A New Approach for Gastrointestinal Tract Findings Detection and Classification: Deep Learning-Based Hybrid Stacking Ensemble Models," Diagnostics, vol. 13, no. 4, p. 720, Feb. 2023, doi: 10.3390/diagnostics13040720.

[20] C. M. Tsai and J.-D. Lee, "Dynamic Ensemble Learning with Gradient-Weighted Class Activation Mapping for Enhanced Gastrointestinal Disease Classification," Electronics, vol. 14, no. 2, p. 305, Jan. 2025, doi: 10.3390/electronics14020305.

[21] M. Wortsman et al., "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," 2022, arXiv. doi: 10.48550/ARXIV.2203.05482.

[22] G. M. Foody, "Challenges in the real-world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient," PLOS ONE, vol. 18, no. 10, p. e0291908, Oct. 2023, doi: 10.1371/journal.pone.0291908.

[23] A. S. Sambyal, U. Niyaz, N. C. Krishnan, and D. R. Bathula, "Understanding calibration of deep neural networks for medical image classification," Comput. Methods Programs Biomed., vol. 242, p. 107816, Dec. 2023, doi: 10.1016/j.cmpb.2023.107816.

[24] S. Ali et al., "Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge," Sci. Rep., vol. 14, no. 1, p. 2032, Jan. 2024, doi: 10.1038/s41598-024-52063-x.

[25] S. Lobanovs, J. Aleksejeva, A. K. Rūtiņa, E. Krustiņš, J. Čižovs, and D. Bļizņuks, "Machine learning in gastrointestinal endoscopy: challenges and opportunities," BMJ Open Gastroenterol., vol. 12, no. 1, p. e001923, Oct. 2025, doi: 10.1136/bmjgast-2025-001923.

[26] A. Sagar, "BUCAN: Bayesian Uncertainty-aware Classification with Attention Networks for Medical Images," Nov. 06, 2025, Health Informatics. doi: 10.1101/2025.11.05.25339638.

[27] S. Mansour et al., "multi-class gastrointestinal disease detection using context-aware deep representation learning with feature fusion approach on biomedical endoscopic images," Eng. Appl. Artif. Intell., vol. 163, p. 113064, Jan. 2026, doi: 10.1016/j.engappai.2025.113064.

[28] O. Attallah, M. F. Aslan, and K. Sabanci, "EndoNet: A Multiscale Deep Learning Framework for Multiple Gastrointestinal Disease Classification via Endoscopic Images," Diagnostics, vol. 15, no. 16, p. 2009, Aug. 2025, doi: 10.3390/diagnostics15162009.
[29] S. Bin Wahid, Z. T. Rothy, R. K. News, and S. A. Rieyan, "Interpretable Deep Learning Approaches for Reliable GI Image Classification: A Study with the HyperKvasir Dataset," Jul. 23, 2025, Gastroenterology. doi: 10.1101/2025.07.22.25332009.
[30] A. Chhetri, J. Korhonen, P. Gyawali, and B. Bhattarai, "NERO: Explainable Out-of-Distribution Detection with Neuron-level Relevance," 2025, arXiv. doi: 10.48550/ARXIV.2506.15404.
[31] Z. M. Lonseko et al., "Deep multi-task learning framework for gastrointestinal lesion-aided diagnosis and severity estimation," Sci. Rep., vol. 15, no. 1, p. 25827, Jul. 2025, doi: 10.1038/s41598-025-09587-7.
[32] D. Agbelese et al., "MEGAN: Mixture of Experts for Robust Uncertainty Estimation in Endoscopy Videos," 2025, arXiv. doi: 10.48550/ARXIV.2509.12772.
[33] M. Fahad et al., "Deep insights into gastrointestinal health: A comprehensive analysis of GastroVision dataset using convolutional neural networks and explainable AI," Biomed. Signal Process. Control, vol. 102, p. 107260, Apr. 2025, doi: 10.1016/j.bspc.2024.107260.
[34] S. Siddiqui, J. A. Khan, and S. Algamdi, "Deep ensemble learning for gastrointestinal diagnosis using endoscopic image classification," PeerJ Comput. Sci., vol. 11, p. e2809, Apr. 2025, doi: 10.7717/peerj-cs.2809.
[35] Q. He, S. Bano, D. Stoyanov, and S. Zuo, "DivGI: delve into digestive endoscopy image classification," Int. J. Comput. Assist. Radiol. Surg., vol. 20, no. 7, pp. 1513–1520, Jun. 2025, doi: 10.1007/s11548-025-03441-x.
[36] A. Şener and B. Ergen, "Automatic detection of gastrointestinal system abnormalities using deep learning-based segmentation and classification methods," Health Inf. Sci. Syst., vol. 13, no. 1, p. 37, May 2025, doi: 10.1007/s13755-025-00354-6.
[37] S. Pokhrel et al., "NCDD: Nearest Centroid Distance Deficit for Out-Of-Distribution Detection in Gastrointestinal Vision," 2024, arXiv. doi: 10.48550/ARXIV.2412.01590.
[38] Md. F. Ahamed et al., "Detection of various gastrointestinal tract diseases through a deep learning method with ensemble ELM and explainable AI," Expert Syst. Appl., vol. 256, p. 124908, Dec. 2024, doi: 10.1016/j.eswa.2024.124908.
[39] A. Quindós, P. Laiz, J. Vitrià, and S. Seguí, "Self-supervised out-of-distribution detection in wireless capsule endoscopy images," Artif. Intell. Med., vol. 143, p. 102606, Sep. 2023, doi: 10.1016/j.artmed.2023.102606.
[40] A. Kamble et al., "Enhanced Multi-Class Classification of Gastrointestinal Endoscopic Images with Interpretable Deep Learning Model," 2025, arXiv. doi: 10.48550/ARXIV.2503.00780.
[41] A. A. Shafi, M. Ahmed, M. S. Rahman, M. S. Hossain, and M. F. Uddin, "Deep Learning for Imbalanced Gastrointestinal Image Classification: A Comparative Study of Architectural Choices," in Proceedings of the 3rd International Conference on Computing Advancements, Dhaka Bangladesh: ACM, Oct. 2024, pp. 741–746. doi: 10.1145/3723178.3723276.
[42] Z. Ozdemir, H. Y. Keles, and O. O. Tanriover, "CLoE: Curriculum Learning on Endoscopic Images for Robust MES Classification," 2025, arXiv. doi: 10.48550/ARXIV.2508.13280.
[43] Y. Yang, Y. Jin, Q. Tian, Y. Yang, W. Qin, and X. Ke, "Enhancing Gastrointestinal Diagnostics with YOLO-Based Deep Learning Techniques," Theor. Nat. Sci., vol. 95, no. 1, pp. 39–46, Feb. 2025, doi: 10.54254/2753-8818/2024.21125