Artificial Intelligence in Gastroenterology

Inga Strümke, Steven A. Hicks, Vajira Thambawita, Debesh Jha, Sravanthi Parasa, Michael A. Riegler, and Pål Halvorsen

Contents

I. Strümke · M. A. Riegler
SimulaMet, Oslo, Norway
e-mail: inga@simula.no; michael@simula.no

S. A. Hicks · V. Thambawita · P. Halvorsen (✉)
SimulaMet, Oslo, Norway

Department of Computer Science, Oslo Metropolitan University, Oslo, Norway
e-mail: steven@simula.no; vajira@simula.no; paalh@simula.no; palh@oslomet.no

D. Jha
SimulaMet, Oslo, Norway

Department of Computer Science, UIT The Arctic University of Norway, Oslo, Norway
e-mail: debesh@simula.no

S. Parasa
Department of Gastroenterology, Swedish Medical Group, Seattle, WA, USA

N. Lidströmer, H. Ashrafian (eds.), *Artificial Intelligence in Medicine*,
https://doi.org/10.1007/978-3-030-58080-3_163-2

Abstract

The holy grail in endoscopy examinations has for a long time been assisted diagnosis using Artificial Intelligence (AI). Recent developments in computer hardware are now enabling technology to equip clinicians with promising tools for computer-assisted diagnosis (CAD) systems. However, creating viable models or architectures, training them, and assessing their ability to diagnose at a human level, are complicated tasks. This is currently an active area of research, and many promising methods have been proposed. In this chapter, we give an overview of the topic. This includes a description of current medical challenges followed by a description of the most commonly used methods in the field. We also present example results from research targeting some of these challenges, and a discussion on open issues and ongoing work is provided. Hopefully, this will inspire and enable readers to future develop CAD systems for gastroenterology.

Keywords

Gastrointestinal endoscopy · Artificial Intelligence · Neural Networks · Hand-crafted features · Anomaly detection · Semantic segmentation · Performance

1 Introduction

Numerous abnormal mucosal findings, ranging from minor annoyances to highly lethal diseases, can be found in the human Gastrointestinal (GI) tract. For example, according to the International Agency for Research on Cancer, about 3.5 million luminal GI (esophageal, stomach, colorectal) cancers are detected yearly in the world [41]. These cancers represent a substantial health challenge for society, with a mortality rate of about 63–65%, resulting in around 2.2 million deaths per year [19, 41]. Overall, Colorectal cancer (CRC) is the third most common cause of cancer mortality for women and men combined [104], and the other most frequently occurring GI cancers are stomach, liver, pancreatic, and esophageal cancers [18].

For diagnosis and treatment of GI diseases, GI endoscopy is the gold-standard procedure used to examine the tract for anomalies, and to a certain extent, the GI diseases may be prevented by improved endoscopic performance and high quality systematic screening in high incidence areas [19]. However, despite the substantial technical improvement of endoscopes over the last two decades, a major limitation of the endoscopic examinations is the endoscope operator variation, depending on the procedural skill, perceptual factors, personality characteristics, experience, knowledge, and attitude deficits [34]. This translates to a substantial inter-observer variation in the detection and assessment of mucosal lesions [64, 108]. This causes, for example, an average 20% polyp miss-rate during colonoscopies [52]. All these factors could potentially, to some extent, be alleviated by substantial educational efforts, but not eliminated [88].

In this context, assisted diagnosis using computers has for a long time been a holy grail. Developments in computer hardware have enabled computationally demanding yet promising technologies like AI, more specifically its sub-field Machine Learning (ML), to provide the clinicians with potentially highly accurate and efficient Computer aided diagnosis (CAD) systems, giving healthcare professionals the tools needed to provide quality care at a large scale [86, 102]. At its core, machine learning involves using algorithms to parse data, learn from it, and then make predictions, in the medical domain this means detect, segment, assess or classify a disease. However, there exist several issues which need to be addressed, both for creating and improving automated diagnosis algorithms. Developing and assessing a computer's ability to diagnose at a human level are complicated tasks, and a potential success depends on various factors which goes beyond simply determining the accuracy of an algorithm. These challenges have been an active area of research for about a decade, and a large number of promising results have been published.

In this chapter, we describe current challenges on the way towards effective computer-based digital assistant systems. In particular, we focus on GI endoscopy. We provide examples of proposed methods and tools employing various techniques, identify current challenges, and give hints for future development and assessment of CAD systems.

2 GI Endoscopy

To examine the esophagus, stomach, duodenum (upper GI), and the large bowel and rectum (lower GI), a long, flexible tube is inserted into the mouth and rectum, respectively. A tiny video camera at the tip of the tube allows the doctor to view inside of the GI tract in real-time, where findings, as depicted in Fig. 1a and b can be found.

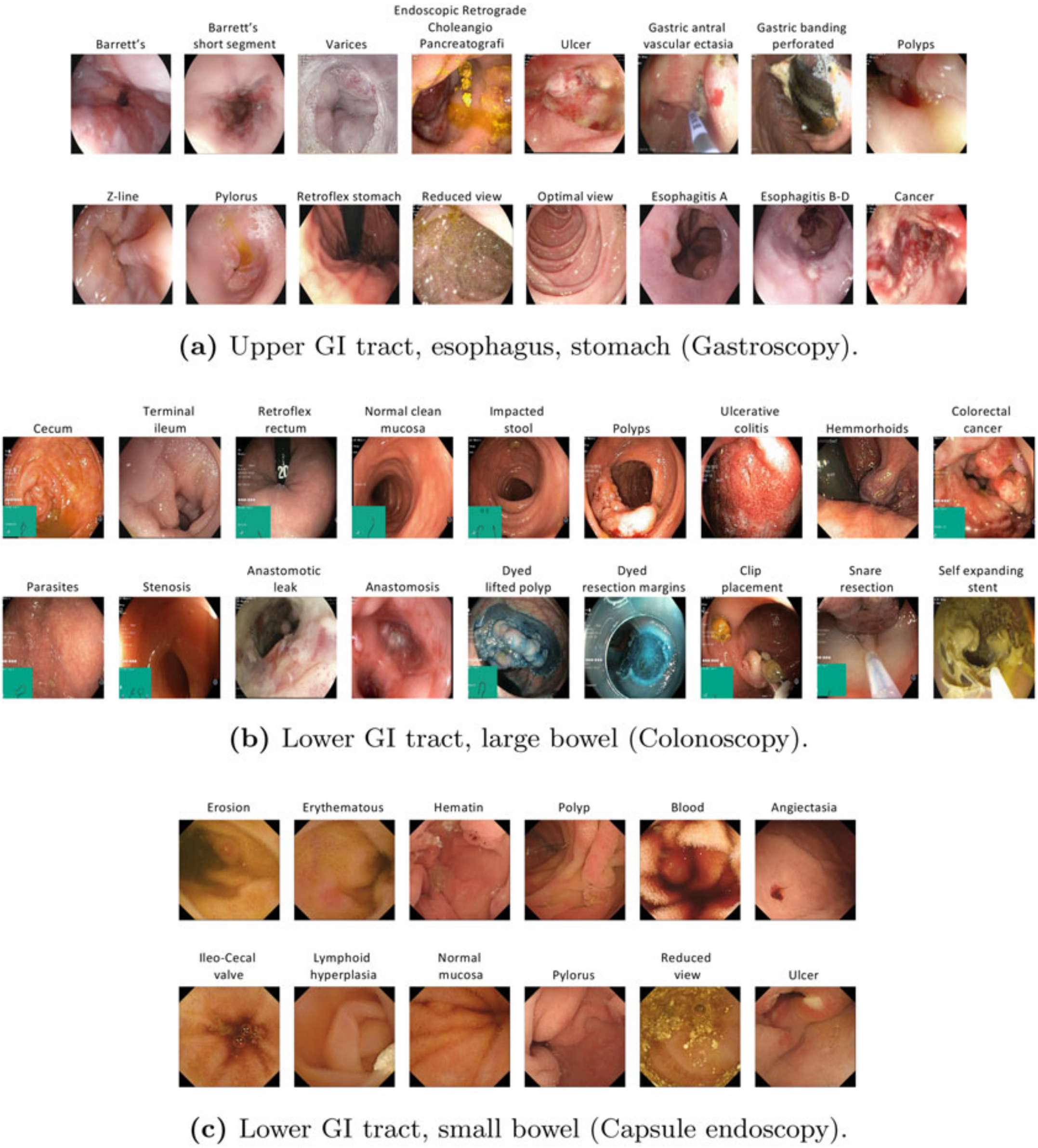
The small bowel is, due to its anatomical location, less accessible for inspection by such flexible endoscopes. To easier access these areas of the GI tract, Video Capsule Endoscopy (VCE) [22] has been introduced as an alternative examination method [25]. A VCE consists of a small capsule containing one or more wide-angle cameras. The capsule is swallowed by the patient, and it captures a video as it moves through the GI tract. The video is extracted, and a medical expert assesses it in a potentially tedious and time-consuming process after the procedure, searching for findings like the ones shown in Fig. 1c.

Even though these examination procedures allow clinicians to detect GI anomalies, there is still ample scope for improvements. Looking at the possible findings depicted in Fig. 1, it is obvious that it can be hard to detect and classify the various anomalies potentially found in the various parts of the GI tract, either live during a gastroscopy or colonoscopy, or in a post-analysis of the VCE video. Moreover, there are large operator variations and anomaly miss-rates reported for both regular endoscopies [34, 52, 64, 108] and capsule endoscopies [20, 88].

Hence, the hope is that automated analysis can *assist* medical experts in real-time anomaly detection, removing variations and increasing detection rates. Moreover, analyzing hours of VCE video, there is also a large potential in saving medical expert time, by analyzing the 4–12 h long videos in a few minutes by a fast computer, compared to the usual 45–60 min error-prone, fast-forward analysis performed by medical personnel today. From an analysis point of view, there are two important requirements for such CAD systems:

1. *High detection or segmentation performance* in the analysis is important in order to address the large human miss-rates and variabilities. It is often measured in terms of metrics like precision, sensitivity (recall), specificity, accuracy, F1 score, Matthews correlation coefficient (MCC) or similar [98]. This requirement aims at finding all anomalies correctly, i.e., detecting all findings without false positives or negatives. A more detailed discussion on metrics is given in Sect. 5.3.
2. An often neglected requirement is *fast processing* in order to give real-time feedback during the endoscopy examination, or in the case of VCE, higher scale of the analysis and a faster feedback on the same amount of processing resources.

Furthermore, in order to be deployable in a clinical environment, all components need to be integrated in a pipeline capturing videos or frames from the endoscopy equipment, via an automatic analysis, to give the clinicians a visual feedback (and potentially also assisting in generating an



(a) Upper GI tract, esophagus, stomach (Gastroscopy).

(b) Lower GI tract, large bowel (Colonoscopy).

(c) Lower GI tract, small bowel (Capsule endoscopy).

Fig. 1 Examples of various findings in the GI tract including anatomical landmarks, pathological findings, normal mucosa, therapeutic interventions and medical instruments [14, 96]

examination report according to medical standards). The system must also be easily integrated into and usable with the current examination procedures, and of course, the various components must meet the medical privacy and security regulations.

3 Existing Methods

As mentioned above, a large number of algorithms and models for automated analysis of GI video and images have already been proposed. In this respect, when we discuss CAD systems for the GI tract today, people often interchangeably

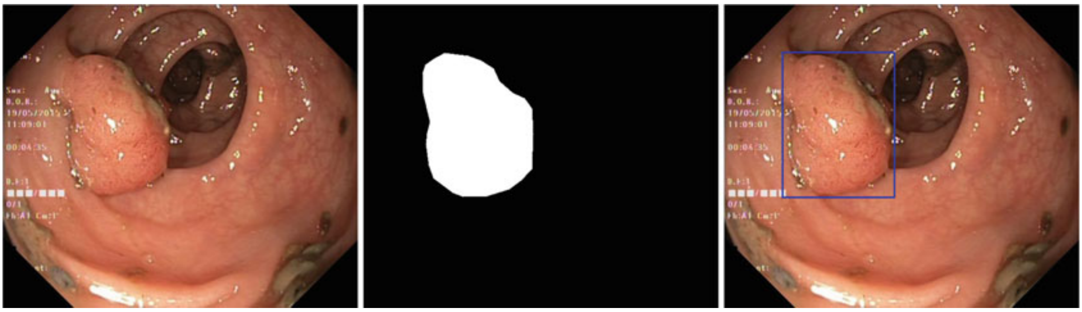


Fig. 2 Various ways of indicating a finding (*left*: detection "just" showing the image; *center*: full segmentation mask showing in white all pixels part of the finding; and *right*: bounding box making a rectangle around the finding)

talk about detection, localization and segmentation. Here, we therefore first try to distinguish between the terms as follows:

- *Detection* is the operation of detecting whether an image belongs to a certain classification or not. This can be a binary "yes or no" for questions whether the image or video frame contains a polyp or not. It also includes systems that *classifies* the input into multiple classes.
- *Localization* is to point into the image where the object is located, e.g., using some type of point markers or making a bounding box around the object of interest.
- *Segmentation* is yet another step further where one determines pixel-wise whether the pixel belongs to a finding or not, e.g., generating an exact segmentation mask of the finding.

Figure 2 shows an example of detection, localization and detection. As localization is often mixed into both detection and segmentation, we here focus on detection and segmentation.

3.1 Hand-Crafted-Feature-Based Approaches

Automatic detection of GI anomalies has been a topic of research long before the success of AI and deep neural networks, using what is nowadays often called traditional computer vision and ML methods, as found in libraries such as *OpenCV* [16] and *LIRE* [69]. Already in 1998, Krishnan et al. [59] proposed detecting polyps using shape-features in a curvature analysis. In the subsequent decade, various approaches using a mix of shape, edge, texture, and color features appeared. For example, Alexandre et al. [2] detected polyps using a support vector machine (SVM) on color patterns. Further, using SVMs, Ameling et al. [5] combined texture and colors, and Park et al. [75] used shape and texture features in a conditional random field classifier.

Two more recent approaches using hand-crafted techniques are Polyp-Alert [111] and EIR [85], where the authors also measured analysis time, with the goal of being able to give real-time feedback during the examination. The Polyp-Alert [111] system combines edge and texture features. The polyp edge detection algorithm mainly relies on edge features obtained from the part-based multi-derivative edge cross-section profile [110]. The EIR [85] system combines a content-based similarity search with statistical classifiers from the training data. A large number of image features are tested [87], ending up with a combination of the joint composite descriptor feature and the Tamura features, due to a good trade-off between the precision and sensitivity (recall), and the speed of the algorithm. A search-based classifier is then used to determine if an image contains a finding of a certain class.

A detailed overview containing earlier example approaches can be found in [85, 111]. However, lately, deep learning approaches have outperformed these hand-crafted approaches and replaced them entirely.

3.2 Deep Learning-Based Approaches

Already in 2001, Karkanis et al. [53] aimed for the detection of lesions in endoscopic video using textural descriptors on the wavelet domain supported by artificial neural network architectures, albeit not using deep architectures. Such early approaches were tested on tiny data sets, in this case 8 images [53]. More recent approaches are usually based on deep learning architectures where Convolutional Neural Networks (CNNs) are clearly the most popular ones.

Where hand-crafted features rely on extracting predefined properties of an image, such as color, texture, or shape, CNNs are neural network architectures using convolutions and pooling operations to automatically learn which features are most relevant. CNNs perform well on many different tasks like image classification, object detection in images, and image generation [56]. Although they are mostly used for image analysis, they have also proven useful in timeseries research and video analysis. In medicine, architectures like U-Net [89] have shown promising results in areas like cardiology, colonoscopy, and radiology [74, 119, 122]. This also includes gastroenterology, where CNNs are currently state-of-the-art for analyzing colonoscopy videos. The most common application is the detection and segmentation of polyps, where many CNN-based approaches have shown excellent results [17, 50, 114]. These approaches have expanded to other findings as well, like detecting and segmenting ulcers [31]. Furthermore, due to limited access to medical image and video data, most approaches use transfer learning. In transfer learning, pre-trained models are used as a starting point, and refined for the given data set by retraining with some layers trainable and some frozen [82].

An automated CAD system for the GI endoscopic image segmentation is a step further than providing “just” detection of anomalies. A predicted segmentation mask (see Fig. 2) can help point out the area of interest in the images (frames) that need to be further examined. However, making such perpixel predictions is also a more complex task. In this respect, there has been a considerable amount of work done so far, especially targeting polyps [32, 45, 47, 48, 50, 71, 81, 100, 109], artifacts [3], and endoscopic instruments [90]. In general, CNN-based approaches perform well with the larger polyps. However, still the major challenges issues in the field are related to adenomatous polyps or small and flat polyps. Recent studies are targeting smaller polyps [50, 63]; however, it is yet an open-challenge to solve.

3.3 Unsupervised and Semi-supervised Approaches

The above presented approaches fall into the category of supervised learning, meaning that we train the models on a data set with an existing ground truth. In this section, we give a glance at newly emerging unsupervised and semi-supervised methods.

Generative Adversarial Networks (GANs), which were introduced by Good-fellow et al. [30] in 2014, are becoming increasingly popular in the medical domain for generating synthetic data. Different advancements to the original GAN architecture, such as conditional GAN [72], pix2pix [43], CycleGAN [123], StyleGANs [54, 55], to mention a few, present different methods, ranging from domain transformation to high definition image generation. ML researchers in the medical domain can use GAN models to generate synthetic data to tackle challenges related to privacy, data deficiency, and data annotation. For example, Younghak et al. [93] use a conditional GAN architecture to generate synthetic polyp images to improve the performance of a deep learning system detecting polyps in the colon. This methodology is still in its early stages, and it has yet to be shown to which extent generated data can replace real data and help to improve performance and shareability.

Another emerging method in the field of medical image analysis, is semi-supervised learning. Here, the goal is to learn from a small set of labelled data combined with a larger amount of unlabeled data. Examples include [7, 67, 70, 116]. These models produce promising results, and could also help overcome the challenge of insufficient labelled data faced by many data-hungry methods. However, these approaches still struggle with challenges such as low accuracy and high entropy during early stages of the training process. The models are also regularized towards high entropy predictions, making it hard to achieve a high accuracy [117, 120]. It will be interesting to see whether these challenges can be overcome, and how useful the results will prove to be in the medical domain.

4 Example Results

High detection or segmentation rates are important in order to be clinically relevant, and the typical way the performance is compared. However, due to factors like different data sets and different equipment, the pure numbers cannot be directly compared. Still, to give some indications of the state-of-the-art performance, we give a set of, by far from complete, examples using standard metrics like precision, sensitivity (recall), specificity, accuracy, F1 score and MCC for detection; and Dice similarity coefficient (DSC), Intersection over Union (IoU), precision and sensitivity for segmentation. A substantial overview of existing approaches can be found in [61], containing 138 different studies. An explanation of the different metrics is given in Table 1 and further discussed in Sect. 5.3. Another source for exploring and comparing different approaches are the popular GI detection, classification and segmentation challenges discussed in Sect. 5.6.

A selection of performance examples are given in Table 2. Looking at the numbers, we see that in the specific tested cases, the computer should be at the level of the best experts with scores above 90%, i.e., potentially being a helpful digital assistant during a GI endoscopy examination. Likewise, example results for lesion segmentation are provided in Table 3, and the numbers are again encouraging in terms of proving that the used models could be of use in a medical setting. However, while the results achieved are promising, there are still several open challenges, including generalizability, overfitting, cross data set testing and explainability of the results. Moreover, as indicated in the Tables, hardly any existing research report the speed of the system, meaning that it is hard to assess the system's capability to provide a live analysis in the clinic.

Table 1 List of commonly used metrics. To define each metric, TP, FP, TN, and FN represent true positives, false positives, true negatives, and false negatives, respectively

Formula	Description
accuracy $= \frac{TP+TN}{TP+FP+TN+FthickmathspaceN}$	Rate of correct classification. Ratio between correctly classified samples and all samples.
precision $= \frac{TP}{TP+FthickmathspaceP}$	Proportion of retrieved samples which are relevant. Ratio between correctly classified positive samples and all samples classified as positive.
sensitivity (also known as recall) $= \frac{TP}{TN+FP}$	Proportion of relevant samples which are retrieved. Ratio between correctly classified positive samples and all positive samples.
specificity $= \frac{TN}{TN+FP}$	Negative class sensitivity. Ratio between correctly classified negative samples and all negative samples.
F1 $= \frac{2\times TP}{2\times TP+FP+FN}$	Harmonic mean of the precision and sensitivity (recall).
MCC $= \frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)d(TP+FP+FN)}}$	Pearson's correlation coefficient [23] for binary classification.
IoU (also known as Jaccard) $= \frac{TP}{TP+FP+FN}$	Similarity between sets from the size of the intersection divided by the size of the union.
DSC $= \frac{2TP}{2TP+FP+FN}$	Quotient of similarity of two sets. Semi-metric as it doesn't satisfy the triangle inequality. Related to the IoU via $\frac{S}{2-S}$.

Table 2 Examples of **detection** performance from different approaches. The results show promising performance with numbers above 90%. Unfortunately, speed is not commonly reported

Paper/ system	Data set used	Sensitivity (recall)	Specificity	Accuracy	Precision	F1	MCC	Speed (fps)
Boughorbel [17]	MICCAI-challenge data sets	86.3	–	–	73.6	–	–	–
Kundu [60]	30 Own data set	95.2	98.3	97.9	88.4	–	–	–
Cho [21]	Seoul National University Hospital	>87	–	>93	–	–	–	
Ghosh [29]	VCE videos data set	99.4	99.2	97.9	95.8	–	–	
Bell [8]	CTC generating 4000 images per patients	89.8	75.5	–	–	–	–	
Pogorelov [76]	Kvasir	83.9	98.5	97.2	84.1	85.6	82.8	46
Billah [13]	Colonoscopy & Endoscopy vision data set	98.7	98.2	98.3	–			
Thambawita [99]	Kvasir	95.8	99.7	95.8	95.9	95.8	95.3	29

Table 3 Some examples of different **segmentation** approaches applied to different data sets. We can clearly see that the performance overall is quite promising (with all metrics in the range of 70 to 95). Speed is unfortunately not commonly reported

Pape/system	Data set used	DSC	IoU (Jaccard)	Sensitivity (recall)	Precision	Speed (fps)
U-Net [89]	MICCAI-PhC-U373	–	92.0	–	–	–
PraNet [26]	CVC-ClinicDB	89.9	84.0	–	–	–
PolypSegNet [71]	CVC-ClinicDB	91.5	86.2	91.1	96.2	–
ResUNet++ [50]	CVC-ClinicDB	79.6	79.6	70.2	87.9	–
PraNet [26]	Kvasir-SEG	89.8	84.0	–	–	–
PolypSegNet [71]	Kvasir-SEG	88.7	82.5	84.5	91.7	–
ResUNet++ [50]	Kvasir-SEG	81.3	79.3	70.6	87.7	–
Double-UNet [45]	CVC-ClinicDB	92.4	86.1	84.6	96.0	–

5 Open Issues and Ongoing Research

Despite impressive results presented in many of the published papers, even exceeding what are reported as average detection rates from clinicians, there are still challenges and open issues. First, for example, Thambawita et al. [98] presented the issue of overfitting to specific data sets and a lack of generalizability. This means that a model that performs well on one data set may not perform at all on another. Furthermore, like other deep neural networks, CNNs are black boxes, and it is not easy to understand why one input gives a particular result. There is also a lack of large open data sets that contain annotations for uncommon abnormalities and rarely documented findings to support data-hungry algorithms like CNNs. Here, we elaborate on a few of these open issues.

5.1 Limited Data Availability

Available medical data is scarce. However, modern deep learning approaches usually require a lot of data to perform well, and often, the more variations in the data, the better the model gets, especially for supervised learning models. Table 4 shows the data sets available in the field of GI endoscopy. Evidently, the number of images used for training and testing is small when compared to the data set from the natural images. This is because it is difficult to obtain data from the medical domain. The data is often protected and unavailable due to legal restrictions and lack of medical personnel for the tedious process of manually extracting and labeling training data. This calls either for better data sharing processes and culture, or methods more capable of handling small amounts of data.

This gives rise to several basic challenges: The amount of data is too small to train a robust model, and the presented results might appear deceivingly good due to overfitting. Moreover, it is hard to compare results if all experiments are performed on different data, and practically impossible to reproduce them. Thus, it is almost impossible to conclude whether one model is better than another. We must therefore aim for more and open data sets. Table 4 contains an overview of know available data sets at the time of writing, making a good starting point for future experiments. Still, more data is needed, especially data containing pathological outcomes.

5.2 Generalizability

One of the open issues in the field is the GI endoscopy is the generalizability of ML models, i.e., their ability to perform well on previously unseen data regardless of source, equipment, etc. Such data can be from either the same distribution as the model was trained on, or from a different distribution. Which of the two a new data sample represents, is not always clear [101, 103]. Although some recent studies address generalizability of ML models for polyp classification [45, 112], this must be addressed for any model or system to be deployed into clinical practice.

Evaluating whether a model is reliable for real world use also requires cross data set testing, to avoid accepting a model which coincidentally works well on one specific set of data. The model developers should in general not have access to the final test data, to avoid bias during testing and development. This process, known as data blinding, is an important tool in many fields of research, including medicine [80]. Ideally, the model should be tested for robustness on data collected separately from the data used during model development and testing.

Furthermore, distinction should be made between data annotated by medical experts, referred to as *soft ground truth*, and data labelled based on a medical test, referred to as *hard ground truth*, e.g., pathological examination of a polyp. The quality of soft ground truth data is limited by how well the medical annotator is trained, and such data is most useful for training models intended to automate processes. On the other hand, data with hard ground truth labels can also be used for automating processes, with the added benefit of avoiding annotator error or bias into the model, but it can furthermore be used for obtaining new knowledge. Note that while, as mentioned above, annotating each image is time consuming, collecting hard ground truth data is even more demanding, resulting in a scarcity of such data sets.

In current endoscopy practices, different hospitals use different endoscope system for diagnosis and therapy. The most common globally available endoscope systems are Olympus (Japan), Pentax 90i series (Japan), Fujinon (Japan), and Karl Storz (Germany) [57]. Moreover, different medical institutes have different protocols. Therefore, designing generalizable CAD systems is essential for performing well on a variety of institutes. Such systems should always be tested on several data sets. Discussions regarding challenges and advantages associated with cross-dataset testing can be found in [98].

5.3 Metrics and Evaluation

Evaluating performance is an important step when creating models for clinical use, and depends strongly on the choice of metric. As shown in

Table 4 Summary of available endoscopic data sets. A further discussion about data sets are found in [14, 96]

Data set	Findings	Location	#Images	#Videos	Bounding box?	Segmentation mask?	Input size	VCE data?	Endoscopic device
Kvasir [77]	various	⇑⇓	8,000	–	–	–	variable	–	†
Nerthus [78]	stool (cleanness)	⇓	5,525	–	–	–	720 × 576	–	†
HyperKvasir [14]	various	⇑⇓	110,079	373	–	–	variable	–	†
KvasirInstrument [46]	instruments	⇓	590	–	✓	✓	variable	–	†
Kvasir-SEG [49]	polyps	⇓	1,000	–	✓	✓	variable	–	†
ASU-Mayo [97]	polyps	⇓	18,781	–	–	–	variable	–	–
CVC-ClinicDB [10]	polyps	⇓	612	–	–	–	384 × 288	–	‡∘
CVC-ColonDB [11]	polyps	⇓	380	–	–	–	574 × 500	–	–
ETIS Larib Polyp DB [95]	polyps	⇓	196	–	–	–	1225 × 966	–	‡∘
SUN Colonoscopy Video DB [73]	polyps	⇓	158,690	–	✓	–	1080 × 1240	–	?
CVCVideoClinicDB [6, 12]	polyps	⇓	11,954	–	–	–	384 × 288	–	–
CAD-CAP [65]	various		25,000	–	–	–	–	✓	–
KID [58]	various		2,371	47	–	–	–	✓	–
KvasirCapsule [96]	various	⇓	4,820,739	118	✓	–	variable	✓	•

Location: ⇑ = upper GI ⇓ = lower GI
Device: † = ScopeGuide, Olympus ‡ = Olympus Q160ALandQ165L
∘ = Exera IIvideoprocessor • = Olympus EC-S10 endocapsule

Table 1, commonly used metrics are precision, sensitivity (recall), specificity, accuracy and F1 score. Some papers also report AUROC (area under the receiver operating characteristics). There are several reasons for going beyond the aforementioned metrics [98]. One challenge frequently encountered in association with medical data sets, is their tendency to be imbalanced between classes, often having far more normal images than images with lesions. Because of this, certain metrics can provide an overly optimistic impression of the actual performance. For instance, a binary classifier can achieve a high accuracy on a data set containing few negative instances, by assigning all instances to the positive class. The AUROC is also known to be deceptive for imbalanced classification [91]. In such cases, the correlation coefficient between the true and predicted classes can be more informative [15], although no single metric is universally informative or suited for any imbalanced data problem. Moreover, for detection purposes, it is also a question whether one report per-frame performance, i.e., giving a decision for every frame in the video, or per-lesion, i.e., giving a correct prediction for at least one of the frames in the video sequence. Looking at the results from a technical point of view, a per-frame analysis of often desired, but from the medical point of view, a per-lesion analysis is often sufficient to notify the clinician of the finding once.

For segmentation performance, commonly used metrics are DSC and the IoU, also known as the Jaccard index. In clinical use, medical experts are usually interested in pixel-wise detail information about the potential lesion. DSC and IoU can be used to compare the pixel-wise similarity between the predicted segmentation maps and the ground truth. In addition, precision and sensitivity are used to evaluate under-segmentation or over-segmentation, where under-segmentation implies that the model predicts less relevant content in some portion of the image compared to the ground truth, and over-segmentation that the predicted image covers more pixels than the ground truth.

As observed in Tables 2 and 3, little research has until now focused on the required real-time capabilities in order to provide live feedback to clinicians during the endoscopy examinations. However, there seems to be reported systems that analyze data faster than the frame-rate threshold, and it has also been given attention in some of the arranged competitions (see Sect. 5.6). Nonetheless, it is often a trade-off between speed (model complexity) and detection performance, indicating that this is still an important issue in future research and development of CAD systems.

5.4 Automatic Report Generation

After the endoscopist finishes an endoscopy, a high-quality report should be generated. This often a time-consuming process, where research shows that approximately one-sixth of U.S. physicians working time is spent on administrative tasks, taking time away from direct-patient care and lessening job satisfaction [115]. Moreover, there are large variations in endoscopists' interpretations of findings as well as reporting styles. This can, and often does, lead to inconsistencies in the final decision [37]. Hence, automated report generation could both save clinical time and help standardize endoscopy reports, and recent development in natural language processing is expected to open up new possibilities in automatic report generation [86].

A method proposed by Jing et al. [51] uses neural image captioning to create reports from x-ray images. In [121], images are analyzed by a neural network, and example images of findings similar to the one at hand and attention maps are combined to reports. Most approaches focus on image analysis as a basis, and combine this with additional information [24, 33, 118]. This of course depends on access to a database containing correct information which can be used in combination with the images. A significant challenge is different reporting standards between countries or even hospitals, making it practically impossible to create a widely adoptable software.

However, for medical experts, automatic text creation might not even be the most crucial feature of such a software: A more important aspect is their ability to understand the reasoning and decision of the underlying model, enabling them to include it in their assessment. This is discussed in the next section.

5.5 Explainability

A well-known challenge associated with deep learning based CAD systems, is limited explainability due to their inherent complexity [4]. This property has caused their notoriety as black boxes whose decision-making processes are unknown, especially to end-users [35]. The need for understanding and explaining how the systems work and which roles the different data features play in the decisions, addresses different needs in the different stages of the system's development and use. The developer of the system needs to understand how data and methods are working together, as understanding and interpretability of the output helps to determine errors in the data as well as enabling targeted failure analysis. Particularly, in the context of this AIM, the medical experts require an explanation of the system's decision to assure that it concurs with the relevant medical knowledge.

Deep learning based systems, such as CNNs, have no inherent ways of providing explanations, meaning that they must either be extended to contain explanation generators, or explanations must be obtained post hoc [1, 38, 39]. A brief overview of approaches to model explanations is shown in Table 5. Models can be designed to provide justifications for their decisions as an additional task, e.g., via a text justification generator as part of the model architecture [62]. Given a model without such a design, different approaches are available: Those which explain the properties of the decision making system itself, and those which treat the system like a black box and provide explanations based on its emergent behavior, referred to as model dependent and model agnostic approaches, respectively. One example of the former is displaying the values of the Deep Neural Network (DNN)'s internal parameters as a heat map superimposed on the classification instance [35]. Interpreted correctly, this can provide an understanding of the system's internal decision making process. Such an approach can also be extended to include information regarding the system induced decision uncertainty (meaning the part of the uncertainty not associated with the data collection and selection process), see [113].

Among the model-agnostic methods, the explanation concept LIME (*L*ocally *I*nterpretable *M*odel-agnostic *E*xplanations) approximates the black-box model using an interpretable model, such as a linear model, decision tree, or falling rule list [84]. This is done in the neighborhood of the instance to explain, making the resulting explanation a local one, given that it applies to a single outcome and is based on the particular instance's characteristics, as is also the case for the aforementioned model-dependent explanations.

In contrast, global explanations capture and explain the model at large, such as feature importance ranking. One class of methods capable of

Table 5 The different model explanation approaches regarding when they are applied: During the model development (in-model) or after the model is finished (post-model). Explanation methods provide insight into model behavior either locally (around a particular prediction) or globally

Category		Description	Ex.
In-model		Justification text generator as part of model architecture	[62]
Post-model	Model dependent	GradCam: Display DNN activations on image	[35]
	Model agnostic	LIME: Yields a locally interpretable model approximating the full model	[84]
		SHAP: Shapley decomposition of a conditional expectation function of the full model	[68]
	Model independent	Global non-parametric Shapley decomposition	[28]

producing global explanations, are those based on the game-theoretic concept of Shapley values [92], which are currently enjoying a surge of interest in the statistics and machine learning literature [27, 40, 44, 66]. Shapley values are obtained by evaluating the model using all possible combinations of the data features. Hence, the computational complexity increases with the number of features $|f|$ as $2^{|f|}$, and the calculation involves re-training the model for each subset of features. The latter is problematic as re-training would result in different model parameters, highlighting that Shapley values are merely model agnostic, not *independent*. The widely used SHAP (*SH*apley *A*dditive ex*P*lanations) package [68] circumvents these challenges in different ways for various model architectures, by calculating approximate values using background samples from the data, and for deep architectures using a similar approach as the per node attribution rules from DeepLIFT [94]. The Shapley decomposition can be computed both globally and locally, and can be formulated [68] as a special case of LIME. Shapley values can also be used to obtain model-independent explanations [28].

5.6 Competitions and Challenges

There have been a series of different challenges related to automatic analysis of endoscopy data [9, 36, 79], where CNN-based approaches have been the top performing methods for the last few years. The various tasks given have been to benchmark and develop automated systems to accurately detect, localize, and segment the abnormalities inside the GI tract. These challenges targeted different tasks from detection, localization, and segmentation of GI anomalies, colorectal polyps to artifacts presence in the GI tract (see Table 6). These regular competitions can help the research community in the field to find to find common standards for evaluating models, benchmarking state-of-the-art methods and tools, and finding new directions to bring the field forward together.

5.7 Clinical Verification and Emerging Commercial Systems

Many research groups have presented promising research results and good performance indicators, and several AI-based commercial systems have emerged, some of which are listed in Table 7. The status of these are mostly unknown, but, for example, the GI Genius system is CE marked, but still lacks US Food and Drug Administration (FDA) approval, and EndoBRAIN-EYE is approved only in Japan. For CAD systems to be deployed for real-time examinations in clinical examination rooms, or to be used for VCE data post analysis, clinical verification is strictly necessary. Still, at the time of writing, such studies are very limited. In August 2020, Repici et al. [83] presented a randomized multi-center trial, concluding that the AI-based CAD increases the adenoma detection rate (ADR), i.e., the percentage of

Table 6 List of GI detection, classification and segmentation challenge examples

Challenge name	URL
MICCAI 2015 Endoscopic Vision	https://polyp.grand-challenge.org/databases/
Medico 2017	http://www.multimediaeval.org/mediaeval2017/medico/
Medico 2018	http://www.multimediaeval.org/mediaeval2018/medico/
GIANA 2018	https://giana.grand-challenge.org/Home/
EAD 2019	https://ead2019.grand-challenge.org/
Biomedia 2019	https://github.com/kelkalot/biomedia-2019
Medico 2020	https://multimediaeval.github.io/editions/2020/tasks/medico/
EndoTect 2020	https://github.com/simula/icpr-endotect-2020
EDD Challenge 2020	https://edd2020.grand-challenge.org/
EndoCV 2020	https://endocv.grand-challenge.org

Table 7 Emerging commercial products

Product	Vendor	Year	URL
GI Genius AI	Medtronic/Cosmo Pharma	2019	https://www.cosmopharma.com/products/gi-genius
EndoBRAIN-EYE	Cybernet	2020	https://www.cybernet.jp/english/documents/pdf/news/press/2020/20200129.pdf
CAD-Eye	Fujifilm	2020	https://www.fujifilm.eu/eu/cadeye
Ai4Gi	Ai4Gi	2016	https://ai4gi.com
UltiVision	DocBot	2018	https://www.docbot.co/gastroenterology-and-health
DISCOVERY	Pentax	2020	https://www.pentaxmedical.com/pentax/en/95/2/DISCOVERY-new
ENDO-AID	Olympus	2020	https://www.olympus.no/medical/en/Products-and-Solutions/Products/Product/ENDO-AID.html
SOMA	Augere Medical	2018	https://augere.md

patients with at least one histologically proven adenoma or carcinoma, demonstrating the potential of such systems. They examined 685 patients: 341 patients using the CAD system and 344 patients using only the traditional manual examination. The system achieved an ADR of 54.8%, and the control group 40.4%. This demonstrates that AI-based systems can help detect adenomas, but that further improvements are required to increase detection rates, and to detect a larger number of sessile serrated lesions (at all). Considering the limitations of the study as well as the presented performance, it is clear that there are still improvements to made, and more clinical studies are in order.

Despite significant interest from the industry, proper standards regarding evaluation methods and reproducibility are widely lacking. In addition, industry applications seem not to have focused on model explainability or model output interpretability. These are all crucial ingredients of trustworthy applications, and industry development will hopefully follow current research trends and focus more on these in the future.

Finally, when a high-performing (research) prototype has been built and tested, meeting the requirements above, it must be approved for medical use. Robust evaluation of AI based software before implementation is needed to reduce patient and health system risk, establish trust to facilitate wide-spread adoption. The common term used for such products is AI based software as a medical device (SaMD). Regulators of the SaMD applications, including the FDA in the United States, have been guided by the Global Harmonization Task Force and International Medical Device Regulators Forum (IMDRF). The IMDRF has proposed four different risk categories for SaMD each with a different set of requirements for assessing scientific and clinical validity of the technology [42]. Within gastroenterology, CADe and CADx technologies have not yet been classified. The current FDA process for SaMD is derived from its approval process for medical devices and will be categorized into three risk categories: Classes I, II, and III (highest risk) [105]. After risk classification, premarket submission as a 510(k) pathway or de novo pathway might be relevant to GI-based AI technologies similar to Osteoidetect [107]. Moreover, given that the AI algorithms are rapidly iterative and continuously learning, it can pose a challenge to the current regulatory process. The FDA proposed a new system of regulation for AI technologies in its Digital Health Innovation Action Plan, focused on AI technologies that rely on continuous learning and adaptation [106]. Regulators around the world have also recognized the challenges involved with AI algorithms when applied to medicine and most countries have initiated efforts to develop policies tailored for SaMD. Many of them share the core principles of designation of risk, review clinical evidence to demonstrate efficacy and safety, practices to incorporate evolving AI systems.

6 Summary and Conclusions

In this work, we have introduced the application of automated data analysis for GI endoscopy, and presented an overview on detection and segmentation based approaches to tackle challenges like large lesion miss-rates and interobserver variability. Recent studies have shown that deep computer vision-based approaches seem to have the potential of improving the accuracy and overall performance in GI endoscopy by providing fully automated CAD systems acting as an additional digital eye. Nevertheless, there are still several open issues and challenges which need to be addressed before automatic analyses can be usefully integrated into clinical practice. These should be regarded as issues requiring research attention in the field.

Acknowledgments This work is funded in part by the Research Council of Norway, project number 282315 (AutoCap).

References

1. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access. 2018;6:52138–60.
2. Alexandre LA, Casteleiro J, Nobreinst N. Polyp detection in endoscopic video using svms. In: Proceeding of Knowledge Discovery in Databases (PKDD). Berlin/Heidelberg: Springer; 2007. p. 358–65.
3. Ali S, Zhou F, Braden B, Bailey A, Yang S, Cheng G, Zhang P, Li X, Kayser M, Soberanis-Mukul R, Albarqouni S, Wang X, Wang C, Watanabe S, Oksuz I, Ning Q, Yang S, Khan MA, Gao X, Rittscher J. An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. Sci Rep. 2020;10:2748.
4. Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Precise4Q consortium: explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak. 2020;20(1):310. https://europepmc.org/articles/PMC7706019
5. Ameling S, Wirth S, Paulus D, Lacey G, Vilarino F. Texture-based polyp detection in colonoscopy. In: Meinzer HP, Deserno TM, Handels H, Tolxdorff T (eds) Bildverarbeitung für die Medizin 2009. Informatik aktuell. Springer, Berlin, Heidelberg; 2009. p. 346–50.
6. Angermann Q, Bernal J, Sánchez-Montes C, Hammami M, Fernández-Esparrach G, Dray X, Romain O, Sánchez FJ, Histace A. Towards real-time polyp detection in colonoscopy videos: adapting still frame-based methodologies for video sequences analysis. In: Cardoso MJ, Arbel T, Luo X, Wesarg S, Reichl T, González Ballester MÁ, McLeod J, Drechsler K, Peters T, Erdt M, Mori K, Linguraru MG, Uhl A, Oyarzun Laura C, Shekhar R, editors. Computer assisted and robotic endoscopy and clinical image-based procedures. Cham: Springer International Publishing; 2017. p. 29–41.
7. Bai W, Oktay O, Sinclair M, Suzuki H, Rajchl M, Tarroni G, Glocker B, King A, Matthews PM, Rueckert D. Semi-supervised learning for network-based cardiac MR image segmentation. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins DL, Duchesne S, editors. Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI). Lecture Notes in Computer Science. Springer International Publishing; 2017. p. 253–60.
8. Bell LT, Gandhi S. A comparison of computer-assisted detection (CAD) programs for the identification of colorectal polyps: performance and sensitivity analysis, current limitations and practical tips for radiologists. Clin Radiol. 2018;73:593.e11–8. https://doi.org/10.1016/j.crad.2018.02.009.
9. Bernal J, Tajkbaksh N, Snchez FJ, Matuszewski BJ, Chen H, Yu L, Angermann Q, Romain O, Rustad B, Balasingham I, Pogorelov K, Choi S, Debard Q, Maier-Hein L, Speidel S, Stoyanov D, Brandao P, Crdova H, Snchez-Montes C, Gurudu SR, Fernndez-Esparrach G, Dray X, Liang J, Histace A. Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. IEEE Trans Med Imaging. 2017;36(6):1231–49.
10. Bernal J, Sánchez FJ, Fernández-Esparrach G, Gil D, Rodríguez C, Vilariño F. Wm-dova maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. Comput Med Imaging Graph. 2015;43:99–111.
11. Bernal J, Sánchez J, Vilarino F. Towards automatic polyp detection with a polyp appearance model. Pattern Recogn. 2012;45(9):3166–82.
12. Bernal JJ, Histace A, Masana M, Angermann Q, Sánchez-Montes C, Rodriguez C, Hammami M, Garcia-Rodriguez A, Córdova H, Romain O, Fernández-Esparrach G, Dray X, Sanchez J. Polyp detection benchmark in colonoscopy videos using GTCreator: a novel fully configurable tool for easy and fast annotation of image databases. In: Proceedings of 32nd CARS conference. Berlin; 2018.
13. Billah M, Waheed S. Gastrointestinal polyp detection in endoscopic images using an improved feature extraction method. Biomed Eng Lett. 2018;8(1):69–75.
14. Borgli H, Thambawita V, Smedsrud PH, Hicks S, Jha D, Eskeland SL, Randel KR, Pogorelov K, Lux M, Nguyen DTD, Johansen D, Griwodz C,

Stensland HK, Garcia-Ceja E, Schmidt PT, Hammer HL, Riegler MA, Halvorsen P, de Lange T. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Sci Data. 2020;7:283. https://doi.org/10.1038/s41597-020-00622-y. Springer Nature

15. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews correlation coefficient metric. PLoS One. 2017;12(6):e0177678.
16. Bradski G. The OpenCV library. Dr Dobb's J Softw Tools. 2000;120:122.
17. Brandao P, Mazomenos E, Ciuti G, Caliò R, Bianchi F, Menciassi A, Dario P, Koulaouzidis A, Arezzo A, Stoyanov D. Fully convolutional neural networks for polyp segmentation in colonoscopy. In: Medical imaging 2017: computer-aided diagnosis, vol. 10134. International Society for Optics and Photonics; 2017. p. 101340F. https://doi.org/10.1117/12.2254361
18. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394–424.
19. Brenner H, Kloor M, Pox CP. Colorectal cancer. Lancet. 2014;383(9927):1490–502.
20. Cave DR, Hakimian S, Patel K. Current controversies concerning capsule endoscopy. Dig Dis Sci. 2019;64 (11):3040–7.
21. Cho M, Kim JH, Kong HJ, Hong KS, Kim S. A novel summary report of colonoscopy: timeline visualization providing meaningful colonoscopy video information. Int J Color Dis. 2018;33(5):549–59.
22. Costamagna G, Shah SK, Riccioni ME, Foschia F, Mutignani M, Perri V, Vecchioli A, Brizi MG, Picciocchi A, Marano P. A prospective trial comparing small bowel radiographs and video capsule endoscopy for suspected small bowel disease. Gastroenterology. 2002;123(4):999–1005.
23. Cramer H. Mathematical methods of statistics. Princeton: Princeton University Press; 1946.
24. Daniels ZA, Metaxas DN. Exploiting visual and report-based information for chest x-ray analysis by jointly learning visual classifiers and topic models. In: Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI). IEEE; 2019. p. 1270–4.
25. Enns RA, Hookey L, Armstrong D, Bernstein CN, Heitman SJ, Teshima C, Leontiadis GI, Tse F, Sadowski D. Clinical practice guidelines for the use of video capsule endoscopy. Gastroenterology. 2017;152(3):497–514.
26. Fan DP, Ji GP, Zhou T, Chen G, Fu H, Shen J, Shao L. PraNet: Parallel reverse attention network for polyp segmentation. arXiv preprint arXiv:2006.11392. 2020.
27. Frye C, Rowat C, Feige I. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. 2020.
28. Fryer D, Strümke I, Nguyen H. Explaining the data or explaining a model? Shapley values that uncover non-linear dependencies. arXiv:abs/2007.06011. 2020.
29. Ghosh T, Fattah SA, Wahid KA. CHOBS: Color Histogram of Block Statistics for automatic bleeding detection in wireless capsule endoscopy video. IEEE J Transl Eng Health Med. 2018;6(May 2017):1800112.
30. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Advances in neural information processing systems. Montreal, Canada. p. 2672–2680 (y).
31. Goyal M, Yap MH, Reeves ND, Rajbhandari S, Spragg J. Fully convolutional networks for diabetic foot ulcer segmentation. In: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC); 2017. p. 618–23.
32. Guo YB, Matuszewski B. GIANA Polyp segmentation with fully convolutional dilation neural networks. In: Proceedings of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications; 2019. p. 632–41.
33. Han Z, Wei B, Leung S, Chung J, Li S. Towards automatic report generation in spine radiology using weakly supervised framework. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2018. p. 185–93.
34. Hewett DG, Kahi CJ, Rex DK. Efficacy and effectiveness of colonoscopy: how do we bridge the gap? Gastrointest Endosc Clin. 2010;20(4):673–84.
35. Hicks S, Riegler M, Pogorelov K, Anonsen KV, de Lange T, Johansen D, Jeppsson M, Ranheim Randel K, Losada Eskeland S, Halvorsen P. Dissecting deep neural networks for better medical image classification and classification understanding. In: Proceedings of IEEE International Symposium on Computer-Based Medical Systems (CBMS); 2018. p. 363–8.
36. Hicks S, Petlund A, de Lange T, Schmidt P, Halvorsen P, Riegler M, Smedsrud P, Haugen T, Randel K, Pogorelov K, Stensland H, Dang Nguyen DT, Lux M. Acm multimedia biomedia 2019 grand challenge overview. In: Proceedings of the ACM International Conference on Multimedia (ACM MM); 2019. p. 2563–7.
37. Hicks S, Smedsrud P, Riegler M, de Lange T, Petlund A, Eskeland S, Pogorelov K, Schmidt P, Halvorsen P. Deep learning for automatic generation of endoscopy reports. Gastrointest Endosc. 2019;89: AB77.
38. Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable ai systems for the medical domain? arXiv preprint arXiv:1712.09923. 2017.
39. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial

intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Disc. 2019;9(4):e1312.

40. Huettner F, Sunder M. Axiomatic arguments for decomposiing goodness of fit according to Shapley and Owen values. Electron J Stat. 2012;6:1239–50.
41. International Agency for Research on Cancer, World Health Organization: Cancer Fact Sheets. 2020. https://gco.iarc.fr/today/fact-sheets-cancers
42. International Medical Device Regulators Forum (IMDRF): Software as a Medical Device (SaMD): key definitions. 2013. http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-131209-samd-key-definitions-140901.pdf
43. Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 1125–34.
44. Israeli O. A Shapley-based decomposition of the R-square of a linear regression. J Econ Inequal. 2007;5:199–212.
45. Jha D, Riegler MA, Johansen D, Halvorsen P, Johansen HD. Doubleu-net: a deep convolutional neural network for medical image segmentation. In: Proceedings of the IEEE International Symposium on Computer-Based Medical Systems (CBMS); 2020. p. 558–64.
46. Jha D, Ali S, Emanuelsen K, Hicks SA, Thambawita V, Garcia-Ceja E, Riegler MA, de Lange T, Schmidt PT, Johansen HD, Johansen D, Halvorsen P. Kvasir-instrument: diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. 2020.
47. Jha D, Ali S, Johansen HD, Johansen D, Rittscher J, Riegler MA, Halvorsen P. Real-time polyp detection, localisation and segmentation in colonoscopy using deep learning. arXiv preprint arXiv:2006.11392. 2020.
48. Jha D, Hicks SA, Emanuelsen K, Johansen HD, Johansen D, de Lange T, Riegler MA, Halvorsen P. Medico multimedia task at mediaeval 2020:automatic polyp segmentation. In: Proceedings of the MediaEval 2020 Workshop; 2020.
49. Jha D, Smedsrud PH, Riegler MA, Halvorsen P, de Lange T, Johansen D, Johansen HD. Kvasir-SEG: a segmented polyp dataset. In: Proceedings of the International Conference on Multimedia Modeling (MMM); 2020. p. 451–62.
50. Jha D, Smedsrud PH, Riegler MA, Johansen D, De Lange T, Halvorsen P, Johansen HD. ResUNet++: an advanced architecture for medical image segmentation. In: Proceedings of International Symposium on Multimedia (ISM); 2019. p. 225–2255.
51. Jing B, Xie P, Xing E. On the automatic generation of medical imaging reports. arXiv preprint arXiv:1711.08195. 2017.
52. Kaminski MF, Regula J, Kraszewska E, Polkowski M, Wojciechowska U, Didkowska J, Zwierko M, Rupinski M, Nowacki MP, Butruk E. Quality indicators for colonoscopy and the risk of interval cancer. N Engl J Med. 2010;362(19):1795–803.
53. Karkanis SA, Iakovidis DK, Karras DA, Maroulis DE. Detection of lesions in endoscopic video using textural descriptors on wavelet domain supported by artificial neural network architectures. In: Proceedings the International Conference on Image Processing; 2001. p. 833–6.
54. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2019. p. 4401–10.
55. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and improving the image quality of StyleGAN. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2020.
56. Khan A, Sohail A, Zahoora U, Qureshi AS. A survey of the recent architectures of deep convolutional neural networks. Artif Intell Rev. 2020;53(8):5455–516.
57. Ko WJ, An P, Ko KH, Hahm KB, Hong SP, Cho JY. Image quality analysis of various gastrointestinal endoscopes: why image quality is a prerequisite for proper diagnostic and therapeutic endoscopy. Clin Endosc. 2015;48(5):374.
58. Koulaouzidis A, Iakovidis DK, Yung DE, Rondonotti E, Kopylov U, Plevris JN, Toth E, Eliakim A, Johansson GW, Marlicz W, Mavrogenis G, Nemeth A, Thorlacius H, Tontini GE. Kid project: an internet-based digital video atlas of capsule endoscopy for research purposes. Endosc Int Open. 2017;5(6):E477–83.
59. Krishnan SM, Yang X, Chan KL, Kumar S, Goh PMY. Intestinal abnormality detection from endoscopic images. In: Proceedings of the IEEE Annual International Conference of the Engineering in Medicine and Biology Society; 1998. p. 895–8.
60. Kundu AK, Fattah SA, Rizve MN. An automatic bleeding frame and region detection scheme for wireless capsule endoscopy videos based on interplane intensity variation profile in normalized RGB color space. J Healthc Eng. 2018;2018:1.
61. Le Berre C, Sandborn WJ, Aridhi S, Devignes MD, Fournier L, Smaïl-Tabbone M, Danese S, Peyrin-Biroulet L. Application of artificial intelligence to gastroenterology and hepatology. Gastroenterology. 2020;158(1):76–94.
62. Lee H, Kim ST, Ro YM. Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis. In: Suzuki K, Reyes M, Syeda-Mahmood T, Glocker B, Wiest R, Gur Y, Greenspan H, Madabhushi A, editors. Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support. Cham: Springer International Publishing; 2019. p. 21–9.
63. Lee JY, Jeong J, Song EM, Ha C, Lee HJ, Koo JE, Yang DH, Kim N, Byeon JS. Real-time detection of

colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets. Sci Rep. 2020;10(1):1–9.

64. Lee S, Jang B, Kim KO, Jeon S, Kwon J, Kim E, Jung J, Park K, Cho K, Kim ES, Park C, Yang C. Endoscopic experience improves interobserver agreement in the grading of esophagitis by los angeles classification: conventional endoscopy and optimal band image system. Gut Liver. 2014;8:154–9.
65. Leenhardt R, Li C, Mouel JP, Rahmi G, Sabourin JC, Cholet F, Boureille A, Amiot X, Delvaux M, Duburque C, Leandri C, Gerard R, Lecleire S, Mesli F, Nion-Larmurier I, Romain O, Sacher-Huvelin S, Simon-Shane C, Vanbiervliet G, Dray X. Cad-cap: a 25,000-image database serving the development of artificial intelligence for capsule endoscopy. Endosc Int Open. 2020;8:E415.
66. Lipovetsky S, Conklin M. Analysis of regression in game theory approach. Appl Stoch Model Bus Ind. 2001;17:319–30.
67. Liu Q, Yu L, Luo L, Dou Q, Heng PA, Heng PA. Semi-supervised medical image classification with relation-driven self-ensembling model. IEEE Trans Med Imaging. 2020;39:3429.
68. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. Advances in neural information processing systems, vol. 30. Curran Associates; 2017. p. 4765–74. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.
69. Lux M, Chatzichristofis SA. Lire: lucene image retrieval: an extensible java cbir library. In: Proceedings of the ACM International Conference on Multimedia (ACM MM); 2008. p. 10851088.
70. Madani A, Moradi M, Karargyris A, Syeda-Mahmood T. Semisupervised learning with generative adversarial networks for chest X-ray classification with ability of data domain adaptation. In: Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI); 2018. p. 1038–42.
71. Mahmud T, Paul B, Fattah SA. PolypSegNet: a modified encoder-decoder architecture for automated polyp segmentation from colonoscopy images. Comput Biol Med. 2020;128:104119.
72. Mirza M, Osindero S. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784. 2014.
73. Misawa M, Kudo SE, Mori Y, Hotta K, Ohtsuka K, Matsuda T, Saito S, Kudo T, Baba T, Ishida F, Itoh H, Oda M, Mori K. Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). Gastrointest Endosc. 2020;93(4):960–967.e3.
74. Norman B, Pedoia V, Majumdar S. Use of 2d u-net convolutional neural networks for automated cartilage and meniscus segmentation of knee mr imaging data to determine relaxometry and morphometry. Radiology. 2018;288(1):177–85.
75. Park SY, Sargent D, Spofford I, Vosburgh KG, A-Rahim Y. A colon video analysis framework for polyp detection. IEEE Trans Biomed Eng. 2012;59 (5):1408–18.
76. Pogorelov K, Riegler M, Halvorsen P, Griwodz C, Lange T, Randel K, Eskeland S, Dang-Nguyen DT, Ostroukhova O, Lux M, Spampinato C. A comparison of deep learning with global features for gastrointestinal disease detection. In: CEUR Workshop Proceedings MediaEval, vol. 1984; 2017. p. 8–10.
77. Pogorelov K, Randel K, Griwodz C, de Lange T, Eskeland S, Johansen D, Spampinato C, Dang Nguyen DT, Lux M, Schmidt P, Riegler M, Halvorsen P. Kvasir: a multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of ACM Multimedia Systems (MMSYS); 2017.
78. Pogorelov K, Randel K, de Lange T, Eskeland S, Johansen D, Griwodz C, Spampinato C, Taschwer M, Lux M, Schmidt P, Riegler M, Halvorsen P. Nerthus: a bowel preparation quality video dataset. In: Proceedings of ACM Multimedia Systems (MMSYS); 2017.
79. Pogorelov K, Riegler M, Halvorsen P, Hicks S, Randel KR, Dang Nguyen DT, Lux M, Ostroukhova O, de Lange T. Medico multimedia task at mediaeval 2018. In: CEUR Workshop Proceedings-MediaEval; 2018.
80. Polit DF. Blinding during the analysis of research data. Int J Nurs Stud. 2011;48(5):636–41. http://www.sciencedirect.com/science/article/pii/S0020748911000496
81. Qadir HA, Balasingham I, Solhusvik J, Bergsland J, Aabakken L, Shin Y. Improving automatic polyp detection using cnn by exploiting temporal dependency in colonoscopy video. IEEE J Biomed Health Inform. 2020;24(1):180–93.
82. Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: understanding transfer learning for medical imaging. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS); 2019. p. 3347–57.
83. Repici A, Badalamenti M, Maselli R, Correale L, Radaelli F, Rondonotti E, Ferrara E, Spadaccini M, Alkandari A, Fugazza A, Anderloni A, Galtieri PA, Pellegatta G, Carrara S, Di Leo M, Craviotto V, Lamonaca L, Lorenzetti R, Andrealli A, Antonelli G, Wallace M, Sharma P, Rosch T, Hassan C. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology. 2020;159(2):512–20. http://www.sciencedirect.com/science/article/pii/S0016508520305837
84. Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16. New York: Association for Computing Machinery; 2016. p. 11351144. https://doi.org/10.1145/2939672.2939778.

85. Riegler M, Pogorelov K, Halvorsen P, de Lange T, Griwodz C, Schmidt PT, Eskeland SL, Johansen D. EIR Efficient computer aided diagnosis framework for gastrointestinal endoscopies. In: Proceeding of the International Workshop on Content-Based Multimedia Indexing (CBMI); 2016. p. 1–6.
86. Riegler M, Lux M, Griwodz C, Spampinato C, de Lange T, Eskeland SL, Pogorelov K, Tavanapong W, Schmidt PT, Gurrin C, Johansen D, Johansen H, Halvorsen P. Multimedia and medicine: teammates for better disease detection and survival. In: Proceedings of the ACM International Conference on Multimedia (ACM MM); 2016. p. 968–77. http://doi.acm.org/10.1145/2964284.2976760.
87. Riegler M, Pogorelov K, Eskeland SL, Schmidt PT, Albisser Z, Johansen D, Griwodz C, Halvorsen P, Lange TD. From annotation to computer-aided diagnosis: detailed evaluation of a medical multimedia system. ACM Trans Multimed Comput Commun Appl. 2017; https://doi.org/10.1145/3079765.
88. Rondonotti E, Soncini M, Girelli CM, Russo A, Ballardini G, Bianchi G, Cant P, Centenara L, Cesari P, Cortelezzi CC, Gozzini C, Lupinacci G, Maino M, Mandelli G, Mantovani N, Moneghini D, Morandi E, Putignano R, Schalling R, Tatarella M, Vitagliano P, Villa F, Zatelli S, Conte D, Masci E, de Franchis R. Can we improve the detection rate and interobserver agreement in capsule endoscopy? Dig Liver Dis. 2012;44(12):1006–11. http://www.sciencedirect.com/science/article/pii/S1590865812002368
89. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Proceeding of the International Conference on Medical image computing and computer-assisted intervention (MICCAI). Springer; 2015. p. 234–41.
90. Ross T, Reinke A, Full PM, Wagner M, Kenngott H, Apitz M, Hempe H, Filimon DM, Scholz P, Tran TN, Bruno P, Arbelez P, Bian GB, Bodenstedt S, Bolmgren JL, Bravo-Snchez L, Chen HB, Gonzlez C, Guo D, Halvorsen P, Heng PA, Hosgor E, Hou ZG, Isensee F, Jha D, Jiang T, Jin Y, Kirtac K, Kletz S, Leger S, Li Z, Maier-Hein KH, Ni ZL, Riegler MA, Schoeffmann K, Shi R, Speidel S, Stenzel M, Twick I, Wang G, Wang J, Wang L, Wang L, Zhang Y, Zhou YJ, Zhu L, Wiesenfarth M, Kopp-Schneider A, Mller-Stich BP, Maier-Hein L. Robust medical instrument segmentation challenge 2019. 2020.
91. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432.
92. Shapley LS. A value for n-person games. Contrib Theory Games. 1953;2(28):307–17.
93. Shin Y, Qadir HA, Balasingham I. Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance. IEEE Access. 2018;6:56007–17.
94. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. CoRR abs/1704.02685. 2017. http://arxiv.org/abs/1704.02685
95. Silva J, Histace A, Romain O, Dray X, Granado B. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. Int J Comput Assist Radiol Surg. 2014;9(2):283–93.
96. Smedsrud PH, Gjestang HL, Nedrejord OO, Næss E, Thambawita V, Hicks SA, Borgli H, Jha D, Berstad TJD, Eskeland SL, Lux M, Espeland H, Petlund A, Dang-Nguyen DT, Garcia-Ceja E, Johansen D, Schmidt PT, Hammer HL, de Lange T, Riegler M, Halvorsen P. Kvasir-capsule, a video capsule endoscopy dataset. OSF Preprints. 2020. https://doi.org/10.31219/osf.io/gr7bn
97. Tajbakhsh N, Gurudu SR, Liang J. Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans Med Imag. 2015;35(2):630–44.
98. Thambawita V, Jha D, Hammer HL, Johansen HD, Johansen D, Halvorsen P, Riegler MA. An extensive study on cross-dataset bias and evaluation metrics interpretation for machine learning applied to gastrointestinal tract abnormality classification. ACM Trans Comput Healthc. 2020;1(3). https://doi.org/10.1145/3386295
99. Thambawita V, Jha D, Riegler M, Halvorsen P, Hammer HL, Johansen H, Johansen D. The medico-task 2018: disease detection in the gastrointestinal tract using global features and deep learning. In: CEUR Workshop Proceedings -MediaEval; 2018.
100. Tomar NK, Jha D, Ali S, Johansen HD, Johansen D, Riegler MA, Halvorsen P. DDANet: Dual Decoder Attention Network forAutomatic Polyp Segmentation. arXiv preprint arXiv:2006.11392. 2020.
101. Tommasi T, Tuytelaars T. A testbed for cross-dataset analysis. In: European Conference on Computer Vision. Springer; 2014. p. 18–31.
102. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56.
103. Torralba A, Efros AA. Unbiased look at dataset bias. In: Proceedings of the International Conference on Pattern Recognition (CVPR). IEEE; 2011. p. 1521–8.
104. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA Cancer J Clin. 2015;65(2):87–108.
105. U.S. Food and Drug Administration: Learn if a medical device has been cleared by FDA for marketing. 2017. https://www.fda.gov/medical-devices/consumers-medical-devices/learn-if-medical-device-has-been-cleared-fda-marketing
106. U.S. Food and Drug Administration: Digital health innovation action plan. 2018. https://www.fda.gov/media/106331/download

107. U.S. Food and Drug Administration: FDA permits marketing of artificial intelligence algorithm for aiding providers in detecting wrist fractures. 2018. https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-algorithm-aiding-providers-detecting-wrist-fractures
108. Van Doorn SC, Hazewinkel Y, East JE, Van Leerdam ME, Rastogi A, Pellisé M, Sanduleanu-Dascalescu S, Bastiaansen BA, Fockens P, Dekker E. Polyp morphology: an interobserver evaluation for the Paris classification among international experts. Am J Gastroenterol. 2015;110(1):180.
109. Wang P, Xiao X, Brown J, Berzin T, Tu M, Xiong F, Hu X, Liu P, Song Y, Zhang D, Yang X, Li L, He J, Yi X, Liu J, Liu X, Lai L. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. Nat Biotechnol. 2018;2: 741–8.
110. Wang Y, Tavanapong W, Wong J, Oh J, de Groen PC. Part-based multiderivative edge cross-sectional profiles for polyp detection in colonoscopy. IEEE J Biomed Health Inform. 2014;18(4):1379–89.
111. Wang Y, Tavanapong W, Wong J, Oh JH, De Groen PC. Polypalert: near real-time feedback during colonoscopy. Comput Methods Progr Biomed. 2015;120 (3):164–79.
112. Wei J, Suriawinata A, Vaickus L, Ren B, Liu X, Lisovsky M, Tomita N, Abdollahi B, Kim A, Snover D, Baron J, Barry E, Hassanpour S. Evaluation of a deep neural network for automated classification of colorectal polyps on histopathologic slides. JAMA Netw Open. 2020;3:e203398.
113. Wickstrm K, Kampffmeyer M, Jenssen R. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. Med Image Anal. 2020;60:101619. http://www.sciencedirect.com/science/article/pii/S1361841519301574
114. Wickstrøm K, Kampffmeyer M, Jenssen R. Uncertainty modeling and interpretability in convolutional neural networks for polyp segmentation. In: Proceedings of the IEEE international workshop on machine learning for signal processing (MLSP). IEEE; 2018. p. 1–6.
115. Woolhandler S, Himmelstein DU. Administrative work consumes one-sixth of U.S. physicians working hours and lowers their career satisfaction. Int J Health Serv. 2014;44(4):63542.
116. Wu H, Prasad S. Semi-supervised deep learning using pseudo labels for hyperspectral image classification. IEEE Trans Image Process. 2018;27(3):1259–70.
117. Xie Q, Luong MT, Hovy E, Le QV. Self-training with Noisy Student improves ImageNet classification. arXiv. 2020. http://arxiv.org/abs/1911.04252
118. Xue Y, Xu T, Long LR, Xue Z, Antani S, Thoma GR, Huang X. Multimodal recurrent model with attention for automated radiology report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2018. p. 457–66.
119. Yang J, Faraji M, Basu A. Robust segmentation of arterial walls in intravascular ultrasound images using dual path u-net. Ultrasonics. 2019;96:24–33.
120. Zhang C, Tavanapong W, Wong J, de Groen PC, Oh J. Real data augmentation for medical image classification. In: Cardoso MJ, Arbel T, Lee SL, Cheplygina V, Balocco S, Mateus D, Zahnd G, Maier-Hein L, Demirci S, Granger E, Duong L, Carbonneau MA, Albarqouni S, Carneiro G, editors. Intravascular imaging and computer assisted stenting, and large-scale annotation of biomedical data and expert label synthesis, vol. 10552. Springer International Publishing; 2017. p. 67–76. http://link.springer.com/10.1007/978-3-319-67534-3_8.
121. Zhang Z, Xie Y, Xing F, McGough M, Yang L. Mdnet: a semantically and visually interpretable medical image diagnosis network. In: Proceedings of IEEE CVPR; 2017. p. 6428–36.
122. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. Unet ++: a nested u-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer; 2018. p. 3–11.
123. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 2223–32.