Machine Learning in Gastrointestinal Tract Imaging: A Comprehensive Review of Techniques and Applications

Phongsakon Mark Konrad[1], Yaser Sabzehmeidani[2], Andrei-Alexandru Popa[2], Serkan Ayvaz[1*]

[1]Centre for Industrial Software, University of Southern Denmark, Sønderborg, 6400, Denmark.
[2]Centre for Industrial Mechanics, University of Southern Denmark, Sønderborg, 6400, Denmark.

*Corresponding author(s). E-mail(s): seay@mmmi.sdu.dk;
Contributing authors: phkon23@student.sdu.dk; yasers@sdu.dk; andrei@sdu.dk;

Gastrointestinal (GI) imaging modalities including endoscopy, colonoscopy, and wireless capsule endoscopy provide vital diagnostic information, yet their manual interpretation remains laborious. Although deep learning approaches, particularly convolutional neural networks (CNNs), hybrid architectures, and transformer-based models, have achieved high accuracy in this domain, their clinical adoption is impeded by data limitations and trust concerns. The current study (1) systematically maps algorithmic trends to specific GI imaging techniques, (2) quantifies the relationship between dataset size and model performance to identify data sufficiency thresholds, and (3) evaluates translational enablers, such as federated learning for data privacy and explainable AI for clinician trust, within established ethical frameworks. By situating our analysis at the intersection of methodological rigor, quantitative assessment, and clinical applicability, this work aims to establish a structured, data-informed baseline that advances beyond purely descriptive surveys and guides future innovation toward impactful clinical AI deployment.

Keywords: Gastrointestinal imaging, Machine learning, Deep learning, Endoscopy, Computer-Aided Diagnosis

List of Abbreviations

Acc. Accuracy
AI Artificial Intelligence
Alg. Type Algorithm Type
AUC Area Under the Curve
Barrett's NeoDe Barrett's Esophagus Neoplasia Detection
Bleeding De Gastrointestinal Bleeding Detection
CADx Computer-Aided Diagnosis
CaNeoCl Cancer and Neoplasia Classification/Detection
CapEndo Capsule Endoscopy
CapEndo Analysis Capsule Endoscopy Analysis
Clin. App. Clinical Application
ClPolypDet Colorectal Polyp Detection
CNN Convolutional Neural Network
ColNeoDe Colorectal Neoplasia Detection
CS Colonoscopy
DL Deep Learning
Early ESCC ID Early Esophageal Squamous Cell Carcinoma Identification
Endo Endoscopy / Endoscopic Imaging
Enteric PoDe Enteric Polyp Detection
Esophageal AbnDe Esophageal Abnormality Detection
FL Federated Learning
FPS Frames Per Second
GaCaDe Gastric Cancer Detection
GaLesDe Gastric Lesion Detection
GI Gastrointestinal
GIAngDe Gastrointestinal Angiodysplasia Detection
GIAbnDe Gastrointestinal Abnormality Detection
GIAbnDe (Crohn's) Gastrointestinal Abnormality Detection for Crohn's Disease
GI MucLesDe Gastrointestinal Mucosal Lesion Detection
GITrAbnId Gastrointestinal Tract Abnormality Identification
GITrImCl Gastrointestinal Tract Imaging Classification
Img. Mod. Imaging Modality
InMetaDe / InMetaGastAt Intestinal Metaplasia Detection / Identification / Gastric Atrophy
Landmark ID Anatomical Landmark Identification
Lower GI Dis Lower Gastrointestinal Disease Classification
ML Machine Learning
N/A Not Available / Not Applicable
Pleomorphic GaLesCl Pleomorphic Gastric Lesion Classification
R-CNN Region-based Convolutional Neural Network
Sens. Sensitivity
SmBwDe Small Bowel Disease Detection
Spec. Specificity
SSD Single Shot MultiBox Detector

SVM Support Vector Machine
UlcDe Ulcer Detection
UlcDe/Hem Ulcer Detection with Hemorrhage
UpGi Abn De Upper Gastrointestinal Abnormality Detection
UpGiEndo Upper Gastrointestinal Endoscopy Imaging Tasks
ViT Vision Transformer
WCE Wireless Capsule Endoscopy
xAI Explainable Artificial Intelligence

1 Introduction

Gastrointestinal (GI) disorders represent a significant global health challenge, characterized by high prevalence rates and substantial associated morbidity and mortality. The timely and accurate diagnosis of various conditions, including inflammatory diseases like Crohn's disease [1–3], potentially pre-malignant lesions such as polyps [4–8], ulcers and bleeding [9], angiectasia [10], and neoplasms in the esophagus, stomach, and colon [11–14], is critical for effective patient management and improved clinical outcomes. Essential diagnostic tools in gastroenterology include various imaging modalities like conventional endoscopy, gastroscopy, colonoscopy, and wireless capsule endoscopy (WCE) [3, 4, 15, 16], which collectively generate vast quantities of visual data requiring expert interpretation.

While manual review of these images by clinicians is the traditional standard, it is inherently labor-intensive and prone to significant inter-observer variability [11]. Such variability can lead to diagnostic inconsistencies and potential delays in treatment. Furthermore, the sheer volume of images, particularly from WCE which can generate tens of thousands of frames per procedure, presents a formidable challenge for manual review [16, 17]. These limitations underscore the need for more efficient, reliable, and standardized methods, positioning automated image analysis via machine learning (ML) as a crucial advancement in modern medical diagnostics [18, 19].

In recent years, the field has witnessed remarkable progress driven by deep learning (DL), especially convolutional neural networks (CNNs) [13, 20]. CNN-based systems excel at automatically learning hierarchical features from image data, enabling highly accurate detection, classification, localization, and even segmentation of pathological findings [4, 11, 21–24]. Numerous studies have demonstrated the potential of these systems, reporting diagnostic performances, including sensitivity and specificity often exceeding 90% and sometimes reaching over 95%, for various tasks such as identifying small-bowel diseases [17], detecting GI angiectasia [10], diagnosing ulcers and hemorrhage [9], classifying gastric mucosal lesions [25, 26], detecting gastric cancer [11], identifying precancerous conditions [27], detecting neoplasia in Barrett's esophagus [12], and locating colorectal polyps [6]. Advanced techniques like attention mechanisms enhance model focus on relevant image areas, while transfer learning allows leveraging knowledge from pre-trained networks, proving beneficial in medical imaging where labeled data can be scarce [8, 28–30].

A significant benefit of ML is its potential integration into real-time clinical workflows. Systems capable of analyzing endoscopic video streams at high frame rates (e.g., up to 30 frames per second (FPS)) have been developed to assist clinicians during live procedures like colonoscopy [6, 15]. These computer-aided diagnosis (CADx) tools can provide immediate feedback, highlight suspicious areas, and potentially reduce the rate of missed lesions, such as polyps [4, 6, 15]. For instance, Urban et al. [6] demonstrated a system achieving approximately 96% accuracy (Area Under the Curve (AUC) 0.991) for real-time polyp localization during screening colonoscopies, while others have focused on achieving high sensitivity for detection [15]. The applicability and robustness of DL models are being further confirmed through large-scale validation studies across diverse patient populations and clinical settings [3, 12, 25]. Artificial

intelligence (AI) is also proving invaluable in analyzing images from less invasive methods like standard or magnetically controlled WCE [1, 2, 10, 16, 17, 26, 31, 32].

Beyond diagnostic accuracy and speed, ML systems offer the promise of enhanced reproducibility and standardization in GI image interpretation. By reducing the subjectivity inherent in manual assessment, automated analysis can lead to more consistent diagnoses and facilitate objective monitoring of disease progression or treatment response over time [11]. The development of hybrid models, which might combine CNNs with other classifiers like Support Vector Machines (SVMs) or utilize ensemble techniques, represents ongoing efforts to further boost diagnostic performance, robustness, and potentially address challenges like cross-dataset bias [33–35].

Figure 1 visually summarizes the evolution of algorithmic approaches in GI imaging literature from 2017 to 2024. The chart illustrates the consistent dominance of CNNs, the rise of hybrid architectures (like CNN+SVM or ensembles), and the more recent exploration of transformer-based models, such as Vision Transformers (ViTs), for analyzing GI images.

Although several reviews have surveyed the landscape of ML in GI imaging [36, 37], they often focus primarily on cataloging applications, summarizing reported accuracies for specific tasks such as polyp detection or cancer classification, or providing a general overview of DL techniques [36]. However, critical gaps remain in systematic review of the field from a multidimensional perspective that integrates methodological trends with quantitative performance analysis and translational readiness. Specifically, few surveys have systematically mapped the evolution and clustering of specific algorithms (CNNs, hybrids, transformers) across different imaging modalities (Endoscopy, CS, WCE) to understand modality-specific adaptations. Furthermore, a quantitative investigation relating model performance metrics reported directly to dataset characteristics (e.g. size, diversity) across a broad range of studies is largely lacking, hindering insights into data sufficiency and generalizability bottlenecks [38].

Additionally, while challenges such as clinical integration and explainability are often mentioned [37, 39], a dedicated synthesis that examines emerging technical solutions (for example, federated learning for privacy [40], energy-efficient architectures for deployment, explainable AI methods for trust [41]) is lacking within the specific constraints of GI imaging workflows and regulatory pathways. This review aims to address these gaps by providing a comprehensive and integrative analysis centered on this methodological–quantitative–translational triad. Our novelty lies in this structured, data-informed synthesis that moves beyond cataloging to offer a baseline for evaluating progress and guiding future development towards clinically viable AI solutions in gastroenterology.

2 Machine Learning in Gastrointestinal Imaging

The literature review was conducted systematically through a semi-structured and carefully curated process that combined digital indexing tools with deliberate manual assessment. To gather relevant studies on ML and DL applications in GI imaging, we employed academic databases such as Semantic Scholar, PubMed, and IEEE Xplore. Query terms included combinations of *"gastrointestinal"*, *"endoscopy"*, *"colonoscopy"*,

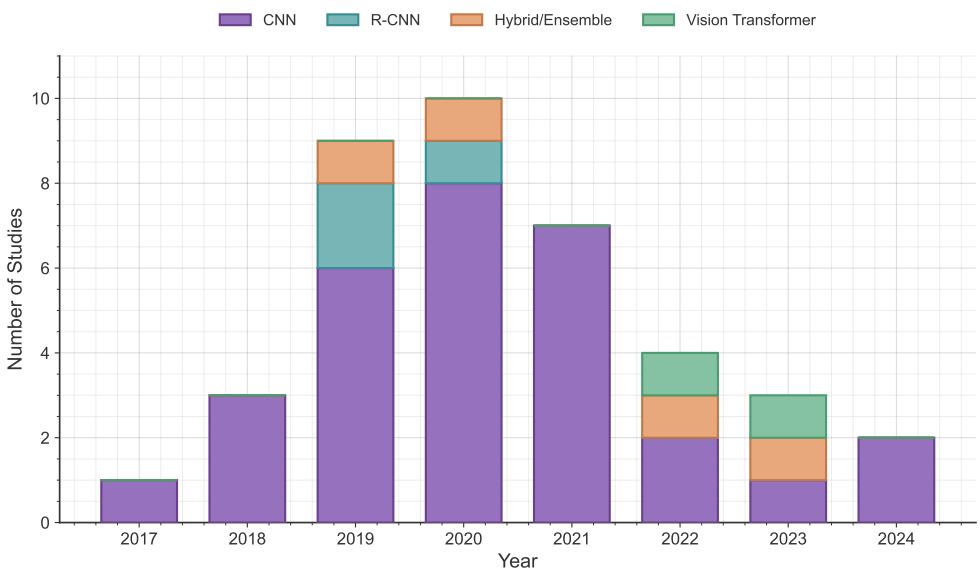

Fig. 1 Evolution of Algorithm Types in GI Imaging Studies (2017–2024). Stacked bar chart showing the annual distribution of machine learning approaches, consolidated into four families: CNN-based methods (purple), R-CNN detection architectures (teal), Hybrid/Ensemble approaches (orange), and Vision Transformers (green).

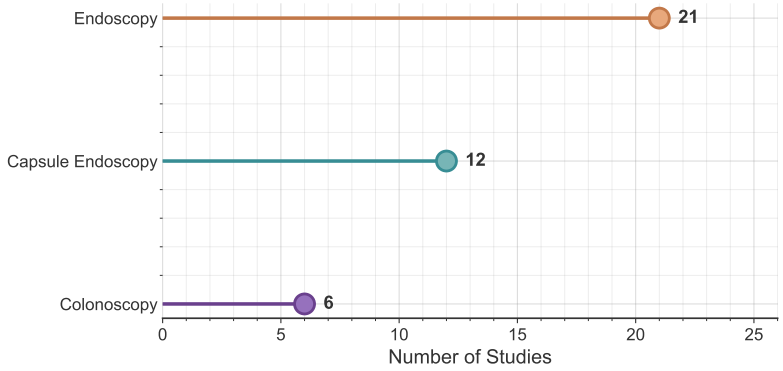

Fig. 2 Distribution of Imaging Modalities in Reviewed Studies. Lollipop chart illustrating the number of studies utilizing each imaging modality: Endoscopy (22 studies, 55%), Capsule Endoscopy (12 studies, 30%), and Colonoscopy (6 studies, 15%).

"capsule endoscopy", *"deep learning"*, *"convolutional neural networks"*, and other relevant clinical or algorithmic terms.

While modern academic search engines facilitated the initial filtering of publications, the review process itself was entirely human-supervised. Each study was manually screened at three levels: title, abstract, and full-text. Inclusion decisions were

made based on predefined criteria: direct relevance to GI imaging and diagnostics, explicit application of ML or DL techniques, presence of quantitative performance metrics (e.g. sensitivity, specificity or AUC) and peer-reviewed publication status. Studies published within the period 2017–2024 were included, with priority given to recent work (2019–2024) to ensure clinical relevance. Earlier high-impact studies, particularly those proposing novel architectures or widely cited benchmarks, were selectively included due to their foundational value.

The result of this multi-stage review process is visually summarized in Figure 2, which presents the proportional distribution of imaging modalities across the selected literature. As illustrated, endoscopy dominates the landscape of GI imaging research, followed by capsule endoscopy and colonoscopy, in alignment with clinical usage patterns. Additionally, Table 1 consolidates methodological and performance metrics across over 40 peer-reviewed studies, detailing algorithm types, clinical applications, dataset sizes, and quantitative outcomes such as sensitivity, specificity, AUC, and accuracy. This curated synthesis provides a comprehensive benchmark for evaluating progress in ML-based GI diagnostics and highlights algorithmic trends that have emerged over time.

Citation	Alg. Type	Img. Mod.	Clin. App.	Data Size	Sens. (%)	Spec. (%)	AUC	Acc. (%)
[13]	Deep Learning	Endo	CaNeoCl	N/A	91.2	80.0	N/A	N/A
[42]	CNN	CapEndo	GIAbnDe	6,000 images	99.3	99.3	0.991	99.2
[28]	Deep Learning	Endo	InMetaDe / InMetaGastAt	21,420 images	99.45	98.90	N/A	99.18
[26]	CNN	CapEndo	Pleomorphic GaLesCl	N/A	97.4	95.9	N/A	96.6
[34]	Hybrid (Stacking Ensemble)	Endo	GITrImCl	N/A	N/A	N/A	N/A	98.42(KvasirV2)
[43]	CNN	Endo	GIAbnDe	N/A	N/A	N/A	N/A	N/A
[33]	Hybrid (CNN+SVM)	Endo	Lower GI Dis	N/A	99.0	100.0	0.9987	99.3
[44]	ViT	Endo	InMetaDe	N/A	N/A	N/A	0.83	N/A
[3]	CNN	CapEndo	UlcDe	N/A	86.7	98.6	0.98	96.6
[45]	CNN	CapEndo	CapEndo Analysis	N/A	95.5	95.8	N/A	95.4
[9]	CNN	CapEndo	UlcDe	N/A	95.5	95.8	N/A	95.4
[14]	Deep Learning	Endo	UpGi Abn De	N/A	89.7	96.9	N/A	93.3
[30]	CNN (Transfer Learning)	Endo	Bleeding De	N/A	N/A	N/A	N/A	97.65
[2]	CNN	CapEndo	UlcDe	N/A	96.2	76.2	0.84	77.1
[25]	CNN	Endo	GI MucLesDe	1,366 patients	N/A	N/A	0.86	N/A
[32]	CNN	CapEndo	Enteric PoDe	N/A	96.8	96.5	N/A	N/A
[16]	Deep Learning (Meta-Analysis)	CapEndo	Various	N/A	95.0	94.0	N/A	>90
[5]	CNN	Endo	ClPolypDet	N/A	N/A	N/A	N/A	N/A
[4]	CNN (SSD)	CS	ClPolypDet	16,418 (train), 7,077 (val)	92.0	N/A	N/A	N/A
[46]	N/A	Endo	UpGiEndo	969,318 images	N/A	N/A	0.96	N/A
[12]	Hybrid (ResNet/U-Net)	Endo	Barrett's NeoDe	494,364 (pretrain); 1,704 (train/val)	90–93	83–88	N/A	88.2(ext. test)
[29]	CNN (Transfer Learning)	Endo	GIAbnDe	N/A	N/A	N/A	N/A	94.6
[35]	CNN (ResNet-152, DenseNet-161)	Endo	GITrImCl	N/A	94.6	N/A	N/A	N/A
[20]	CNN	CapEndo	UlcDe/Hem	N/A	95.5	95.8	N/A	95.4
[21]	CNN (R-CNN, ResNet101, SVM)	Endo	GITrAbnId	N/A	N/A	N/A	N/A	99.13
[1]	CNN	CapEndo	GIAbnDe (Crohn's)	17,640 images	N/A	N/A	0.99	N/A
[18]	CNN	CS	ClPolypDet	N/A	N/A	N/A	N/A	94.0
[31]	CNN & Faster R-CNN	Endo	GaLesDe	N/A	N/A	N/A	N/A	N/A
[17]	CNN	CapEndo	SmBwDe	113,426,569 images	99.88(patient)	N/A	N/A	N/A
[10]	CNN	CapEndo	GIAngDe	N/A	100.0	96.0	N/A	N/A
[27]	CNN (VGG16)	Endo	GI MucLesDe	200/70 images	93.0	95.0	0.98	N/A
[15]	Faster R-CNN (VGG16)	CS	ColNeoDe	139,961 images	97.3	99.0	0.975	N/A
[7]	CNN	CapEndo/CS	ClPolypDet	255 participants	97.1	93.3	N/A	96.4
[19]	CNN	Endo	Early ESCC ID	Varied	N/A	N/A	N/A	N/A
[23]	CNN	Endo	Landmark ID	8,000 images	N/A	N/A	N/A	98.45(Inception-v4)
[24]	Faster R-CNN w/ DenseNets	Endo	Esophageal AbnDe	N/A	90.2–95.0	N/A	N/A	N/A
[11]	CNN (SSD)	Endo	GaCaDe	13,584 (train); 2,296 (test)	92.2	N/A	N/A	N/A
[6]	CNN	CS	ClPolypDet	8,641 images	93.0	93.0	0.991	96.4
[22]	CNN	Endo	GIAbnDe	N/A	>80	N/A	N/A	N/A
[8]	CNN (Transfer Learning)	CS	ClPolypDet	1,930 images	87.6	N/A	N/A	85.9

Table 1 Summary table of GI imaging studies reporting various ML approaches, sorted by recency. Performance metrics are included when reported. "N/A" indicates data not reported or not applicable. Abbreviations are defined in the List of Abbreviations.

3 GI Tract Imaging and Machine Learning: Background and Foundations

Advances in AI and DL have steadily transformed the field of GI imaging over the past decade. The rapid development and deployment of ML models applied to endoscopic, colonoscopic, and capsule endoscopic data have yielded significant improvements in diagnostic accuracy, clinical decision support, and workflow efficiency. In this section, we provide a comprehensive overview of the primary imaging modalities and the evolution of ML approaches that have shaped the current landscape of GI diagnostics.

Figure 3 illustrates the general workflow of applying ML in GI imaging, encompassing data acquisition, AI analysis methodologies, the generation of clinical findings, and the overall impact and challenges in clinical practice.

Figure 4 offers a cross-sectional view of algorithmic deployments across different GI imaging modalities. The heatmap underscores the dominance of CNNs across all imaging types, with notable clusters in endoscopy and capsule endoscopy (CapEndo). Hybrid models and transformer-based methods, while less prevalent, are beginning to appear in the literature, signaling diversification in methodological strategies. This trend highlights not only the adaptability of CNNs to varied image inputs but also the growing exploration of ensemble and attention-based models for specialized diagnostic tasks. Figure 5 further illustrates the performance distribution across algorithm families, showing that CNN-based methods exhibit the widest accuracy range while Vision Transformers and Hybrid approaches cluster at higher performance levels.

3.1 Imaging Modalities

GI imaging now encompasses a wide range of advanced techniques, each critical for visualizing distinct regions of the GI tract. Ongoing improvements in hardware and imaging software have enabled clinicians to detect pathologies at earlier stages, thereby improving patient outcomes.

Upper Endoscopy

Upper endoscopy allows high-resolution visualization of the esophagus, stomach, and duodenum. This modality is central to diagnosing conditions such as gastritis, peptic ulcers, and early neoplastic changes. Technological enhancements including narrow-band imaging, high-definition sensors, and magnification endoscopy have further increased lesion contrast, facilitating more accurate tissue classification when combined with ML algorithms [11]. For instance, multicenter studies have demonstrated that DL models can standardize analysis, reduce inter-observer variability, and ultimately improve diagnostic consistency [12, 25].

Colonoscopy (CS)

Colonoscopy remains the gold standard for colorectal cancer screening. Its ability to inspect the entire colon in real time allows for the early detection of adenomatous polyps and colorectal neoplasia. Modern ML approaches, especially CNN-based systems, have enabled real-time polyp detection and localization, with studies reporting

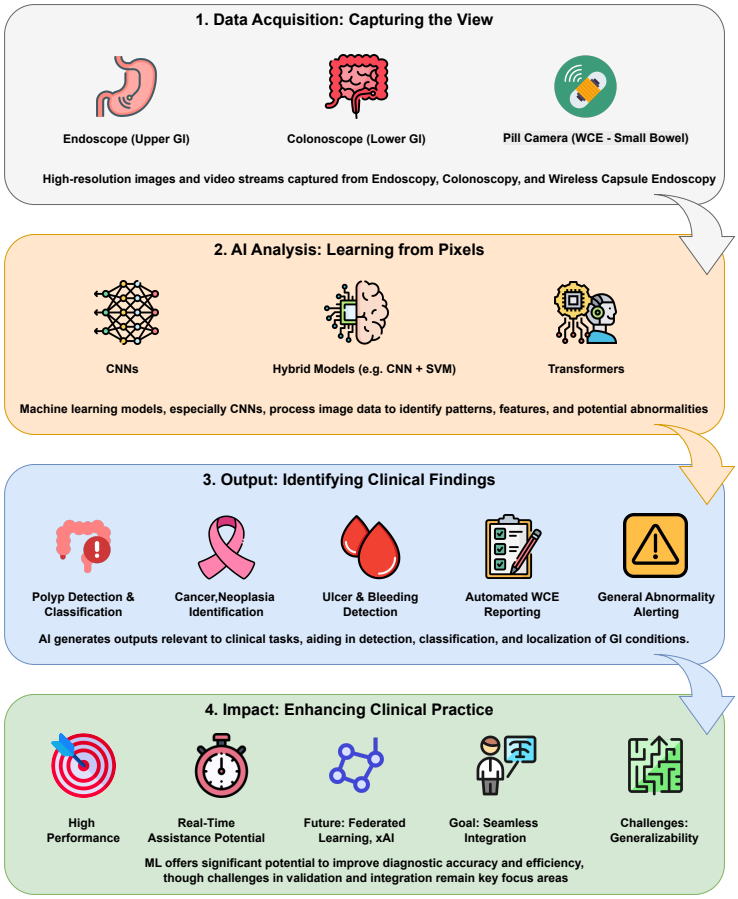

Fig. 3 Conceptual Workflow of Machine Learning Integration in Gastrointestinal Imaging. Overview of the typical process, from data acquisition via various endoscopic modalities (1), through analysis by different AI model architectures (2), to the generation of clinically relevant outputs such as lesion detection or classification (3). The workflow culminates in potential impacts on clinical practice, including enhanced diagnostic performance and real-time assistance, while also highlighting persistent challenges like model generalizability (4). Abbreviations are defined in the List of Abbreviations.

sensitivities exceeding 95% and AUC values approaching 1.0 [6, 15]. Furthermore, the integration of hybrid models that combine CNN-based feature extraction with classical classifiers (e.g. SVM) has been shown to mitigate false positive rates, thus improving both specificity and overall diagnostic value [21, 33].

Wireless Capsule Endoscopy (WCE)

WCE offers a minimally invasive means of imaging the small bowel, an area that poses significant challenges to conventional endoscopy. A single capsule procedure may generate tens of thousands of images, necessitating automated review. State-of-the-art

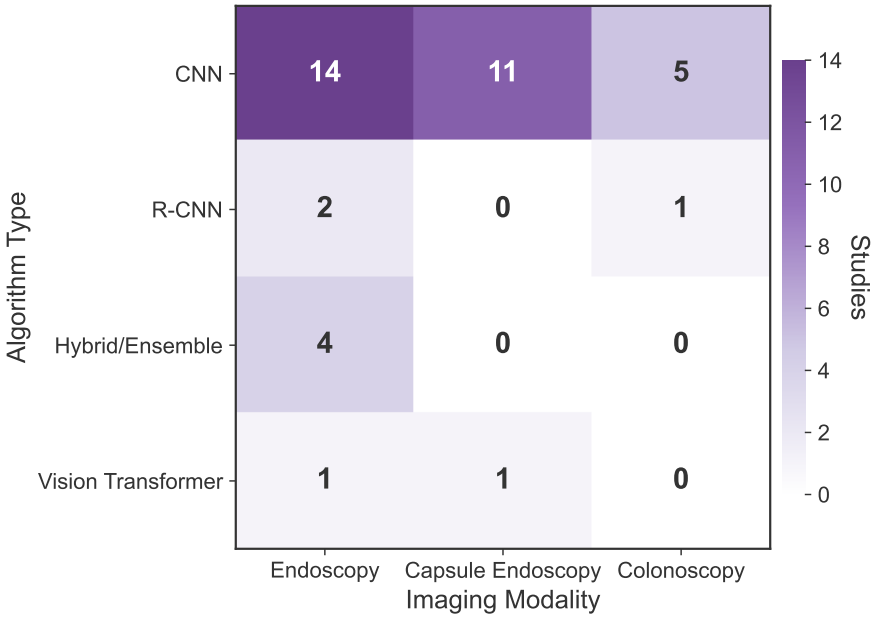

Fig. 4 Algorithm–Modality Distribution Matrix. Heatmap showing the co-occurrence of algorithm types and imaging modalities. CNN-based methods dominate across all modalities, with the highest concentration in standard endoscopy (14 studies).

CNN models have demonstrated the ability to detect bleeding, ulcers, and neoplasms with high sensitivity and reproducibility [10, 17]. Additionally, as datasets expand to encompass diverse patient populations, these algorithms have begun to support multicenter validations that further confirm their clinical applicability.

3.2 Machine Learning Approaches

The evolution of ML in GI imaging has been characterized by a progression from conventional CNN architectures to more sophisticated hybrid, attention-based, and Transformer-enhanced models. This progression reflects a dual drive: to improve raw diagnostic performance and to increase model interpretability – both key for clinical acceptance.

Standard CNN Architectures

CNNs such as VGG16, ResNet, and Inception have served as the foundational models in many GI imaging studies. These models are designed to extract hierarchical image features, thereby capturing both low-level textures and high-level anatomical patterns. In practice, many studies use transfer learning wherein models pretrained on large non-medical datasets (e.g., ImageNet) are fine-tuned on GI-specific data [8, 25, 27]. Their relative ease of training and robust performance in tasks like polyp detection and lesion classification make them a popular starting point.

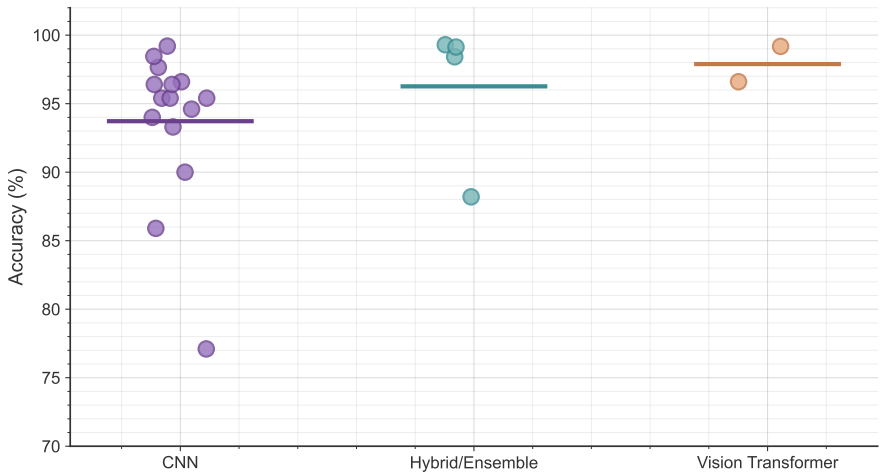

Fig. 5 Performance Distribution by Algorithm Type. Strip plot showing reported accuracy values for each algorithm family. Horizontal lines indicate mean performance. CNN-based methods show the widest range (77–99%), while Vision Transformers and Hybrid approaches cluster at higher accuracy levels (>96%).

Hybrid Models and Ensemble Techniques

To overcome challenges such as overfitting and high false-positive rates, researchers have developed hybrid approaches that blend conventional CNNs with classifiers like SVMs or use ensemble methods to merge predictions from multiple networks. For instance, hybrid (CNN+SVM) models combine the deep feature extraction capabilities of CNNs with the discriminative power of SVMs, while stacking ensemble methods leverage diverse model strengths to improve overall diagnostic accuracy [33–35]. Such methods are often more robust when applied to heterogeneous datasets and different clinical settings.

Attention Mechanisms and Transformers

More recently, attention-based frameworks and ViTs have been introduced into GI imaging. Unlike conventional CNNs, these approaches are designed to capture long-range dependencies and complex spatial relationships within images. For example, Yang et al. [28] integrated local attention grouping for detecting intestinal metaplasia, achieving sensitivity near 99%, while Neto et al. [44] demonstrated the utility of a ViT-based model for gastric lesion detection (GaLesDe) with an AUC around 0.83 despite limited data. While CNNs have dominated research, these emerging architectures hold significant potential, possibly enhancing feature extraction by capturing global context more effectively, which is vital for identifying subtle abnormalities. Architectures like ViTs, with inherent attention mechanisms, could also improve model interpretability [28]. Although these methods show promising improvements, extensive benchmarking on larger, diverse datasets is essential to confirm their superiority over traditional CNNs.

Additional Algorithms and Architectures

Beyond these general categories, various modifications and specialized architectures have been applied. Approaches such as Faster Region-based Convolutional Neural Network (R-CNN) and Single Shot MultiBox Detector (SSD) have been adapted to localize and detect lesions or polyps within complex endoscopic images [11, 15]. Given the unique imaging characteristics of WCE, models are often tailored or augmented to handle challenges associated with lower resolution or variable view angles (e.g., [3, 42, 45]). Furthermore, various architectures have been fine-tuned for specialized clinical tasks such as Barrett's esophagus neoplasia detection (Barrett's NeoDe), ulcer detection with hemorrhage (UlcDe/Hem), and detection of Crohn's disease-specific abnormalities (GIAbnDe (Crohn's)), underscoring the need to balance model complexity with clinical utility. A detailed comparison of methods and performance is provided in Table 1.

Advanced Preprocessing and Data Augmentation

Variability in imaging quality due to differences in equipment and patient-specific factors necessitates robust preprocessing pipelines. Techniques such as contrast enhancement, edge detection, and multi-scale analysis help standardize images before they are fed into ML systems. The MAPGI framework, for example, has been used to pre-process GI images by enhancing critical features and reducing noise [23]. Data augmentation strategies (e.g., rotation, scaling, and color jitter) also enrich training sets, improving model generalizability and reducing overfitting, which is especially important when dealing with limited datasets [29].

Emerging Methodologies: Federated Learning, Real-Time Systems, and Explainability

A forward-looking development is the application of federated learning (FL) to GI imaging. FL allows for distributed model training on data from multiple centers without requiring centralized data aggregation, thereby preserving patient privacy and potentially enhancing model robustness across diverse populations [17]. While still nascent in GI imaging, FL could significantly improve generalizability and compliance with data protection regulations. Future research is expected to integrate FL approaches to further validate AI systems' clinical utility.

In parallel, there is a strong focus on real-time applications. Systems developed for real-time polyp detection, for example, achieve frame rates of up to 30 FPS, demonstrating the practical feasibility of integrating ML into endoscopic procedures [6, 15]. However, as models increase in complexity, their computational and energy demands escalate, posing deployment challenges, particularly in resource-constrained settings or for edge computing. Developing resource-efficient algorithms through techniques like model compression, knowledge distillation, and designing lightweight architectures is therefore imperative for scalable and sustainable deployment without prohibitive costs. Rigorous benchmarking of these novel, efficient architectures against traditional CNNs is needed.

Furthermore, clinical acceptance is often hampered by the "black box" nature of complex models. Explainable Artificial Intelligence (xAI) techniques are critical

for bridging this gap by providing insights into the model's decision-making process, enhancing clinician trust, and facilitating safe adoption [28]. Future research should move beyond traditional saliency maps towards more advanced interpretability methods, such as local attention grouping, layer-wise relevance propagation, and counterfactual explanations. Integrating these interpretable outputs directly into clinical user interfaces is vital for validating model decisions and enabling clinician feedback for continuous improvement. Clinicians are more likely to rely on ML systems when they understand the basis of their outputs.

In summary, the evolution of ML in GI imaging—spanning traditional CNNs, hybrid architectures, attention methods, and burgeoning frameworks like FL—underscores a transformative shift. These advancements, supported by numerous studies [6, 10, 15, 17, 25], set the stage for improved clinical outcomes through faster, more reliable, and increasingly interpretable diagnostic systems.

4 Datasets and Performance Evaluation

Evaluating the performance of ML algorithms in GI imaging is fundamental to demonstrating their clinical utility and reliability. This section reviews dataset characteristics, performance metrics, evaluation methodologies, and discusses critical challenges related to data quality, model validation, and clinical integration.

4.1 Dataset Characteristics, Preprocessing, and Augmentation

Studies in GI imaging utilize datasets ranging from single-center collections to extensive, multicenter repositories. For instance, WCE studies may involve tens of thousands of images per patient [17], whereas datasets for colorectal polyp detection might comprise several thousand annotated images [6]. Beyond size and source variations, datasets often exhibit heterogeneity due to differences in image resolution, acquisition protocols, and patient demographics [17]. This heterogeneity significantly impacts the external validity of developed models.

Common preprocessing steps aim to mitigate this variability. Techniques such as histogram equalization, contrast enhancement, and geometric transformations help standardize image properties. Transfer learning, where models pre-trained on large datasets like ImageNet are fine-tuned for GI imaging [8], is frequently employed. Data augmentation, including rotations, flips, random cropping, and color jittering, artificially expands training sets, which is particularly crucial for smaller datasets (e.g., under 5,000 images) to enhance model robustness and reduce the risk of memorizing dataset-specific artifacts [4, 29]. Ongoing efforts also advocate for uniform image acquisition and annotation protocols to improve comparability across studies. In many studies, data are typically split into training and validation sets (e.g., 80%/20%), with some reserving a separate test set to better assess generalization and mitigate overfitting [4].

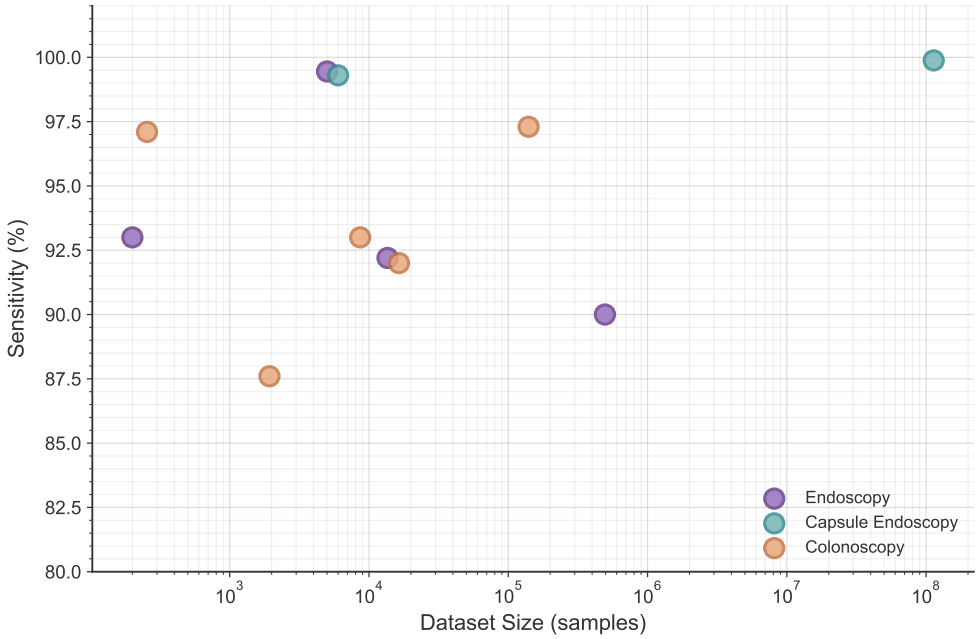

Fig. 6 Performance versus Dataset Size. Scatter plot correlating sensitivity (%) with dataset size (log scale). Points are colored by imaging modality. No strong correlation is observed, suggesting that model architecture and data quality may be more influential than dataset size alone.

4.2 Performance Metrics and Comparative Analysis

Performance evaluation relies heavily on standard metrics such as sensitivity, specificity, accuracy, and AUC-ROC. Many studies report high performance, with CNN-based models often yielding sensitivity and specificity values exceeding 90%, sometimes reaching over 95% [10, 16, 20]. High AUC values, typically above 0.95, further underscore robust discriminative power. For class-imbalanced tasks, metrics like precision, recall (sensitivity), F1 score, and precision-recall curves provide more nuanced insights into model performance.

As illustrated in Figure 6, a comparison of reported metrics against dataset size reveals interesting patterns. Clear performance clustering is observed in mature applications like colorectal polyp detection (ClPolypDet), where models like Urban et al. [6] consistently achieve high AUC values (near 0.99), even with moderately sized datasets. In contrast, tasks such as ulcer classification (UlcDe) and small-bowel lesion detection (SmBwDe) show more scattered results, often requiring larger datasets to reach comparable performance levels. This disparity highlights the importance of large-scale, standardized datasets and tailored model architectures for specific, complex GI applications.

4.3 Addressing Overfitting, Bias, and Annotation Quality

Despite impressive reported accuracies, often exceeding 90%, a critical concern is the reliance of many studies on relatively small or highly curated datasets. This increases the risk of *overfitting*, where models learn dataset-specific noise rather than generalizable patterns. Heterogeneity in imaging devices (resolution, illumination) across different centers can also introduce *device-induced bias*, limiting the model's applicability in new settings.

Annotation consistency represents another significant challenge. Inter-observer variability among expert gastroenterologists can lead to noisy or inconsistent labels, yet few studies explicitly quantify or correct for these discrepancies during training or evaluation. Furthermore, demographic bias is often overlooked; models trained on datasets lacking representation across diverse age groups, ethnicities, or sexes may not perform equitably or reliably for all patient populations. Addressing these issues requires a concerted effort towards standardized, multi-expert annotation protocols, transparent reporting of dataset demographics, and the development of techniques to actively identify and mitigate biases during model training. Ensuring equitable performance across diverse groups is not only a technical necessity but also an ethical imperative to prevent exacerbating health disparities.

4.4 Challenges in Validation, Clinical Integration, and Long-Term Viability

While ML models frequently demonstrate high performance in retrospective, single-center evaluations, their translation to real-world clinical practice faces substantial hurdles. A primary challenge is ensuring model robustness and generalizability. Rigorous prospective and multicenter evaluations are essential, as they consistently reveal performance degradation due to domain shifts—variations in data characteristics across different clinical environments [3, 12]. For instance, Ribeiro et al. documented sensitivity reductions when deploying a CNN model across institutions, linked to device variations and population differences [3]. Similarly, de Groof et al. highlighted the need for site-specific fine-tuning for a Barrett's neoplasia detection system [12]. These findings underscore the profound impact of data heterogeneity. Key obstacles include the limited availability of large-scale, consistently annotated multicenter datasets, variations across sites requiring sophisticated domain adaptation strategies, and the lack of standardized evaluation frameworks. Future efforts must prioritize developing such standards and exploring methodologies like FL to train models across diverse datasets securely [17].

Furthermore, demonstrating real-world value requires moving beyond retrospective analysis. Large-scale, prospective clinical trials are necessary to validate ML models in live clinical settings, confirming that theoretical performance translates into tangible benefits like improved patient outcomes and workflow efficiencies [25]. Trust is also undermined by potential publication bias favouring positive results; encouraging publication of replication studies, negative results, and adhering to comprehensive performance metrics (e.g., F1-scores, precision-recall curves, uncertainty quantification) is crucial for a balanced assessment.

Beyond validation, integrating validated systems into dynamic clinical workflows presents practical challenges. Algorithms must operate with minimal latency for real-time assistance [6, 15]. Fostering clinician trust necessitates enhancing model interpretability (as discussed in Section 3.2) and designing intuitive user interfaces. Moreover, deployed ML systems cannot remain static; they must adapt to evolving data distributions and clinical protocols. Implementing models capable of incremental or continuous learning is essential for maintaining performance over time. Alongside technical adaptation, ethical considerations are paramount, including maintaining patient confidentiality through robust data privacy measures (like differential privacy or secure multi-party computation) and adhering to regulatory requirements. Ultimately, successful deployment is a multidisciplinary endeavor requiring collaboration among clinicians, engineers, data scientists, and regulatory bodies.

5 Key Applications of Machine Learning in GI Imaging

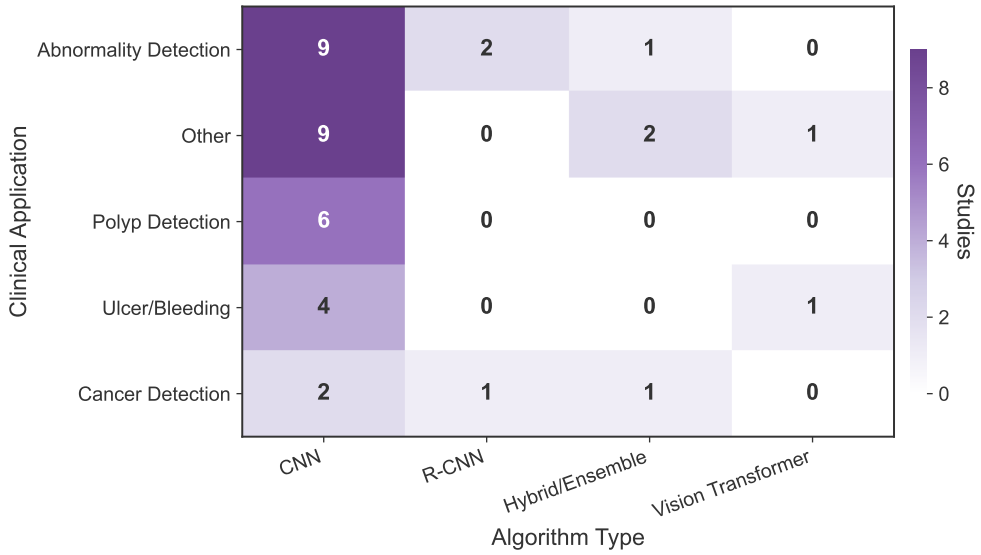

Fig. 7 Clinical Application versus Algorithm Type. Heatmap visualizing the frequency of algorithm deployment across clinical tasks. Polyp detection shows the broadest algorithmic diversity, while ulcer/bleeding detection is predominantly addressed using CNN-based methods.

Advances in ML have revolutionized GI imaging, enabling the development of systems that increase diagnostic accuracy, reduce review times, and standardize interpretations across institutions. The clinical applications span a broad range, from automated lesion detection to real-time decision support, transforming endoscopic procedures.

Figure 7 provides a structured visualization of how specific algorithm types have been employed across diverse GI clinical applications. Notably, CNNs are heavily concentrated in tasks such as ClPolypDet, general GI abnormality detection (GIAbnDe), and UlcDe, reflecting their adaptability. In contrast, hybrid models—like CNNs combined with SVMs or ensemble frameworks—have emerged in more specialized domains such as pleomorphic gastric lesion classification (Pleomorphic GaLesCl). The recent inclusion of ViTs in areas like GI tract abnormality identification (GITrAbnId) underscores the field's ongoing diversification. This mapping reveals how model selection aligns with the complexity and demands of different clinical problems.

5.1 Polyp Detection and Classification

Polyp detection is one of the most intensively studied applications. Early CNN-based systems, like that of Urban et al. [6], achieved real-time detection with high sensitivity and specificity (93%) and an outstanding AUC of 0.991, processing video at 25–30 FPS to provide near-instantaneous feedback during colonoscopy. Later studies, such as Yamada et al. [15], reported further improvements with systems operating at 30 FPS and sensitivities often exceeding 95%. Transfer learning has also been successfully applied (e.g., Zhang et al. [8]) to leverage large non-medical datasets, improving classification robustness. Hybrid and ensemble models integrating CNN outputs with classifiers like SVMs have emerged to further reduce false positives and enhance overall accuracy [16, 21, 33].

5.2 Gastric Cancer and Precancerous Conditions

Early detection of gastric cancer and related precancerous conditions is vital. Hirasawa et al. [11] demonstrated a DL system achieving 92.2% sensitivity for gastric cancer detection (GaCaDe). Nam et al. [25] reported an AUC of 0.86 for diagnosing GI mucosal lesions (GI MucLesDe), highlighting the challenge of differentiating subtle changes. Kikuchi et al. [13] showed high sensitivity (>90%) for cancer/neoplasia classification (CaNeoCl), though specificity sometimes remained lower (80%), indicating room for refinement. Integrating advanced imaging techniques (e.g., magnification endoscopy) with deep feature fusion may further bolster diagnostic performance.

5.3 Capsule Endoscopy for Small Bowel Pathologies

WCE generates vast image volumes, making automated ML analysis indispensable for reviewing the small bowel. Ding et al. [17] reported patient-level sensitivities up to 99.88% for detecting small-bowel diseases (SmBwDe). Leenhardt et al. [10] achieved near-perfect sensitivity (100%) for GI angiodysplasia detection (GIAngDe), with high specificity (96%). Prospective evaluations (e.g., Aoki et al. [45]) have demonstrated robust WCE system performance in clinical scenarios, underscoring ML's utility in handling large datasets for consistent outcomes.

5.4 Real-Time Clinical Decision Support

Integrating ML into real-time clinical workflows is a key frontier. Decision support tools processing images at up to 30 FPS provide immediate feedback during procedures like CS [15]. These systems not only improve lesion detection but also help standardize interpretation across clinicians [15]. Reviews emphasize AI's potential to improve clinical outcomes by reducing diagnostic delays and enhancing procedural efficiency [6, 25]. As interpretability methods mature (e.g., Grad-CAM, attention maps), clinician trust and adoption are likely to increase.

5.5 Emerging and Additional Applications

Beyond the major applications, ML techniques are being extended to other tasks. Several studies explore multi-class lesion detection, simultaneously identifying abnormalities like bleeding, polyps, and early cancers [16, 23]. Recent research integrates novel architectures like ViTs to capture long-range image dependencies [44]. Advanced ensemble and hybrid methods combining different neural networks continue to emerge, enhancing diagnostic precision and robustness [34, 35]. As mentioned earlier, FL is also being explored to address data variability and privacy concerns collaboratively.

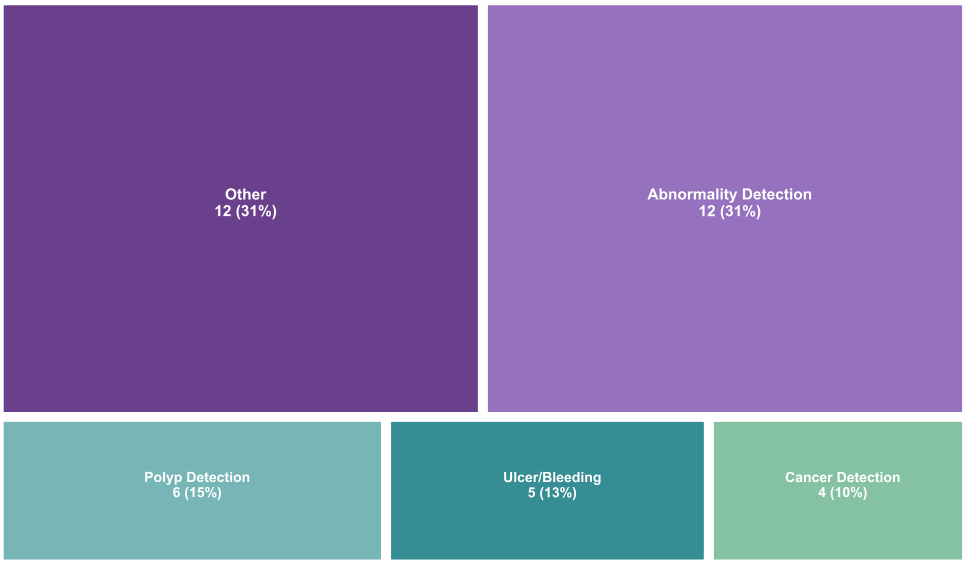

Fig. 8 Distribution of Clinical Applications. Treemap showing the relative focus of reviewed studies across consolidated categories. Abnormality detection (30%) and polyp detection (15%) together account for 45% of the literature, while diverse specialized applications comprise approximately one-third of studies.

Figure 8 shows the distribution of clinical focus areas. Abnormality detection represents the largest consolidated category (30%), followed by polyp detection

(15%) and ulcer/bleeding detection (12.5%). Notably, diverse specialized applications—including small-bowel disease detection (SmBwDe), gastric lesion classification, and tract imaging classification—collectively account for approximately one-third of studies, suggesting broadening ML use cases beyond traditional screening tasks.

In summary, ML applications in GI imaging show a transformative progression—from initial polyp/cancer detection systems to sophisticated, multi-task platforms capable of real-time analysis and broad clinical workflow integration. The adoption of hybrid models, attention networks, and FL suggests a future where AI augments clinical decision-making across diverse GI pathologies.

6 Conclusion

Machine learning has emerged as a transformative force in gastrointestinal imaging, offering substantial improvements in diagnostic accuracy, consistency, and efficiency, alongside the potential for real-time clinical decision support. Modern deep learning models, particularly CNNs and increasingly sophisticated architectures, consistently achieve high performance metrics, often exceeding 90% sensitivity and specificity for various tasks, enabling clinicians to detect and classify abnormalities more effectively and potentially earlier during endoscopic procedures.

Despite significant progress, the path toward widespread and seamless clinical adoption is ongoing. Key challenges remain, particularly concerning model generalizability across diverse clinical settings and patient populations, the need for robust validation through large-scale prospective trials, and ensuring model transparency and interpretability to foster clinician trust. Integrating these powerful tools effectively into complex clinical workflows while addressing computational demands and ethical considerations, such as data privacy and bias mitigation, requires continued multidisciplinary effort.

Future developments should focus on enhancing model robustness, perhaps through federated learning and advanced domain adaptation techniques. Continued research into explainable AI and the creation of resource-efficient models suitable for real-time deployment are also crucial. Addressing these technical, validation, and ethical hurdles will be essential for realizing the full potential of ML. Overall, the continuous evolution and rigorous validation of ML in GI imaging hold exceptional promise for elevating diagnostic standards, streamlining clinical workflows, and ultimately improving patient outcomes in gastroenterology.

References

[1] Klang, E., Barash, Y., Margalit, R.Y., Soffer, S., Shimon, O., Albshesh, A., Ben-Horin, S., Amitai, M.M., Eliakim, R., Kopylov, U.: Deep learning algorithms for automated detection of Crohn's disease ulcers by video capsule endoscopy. Gastrointestinal Endoscopy **91**(3), 606–6132 (2020) https://doi.org/10.1016/j.gie.2019.11.012

[2] J., F., Miguel, J.d.Q.E.C.d.M.S., J., A., T., R., Hélder, C., Ana, P.R.A., Miguel, N.G.d.M.S., M., P., R., N.N.J., Susana, I.O.L., Guilherme, M.G.d.M.: Identification of ulcers and erosions by the novel PillcamTM Crohn's Capsule using a convolutional neural network: a multicentre pilot study. Journal of Crohn's & Colitis (2021) https://doi.org/10.1093/ecco-jcc/jjab117

[3] T., R., M., M., J., A., H., C., A., A., S., L., M., M.S., J., F., G., M.: P156 A multicentric study on the development and application of a deep learning algorithm for automatic detection of ulcers and erosions in the novel PillCamTM Crohn's capsule. Journal of Crohn's & Colitis (2022) https://doi.org/10.1093/ecco-jcc/jjab232.284

[4] Ozawa, T., Ishihara, S., Fujishiro, M., Kumagai, Y., Shichijo, S., Tada, T.: Automated endoscopic detection and classification of colorectal polyps using convolutional neural networks. Therapeutic Advances in Gastroenterology **13** (2020) https://doi.org/10.1177/1756284820910659

[5] Sánchez-Peralta, L.F., Bote-Curiel, L., Picón, A., Sánchez-Margallo, F.M., Pagador, J.B.: Deep learning to find colorectal polyps in colonoscopy: A systematic literature review. Artificial Intelligence in Medicine **108**, 101923 (2020) https://doi.org/10.1016/j.artmed.2020.101923

[6] Urban, G., Tripathi, P., Alkayali, T., Mittal, M., Jalali, F., Karnes, W., Baldi, P.: Deep Learning Localizes and Identifies Polyps in Real Time With 96% Accuracy in Screening Colonoscopy. Gastroenterology **155**(4), 1069–10788 (2018) https://doi.org/10.1053/j.gastro.2018.06.037

[7] Blanes-Vidal, V., Baatrup, G., Nadimi, E.S.: Addressing priority challenges in the detection and assessment of colorectal polyps from capsule endoscopy and colonoscopy in colorectal cancer screening using machine learning. Acta Oncologica **58**(sup1), 29–36 (2019) https://doi.org/10.1080/0284186x.2019.1584404

[8] Zhang, R., Zheng, Y., Mak, T.W.C., Yu, R., Wong, S.H., Lau, J.Y.W., Poon, C.C.Y.: Automatic Detection and Classification of Colorectal Polyps by Transferring Low-Level CNN Features From Nonmedical Domain. IEEE Journal of Biomedical and Health Informatics **21**(1), 41–47 (2017) https://doi.org/10.1109/jbhi.2016.2635662

[9] Mohan, B.P., Khan, S.R., Kassab, L.L., Ponnada, S., Chandan, S., Ali, T., Dulai, P.S., Adler, D.G., Kochhar, G.S.: High pooled performance of convolutional neural networks in computer-aided diagnosis of GI ulcers and/or hemorrhage on wireless capsule endoscopy images: a systematic review and meta-analysis. Gastrointestinal Endoscopy **93**(2), 356–3644 (2021) https://doi.org/10.1016/j.gie.2020.07.038

[10] Leenhardt, R., Vasseur, P., Li, C., Saurin, J.C., Rahmi, G., Cholet, F., Becq, A., Marteau, P., Histace, A., Dray, X., Sacher-Huvelin, S., Mesli, F., Leandri,

C., Nion-Larmurier, I., Lecleire, S., Gerard, R., Duburque, C., Vanbiervliet, G., Amiot, X., Philippe Le Mouel, J., Delvaux, M., Jacob, P., Simon-Shane, C., Romain, O.: A neural network algorithm for detection of GI angiectasia during small-bowel capsule endoscopy. Gastrointestinal Endoscopy **89**(1), 189–194 (2019) https://doi.org/10.1016/j.gie.2018.06.036

[11] Hirasawa, T., Aoyama, K., Tanimoto, T., Ishihara, S., Shichijo, S., Ozawa, T., Ohnishi, T., Fujishiro, M., Matsuo, K., Fujisaki, J., Tada, T.: Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. Gastric Cancer **21**(4), 653–660 (2018) https://doi.org/10.1007/s10120-018-0793-2

[12] Groof, A.J., Struyvenberg, M.R., Putten, J., Sommen, F., Fockens, K.N., Curvers, W.L., Zinger, S., Pouw, R.E., Coron, E., Baldaque-Silva, F., Pech, O., Weusten, B., Meining, A., Neuhaus, H., Bisschops, R., Dent, J., Schoon, E.J., With, P.H., Bergman, J.J.: Deep-Learning System Detects Neoplasia in Patients With Barrett's Esophagus With Higher Accuracy Than Endoscopists in a Multistep Training and Validation Study With Benchmarking. Gastroenterology **158**(4), 915–9294 (2020) https://doi.org/10.1053/j.gastro.2019.11.030

[13] Kikuchi, R., Okamoto, K., Ozawa, T., Shibata, J., Ishihara, S., Tada, T.: Endoscopic Artificial Intelligence for Image Analysis in Gastrointestinal Neoplasms. Digestion **105**(6), 419–435 (2024) https://doi.org/10.1159/000540251

[14] M., E., Lcgp, H., W., L., I., L., O., A., M., B., C., M., D., A., H., H., D., G., Tom, K.M.V., L., L., S., O., S., K., HsiuPo, W., Wen-Lun, W., R., H.: Artificial intelligence for the real-time classification of intrapapillary capillary loop patterns in the endoscopic diagnosis of early oesophageal squamous cell carcinoma: A proof-of-concept study. United European Gastroenterology journal (2019) https://doi.org/10.1177/2050640618821800

[15] Yamada, M., Saito, Y., Imaoka, H., Saiko, M., Yamada, S., Kondo, H., Takamaru, H., Sakamoto, T., Sese, J., Kuchiba, A., Shibata, T., Hamamoto, R.: Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy. Scientific Reports **9**(1) (2019) https://doi.org/10.1038/s41598-019-50567-5

[16] Soffer, S., Klang, E., Shimon, O., Nachmias, N., Eliakim, R., Ben-Horin, S., Kopylov, U., Barash, Y.: Deep learning for wireless capsule endoscopy: a systematic review and meta-analysis. Gastrointestinal Endoscopy **92**(4), 831–8398 (2020) https://doi.org/10.1016/j.gie.2020.04.039

[17] Ding, Z., Shi, H., Zhang, H., Meng, L., Fan, M., Han, C., Zhang, K., Ming, F., Xie, X., Liu, H., Liu, J., Lin, R., Hou, X.: Gastroenterologist-Level Identification of Small-Bowel Diseases and Normal Variants by Capsule Endoscopy Using a Deep-Learning Model. Gastroenterology **157**(4), 1044–10545 (2019) https://doi.org/10.1053/j.gastro.2019.06.025

[18] S., G., Matthew, B.M.: Siri here, cecum reached, but please wash that fold: Will artificial intelligence improve gastroenterology? Gastrointestinal Endoscopy (2020) https://doi.org/10.1016/j.gie.2019.10.027

[19] Ebigbo, A., Messmann, H.: Artificial intelligence in the upper GI tract: the future is fast approaching. Gastrointestinal Endoscopy (2021) https://doi.org/10.1016/j.gie.2021.01.012

[20] Mohan, B.P., Khan, S.R., Kassab, L.L., Ponnada, S., Dulai, P.S., Kochhar, G.S.: Accuracy of convolutional neural network-based artificial intelligence in diagnosis of gastrointestinal lesions based on endoscopic images: A systematic review and meta-analysis. Endoscopy International Open **08**(11), 1584–1594 (2020) https://doi.org/10.1055/a-1236-3007

[21] Khan, M.A., Khan, M.A., Ahmed, F., Mittal, M., Goyal, L.M., Jude Hemanth, D., Satapathy, S.C.: Gastrointestinal diseases segmentation and classification based on duo-deep architectures. Pattern Recognition Letters **131**, 193–204 (2020) https://doi.org/10.1016/j.patrec.2019.12.024

[22] Iakovidis, D.K., Georgakopoulos, S.V., Vasilakakis, M., Koulaouzidis, A., Plagianakos, V.P.: Detecting and Locating Gastrointestinal Anomalies Using Deep Learning and Iterative Cluster Unification. IEEE Transactions on Medical Imaging **37**(10), 2196–2210 (2018) https://doi.org/10.1109/tmi.2018.2837002

[23] Cogan, T., Cogan, M., Tamil, L.: Mapgi: Accurate identification of anatomical landmarks and diseased tissue in gastrointestinal tract using deep learning. Comput. Biol. Medicine (2019) https://doi.org/10.1016/j.compbiomed.2019.103351

[24] Ghatwary, N., Ye, X., Zolgharni, M.: Esophageal Abnormality Detection Using DenseNet Based Faster R-CNN With Gabor Features. IEEE Access **7**, 84374–84385 (2019) https://doi.org/10.1109/access.2019.2925585

[25] Nam, J.Y., Chung, H.J., Choi, K.S., Lee, H., Han, S.J., Kim, T.J., Soh, H., Kang, E.A., Cho, S.-J., Ye, J.C., Im, J.P., Kim, S.G., Kim, Y.J., Kim, J.S., Yoon, J.-H., Chung, H., Lee, J.-H.: A Deep Learning Model for Diagnosing Gastric Mucosal Lesions Using Endoscopic Images: Development, Validation, and Method Comparison. SSRN Electronic Journal (2021) https://doi.org/10.2139/ssrn.3766771

[26] M., M., F., M., T., R., J., A., P., C., M., M., Hélder, C., P., A., J., F., M., M.M.S., G., M.: Deep Learning and Minimally Invasive Endoscopy: Automatic Classification of Pleomorphic Gastric Lesions in Capsule Endoscopy. Clinical and Translational Gastroenterology (2023) https://doi.org/10.14309/ctg.0000000000000609

[27] Guimarães, P., Keller, A., Fehlmann, T., Lammert, F., Casper, M.: Deep-learning based detection of gastric precancerous conditions. Gut **69**(1), 4–6 (2019) https:

//doi.org/10.1136/gutjnl-2019-319347

[28] Yang, J., Ou, Y., Chen, Z., Liao, J., Sun, W., Luo, Y., Luo, C.: A Benchmark Dataset of Endoscopic Images and Novel Deep Learning Method to Detect Intestinal Metaplasia and Gastritis Atrophy. IEEE Journal of Biomedical and Health Informatics **27**(1), 7–16 (2023) https://doi.org/10.1109/jbhi.2022.3217944

[29] Escobar, J.P., Gomez, N., Sanchez, K., Arguello, H.: Transfer Learning with Convolutional Neural Network for Gastrointestinal Diseases Detection using Endoscopic Images. In: 2020 IEEE Colombian Conference on Applications of Computational Intelligence (IEEE ColCACI 2020), pp. 1–6 (2020). https://doi.org/10.1109/colcaci50549.2020.9247847 . IEEE. http://dx.doi.org/10.1109/ColCACI50549.2020.9247847

[30] Caroppo, A., Leone, A., Siciliano, P.: Deep transfer learning approaches for bleeding detection in endoscopy images. Comput. Medical Imaging Graph. (2021) https://doi.org/10.1016/j.compmedimag.2020.101852

[31] J., X., T., X., Jun, P., Fei, G., Shuang, W., Yangyang, Q., Heng, W., Jie, Z., Xi, J., WenBin, Z., Yuan-Chen, W., W., Z., Zhaoshen, L., Z., L.: Use of artificial intelligence for detection of gastric lesions by magnetically controlled capsule endoscopy. Gastrointestinal Endoscopy (2020) https://doi.org/10.1016/j.gie.2020.05.027

[32] Mascarenhas Saraiva, M., Afonso, J., Ribeiro, T., Ferreira, J., Cardoso, H., Andrade, P., Gonalves, R., Cardoso, P., Parente, M., Jorge, R., Macedo, G.: Artificial intelligence and capsule endoscopy: automatic detection of enteric protruding lesions using a convolutional neural network. Revista Española de Enfermedades Digestivas (2021) https://doi.org/10.17235/reed.2021.7979/2021

[33] Fati, S.M., Senan, E.M., Azar, A.T.: Hybrid and Deep Learning Approach for Early Diagnosis of Lower Gastrointestinal Diseases. Italian National Conference on Sensors (2022) https://doi.org/10.3390/s22114079

[34] Sivari, E., Bostanci, E., Guzel, M.S., Acici, K., Asuroglu, T., Ercelebi Ayyildiz, T.: A New Approach for Gastrointestinal Tract Findings Detection and Classification: Deep Learning-Based Hybrid Stacking Ensemble Models. Diagnostics **13**(4), 720 (2023) https://doi.org/10.3390/diagnostics13040720

[35] Thambawita, V., Jha, D., Hammer, H.L., Johansen, H.D., Johansen, D., Halvorsen, P., Riegler, M.A.: An Extensive Study on Cross-Dataset Bias and Evaluation Metrics Interpretation for Machine Learning Applied to Gastrointestinal Tract Abnormality Classification. ACM Transactions on Computing for Healthcare **1**(3), 1–29 (2020) https://doi.org/10.1145/3386295

[36] Byrne, M.F., Chapados, N., Soudan, F., Oertel, C., Pérez-Cuadrado Robles, L., Omazic, B., Ahmad, O.F., Aabakken, L., Barkun, A.N., Jover, R., Pohl, H., Rösch, T., Wallace, M.B.: Real-time differentiation of adenomatous and

hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. Gut **68**(1), 94–100 (2019) https://doi.org/10.1136/gutjnl-2017-314547

[37] Mori, Y., Kudo, S.-e., Misawa, M., Saito, Y., Ikematsu, H., Hotta, K., Ohtsuka, K., Urushibara, F., Kataoka, S., Ogawa, Y., Maeda, Y., Takeda, K., Nakamura, H., Ichimasa, K., Kudo, T., Hayashi, T., Wakamura, K., Baba, T., Ishida, F., Inoue, H., Itoh, H., Oda, M., Igarashi, M.: Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy: A prospective study. Annals of Internal Medicine **169**(6), 357–366 (2018) https://doi.org/10.7326/M18-0249

[38] Varoquaux, G., Cheplygina, V.: Machine learning for medical imaging: methodological failures and recommendations for the future. npj Digital Medicine **5**(1), 48 (2022) https://doi.org/10.1038/s41746-022-00592-y

[39] Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D.: Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine **17**(1), 195 (2019) https://doi.org/10.1186/s12916-019-1426-2

[40] Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R.M., Trask, A., Xu, D., Baust, M., Cardoso, M.J.: The future of digital health with federated learning. npj Digital Medicine **3**(1), 119 (2020) https://doi.org/10.1038/s41746-020-00323-1

[41] Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Müller, K.-R.: Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE **109**(3), 247–278 (2021) https://doi.org/10.1109/JPROC.2021.3060483

[42] S., R., C., S.H., J., G., V., S., Naīm, A., Ahmed, S.B., G., K.K., P., D., G., E.C., V., S.: Gastronet: A CNN based system for detection of abnormalities in gastrointestinal tract from wireless capsule endoscopy images. AIP Advances (2024) https://doi.org/10.1063/5.0208691

[43] Iqbal, I., Walayat, K., Kakar, M.U., Ma, J.: Automated identification of human gastrointestinal tract abnormalities based on deep convolutional neural network with endoscopic images. Intelligent Systems with Applications **16**, 200149 (2022) https://doi.org/10.1016/j.iswa.2022.200149

[44] Neto, A., Ferreira, S., Libânio, D., Dinis-Ribeiro, M., Coimbra, M., Cunha, A.: Preliminary Study of Deep Learning Algorithms for Metaplasia Detection in Upper Gastrointestinal Endoscopy. International ICST Conference on Wireless Mobile Communication and Healthcare (2022) https://doi.org/10.1007/978-3-031-32029-3_4

[45] Aoki, T., Yamada, A., Koike, K.: The exceptional performance of deep learning for capsule endoscopy: Will such quality be maintained in clinical scenarios? Gastrointestinal Endoscopy **93**(2), 365–366 (2021) https://doi.org/10.1016/j.gie.2020.08.014

[46] Lui, T.K.L., Tsui, V.W.M., Leung, W.K.: Accuracy of artificial intelligence–assisted detection of upper GI lesions: a systematic review and meta-analysis. Gastrointestinal Endoscopy **92**(4), 821–8309 (2020) https://doi.org/10.1016/j.gie.2020.06.034