|  |  |  |
| --- | --- | --- |
|  | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2022-2024 | FG-AI4H-P-025-A01 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | Plenary | Helsinki, 20-22 September 2022 |
| **DOCUMENT** |
| **Source:** | TG-Endoscopy Topic Driver |
| **Title:** | Att.1 – TDD update (TG-Endoscopy) |
| **Purpose:** | Discussion |
| **Contact:** | Jianrong WuTG-EndoscopyTencent Healthcare (Shenzhen) Co.,LtdChina | E-Mail: edwinjrwu@tencent.com |
| **Contact:** | Shan XuTG-EndoscopyCAICTChina | E-Mail: xushan@caict.ac.cn  |

|  |  |
| --- | --- |
| **Abstract:** | This topic description document (TDD) specifies a standardized benchmarking for AI-based Endoscopy. It covers all scientific, technical, and administrative aspects relevant for setting up this benchmarking (and follows the template structure defined in document FGAI4H-J-105). The creation of this TDD is an ongoing iterative process until it is approved by the Focus Group on AI for Health (FG-AI4H) as deliverable No. DEL 10.20. This draft will be a continuous input and output document. |

|  |  |
| --- | --- |
| **Change notes:** | *Version 1* (submitted as FGAI4H-J-025-A01-R01 to meeting J (e-meeting), 30 Sept – 2 Oct 2020)* Drafted this TDD.

*Version 2* (submitted as FGAI4H-K-025-A01-R01 to meeting K (e-meeting), 27 – 29 Jan 2021)* Arranged and aligned the TDD following the new template J-105.

*Version 3* (submitted as FGAI4H-M-025-A01-R01 to meeting M (e-meeting), 28 – 30 Sept 2021)* Introduced 3 new participants, and 1 new subtopic of endoscopic ultrasound.
* Modified the chapter of Method, including additional content to chapter of Input data structure, new chapter of Output data structure, new chapter of Annotation procedure combining content from previous chapters of Information requirements of annotation, Annotation of detection, Annotation of classification, Annotation of segmentation
* Deleted chapter of Existing work on benchmarking, move corresponding content to the chapter of Evaluation score and metrics.

*Version 4* (submitted as FGAI4H-N-025-A01-R01 to meeting N (e-meeting), 15 – 17 Feb 2022)* Introduced 1 new participants
* Modified the structure of the TDD, such as adding new chapter of Definition of the AI task, Current gold standard, Relevance and impact of an AI solution
* Updated content of chapter Status of this topic group, Definition of the AI task, Current gold standard, Relevance and impact of an AI solution, Existing AI solutions, Annotation of classification and Annotation of segmentation

*Version 5* (submitted as FGAI4H-P-025-A01-R01 to meeting P, 19 – 22 Sep 2022)* Updated content of chapter Status of this topic group, Subtopics, Relevance and impact of an AI solution, Existing work on benchmarking, Available public data and undisclosed test data set collection, Ethical considerations
 |

**Contributors**

|  |  |
| --- | --- |
| Yajun ZhangTencent Technology (Shenzhen) Co.,LtdChinaYi CaiOlympus Medical Systems Corp. JapanJunbo LiSuzhou Institute of Biomedical Engineering and Technology, Chinese Academy of SciencesChinaYanchun ZhuChina Unicom (Guangdong) Industrial Internet Co., LtdChina | Email: yajunzhang@tencent.com Email: i.sai@olympus.comEmail: Lijb@sibet.ac.cnEmail: zhuyc82@chinaunicom.cn |

**CONTENTS**

|  Page |
| --- |
| [1 Introduction 6](#_Toc113974658)[1.1 About the FG-AI4H topic group on Endoscopy 6](#_Toc113974659)[1.1.1 Documentation 6](#_Toc113974660)[1.1.2 Status of this topic group 7](#_Toc113974661)[1.1.3 Topic group participation 8](#_Toc113974662)[1.1.4 Topic description 8](#_Toc113974663)[1.1.5 Subtopics 8](#_Toc113974664)[1.1.6 Definition of the AI task 9](#_Toc113974665)[1.1.7 Current gold standard 10](#_Toc113974666)[1.1.8 Relevance and impact of an AI solution 10](#_Toc113974667)[1.2 Ethical considerations 11](#_Toc113974668)[1.3 Regulatory considerations 11](#_Toc113974669)[2 Existing AI solutions 12](#_Toc113974670)[3 Existing work on benchmarking 13](#_Toc113974671)[3.1 EndoCV 13](#_Toc113974672)[3.2 EndoVis 14](#_Toc113974673)[3.2.1 GIANA 14](#_Toc113974674)[3.2.2 CATARACTS 15](#_Toc113974675)[3.2.3 DAGI 15](#_Toc113974676)[3.2.4 APDCV 16](#_Toc113974677)[3.3 EndoTect 16](#_Toc113974678)[4 Benchmarking by the topic group 17](#_Toc113974679)[4.1 Method 17](#_Toc113974680)[4.1.1 Input data structure 17](#_Toc113974681)[4.1.2 Output data structure 18](#_Toc113974682)[4.1.3 Annotation procedure 19](#_Toc113974683)[4.2 Score and metrics 22](#_Toc113974684)[4.2.1 Evaluation score and metrics 22](#_Toc113974685)[4.2.2 Other quality score and metrics criteria 23](#_Toc113974686)[5 Overall discussion of the benchmarking 25](#_Toc113974687)[6 Regulatory considerations 25](#_Toc113974688)[6.1 Existing applicable regulatory frameworks 26](#_Toc113974689)[6.2 Regulatory features to be reported by benchmarking participants 26](#_Toc113974690)[6.3 Regulatory requirements for the benchmarking systems 26](#_Toc113974691)[6.4 Regulatory approach for the topic group 26](#_Toc113974692)[7 References 26](#_Toc113974693)[Annex A: Glossary 29](#_Toc113974694)[Annex B: Declaration of conflict of interests 31](#_Toc113974695) |

**List of Tables**

[Table 1: Topic group output documents 6](#_Toc113973938)

[Table 2: Gastrointestinal Image ANAlysis (GIANA) 14](#_Toc113973939)

[Table 3: Benchmarking metrics 22](#_Toc113973940)

[Table 4: Available public data 23](#_Toc113973941)

**List of Figures**

[Figure 1: Illustration of annotation procedure for detection 23](#_Toc95388875)

[Figure 2: Illustration of annotation procedure for classification 25](#_Toc95388876)

FG-AI4H Topic Description Document

Topic group - AI for Endoscopy

# Introduction

Endoscopy is the core technical means for early diagnosis and screening of digestive cancer. Implementing endoscopic screening for digestive cancer can detect and treat precancerous lesions, which can drastically reduce the incidence and mortality of digestive cancer. Due to factors such as the endoscopic doctor's operating, the ability to identify lesions, and visual fatigue, a considerable proportion of lesions in clinical diagnosis, including even advanced and precancerous lesions, may be missed by the endoscopic doctor.

In recent years, with the breakthrough of the new generation of artificial intelligence technology represented by deep learning, revolutionary progress has been made in the field of automatic recognition of medical images. The real-time assistance of artificial intelligence to detect and classify gastrointestinal lesions is expected to help clinicians improve their examination quality and reduction of missed diagnoses.

This topic description document specifies the standardized benchmarking for Endoscopy systems. It serves as deliverable No. DEL 10.20 of the ITU/WHO Focus Group on AI for Health (FG-AI4H).

## About the FG-AI4H topic group on Endoscopy

The introduction highlights the potential of a standardized benchmarking of AI systems for Endoscopy to help solve important health issues and provide decision-makers with the necessary insight to successfully address these challenges.

To develop this benchmarking framework, FG-AI4H decided to create the TG-Endoscopy at the meeting I e-meeting, 7-8 May 2020​​.

FG-AI4H assigns a *topic driver* to each topic group (similar to a moderator) who coordinates the collaboration of all topic group members on the TDD.During the FG-AI4H meeting I e-meeting, 7-8 May 2020​​, Dr Jianrong Wu from Tencent Healthcare was nominated as the topic driver for the TG-Endoscopy.

### Documentation

This document is the TDD for the TG-Endoscopy. It will be developed cooperatively by all members of the topic group over time and updated TDD iterations are expected to be presented at each FG-AI4H meeting.

The final version of this TDD will be released as deliverable “DEL 10.20 Endoscopy (TG- Endoscopy).” The topic group is expected to submit input documents reflecting updates to the work on this deliverable **(Table 1)** to each FG-AI4H meeting.

Table 1: Topic group output documents

| Number | Title |
| --- | --- |
| FGAI4H-P-025-A01-R01 | Latest update of the Topic Description Document of the TG-Endoscopy  |
| FGAI4H-J-025-A02 | Latest update of the Call for Topic Group Participation (CfTGP) |
| N/A | The presentation summarizing the latest update of the Topic Description Document of the TG-Endoscopy |

The working version of this document can be found in the official topic group SharePoint directory.

* https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Endoscopy.aspx

The TDD document is structured into the following sections:

* **Section 1** introduces the general information of FG-AI4H topic group on endoscopy, ethical considerations and regulatory consideration..
* **Section 2** provides an overview of the existing AI solutions for Endoscopy.
* **Section 3** describes the existing approaches for assessing the quality of AI-based Endoscopy systems and details that are likely relevant for setting up a new standardized benchmarking.
* **Section 4** specifies the actual benchmarking methods for all subtopics at a level of detail that includes technological and operational implementation.
* **Section 5** discusses the overall insights gained from benchmarking work in this topic group.
* **Section 6** specifieshow the different regulatory aspects related to TG-Endoscopy.
* **Section 7** provides the references materials.

### Status of this topic group

The following subsections describe the update of the collaboration within the TG-Endoscopy for the official focus group meetings.

#### Status update for meeting I

* Discussed the proposal from Tencent
* Approved AI for Endoscopy as a use case for FG-AI4H
* Established the topic group at Meeting I (online, 7-8 May 2020)
* Nominated the topic group driver

#### Status update for meeting J

* Start the draft of TDD
* Start the draft of the call for participation
* Present the initial documents for AI for endoscopy (TG-Endoscopy)

#### Status update for meeting K

* Updates the initial documents of AI for endoscopy
* Reach out for new participants.

#### Status update for meeting L

* Transform the initial document of TG-Endoscopy into TDD template format
* Reach out for new participants.

#### Status update for meeting M

* Update the TDD for AI for endoscopy
* Add new subtopic as endoscopic ultrasound
* Invite new Participants.

#### Status update for meeting N

* Update the TDD for AI for endoscopy
* Modified the structure of chapters
* Invite new participants

#### Status update for meeting P

* Update the TDD for AI for endoscopy

### Topic group participation

Participation in both, the Focus Group on AI for Health and in a TG is generally open to anyone (with a free ITU account). For this TG, the corresponding ‘Call for TG participation’ (CfTGP) can be found here:

* <https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/tg/CfP-TG-Endoscopy.pdf>

Each topic group also has a corresponding subpage on the ITU collaboration site. The subpage for this topic group can be found here:

* <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Endoscopy.aspx>

For participation in this topic group, interested parties can also join the regular online meetings. For all TGs, the link will be the standard ITU-TG ‘zoom’ link:

* <https://itu.zoom.us/my/fgai4h>

All relevant administrative information about FG-AI4H—like upcoming meetings or document deadlines—will be announced via the general FG-AI4H mailing list fgai4h@lists.itu.int.

All TG members should subscribe to this mailing list as part of the registration process for their ITU user account by following the instructions in the ‘Call for Topic Group participation’ and this link:

* <https://itu.int/go/fgai4h/join>

Regular FG-AI4H workshops and meetings proceed about every two months at changing locations around the globe or remotely. More information can be found on the official FG-AI4H website:

* <https://itu.int/go/fgai4h>

### Topic description

This topic contains a detailed description and background information of the specific health topic for the benchmarking of AI for Endoscopy and how this can help to solve a relevant ‘real-world’ problem. This topic groups summarize uses cases of a certain health topic or problem and similar AI benchmarking requirements for Endoscopy.

### Subtopics

TG-Endoscopy started without separate subtopic groups. And by the efforts of participants, there are two subtopics in progress, including colonoscopy and endoscopic ultrasound. However, it is possible that future subtopics for gastroscopy, laryngoscopy, rhinoscopy might be introduced to pursue more different topic-specific specializations, and the topic group driver will moderate the activities within the topic group.

#### Colonoscopy

Colorectal cancers (CRC) are the third most prevalent cancer and the second-highest cause of cancer deaths worldwide. Colonoscopy is the gold standard for CRC screening. In clinical practice, colonoscopy requires both manipulation and observation at the same time, and it cannot detect all colonic polyps, some of which may be neoplasms. Colonoscopy has been reported to miss 17%-48% of adenomas which are considered to be 50%-60% causes of interval cancers.

Over the last two decades, computer-assisted polyp detection has been actively explored to improve inspecting quality and reduce adenoma miss rates (AMR). Recently, artificial intelligence (AI) has made remarkable breakthroughs in medical fields with deep learning and convolutional neural networks (CNNs). With enough qualified learning materials, CNNs can reach even higher real-time detecting accuracy than human experts, which demonstrate that computer assisted-detection systems (CADe) might have the potential to serve as real-time ‘experts’ to improve the quality of colonoscopies.

#### Endoscopic Ultrasound

Endoscopic ultrasound (EUS) is a minimally invasive procedure in which endoscopy is combined with ultrasound to obtain images of the internal organs. EUS has emerged as an important imaging modality for the diagnosis and staging of benign and malignant lesions in the upper digestive tract and the respiratory system. EUS is most commonly used for staging of GI malignancies, evaluating pancreaticobiliary disease, evaluating subepithelial abnormalities, evaluating extraluminal abnormalities, staging of lung cancer and image guidance for therapeutic procedures [1]. Research has shown remarkable improvement of sensitivity and specificity in various tumor diagnoses using EUS. For instance, EUS is capable to identify small pancreatic tumours with a staging sensitivity greater than 90% [2]. Endobronchial ultrasound (EBUS) has been used for lung cancer staging with a diagnostic accuracy of 90–100% [3].

Research on artificial intelligence (AI) in EUS is still limited [4]. Only a handful of reports were published based on limited clinical data through retrospective or prospective studies, with a focus on pancreatic diseases. Currently, there’s no commercial AI product for EUS on the market.

### Definition of the AI task

The application of AI in the field of endoscopy varies according to different clinical goals. In general, it is mainly divided into the following three categories: classification, detection, and segmentation.

#### Classification

Classification is a machine learning task for determining which classes are in an image, video or other types of data. It refers to training machine learning models with the intent of finding out which classes are present. In practical clinical applications, it is possible to classify with/without lesions in endoscopic entire image data from a patient. It is also possible to evaluate the image quality of all endoscopic images of the patient, from which the quality of the image meets the diagnostic quality requirements.

#### Detection

Object detection combines classification and localization to determine what objects are in the image or video and specify where they are in the image. It applies classification to distinct objects and uses bounding boxes. Object detection is useful in identifying objects in an image or video. In endoscopy, detection can effectively reduce the missed rate of screening. For example, in colorectal examination, the detection rate of polyps can be improved.

#### Segmentation

Image segmentation separates an image into regions in pixel level, with particular shape and border, delineating potentially meaningful areas for further processing, such as measurement, classification and object detection. The regions may not take up the entire image, but the goal of image segmentation is to highlight foreground elements and make it easier to be evaluated. Image segmentation provides pixel-by-pixel details of an object, making it different from classification and object detection. For example, in endoscopy, the tumor size and area can be automatically calculated based on the image segmentation results.

### Current gold standard

#### Colonoscopy

Colonoscopy is considered the gold standard for CRC screening to detect and remove the polyps and adenomas in the colorectum[5]. To build an AI solution for colonoscopy, such as CADe and computer-aided diagnosis (CADx) system, gold standard is necessary besides colonoscopic images, which could be annotated by objective or subjective methods. By objective methods, the gold standard would be annotated by information from clinical diagnosis report. For example, to train a CADx classifying the nature of polyps by image, the nature of polyps in pathology report could be used for gold standard[6][7][8]. By subjective methods, the gold standard would be made by colonoscopists manually[9]. Usually, subjective method is used for polyp detection[10][11] and segmentation tasks[9][12][13][14], whose annotation results are bounding-box and mask respectively. The subjective method might involve one or multiple colonoscopists in single-step or multi-steps[10] procedures of annotation.

#### Endoscopic Ultrasound

Endoscopic ultrasound (EUS) has become a widely used diagnostic and therapeutic procedure for various types of pancreatic diseases throughout the world. Clinical evidences have shown the benefits of EUS over the potential adverse events (AEs) and clinical guidelines have been published and continuously updated to ensure the safely use of the procedures [15][16][17][18]. European Society of Gastrointestinal Endoscopy (ESGE) has suggested EUS for pancreatic cancer screening in selected high-risk patients [15], recommended EUS-guided sampling for pancreatic solid lesions as first line procedure and EUS-guided sampling for biochemical analysis plus cytopathologic examination for pancreatic cystic lesions, etc. [16], and recommended EUS as therapeutic procedures over various types of diseases, including percutaneous transhepatic biliary drainage (PTBD), pancreatic duct (PD) drainage, etc [17].

### Relevance and impact of an AI solution

#### Colonoscopy

The significance of screening colonoscopy largely lies in the detection and removal of colorectal polyps. In clinical practice, colonoscopy is a highly operator-dependent procedure and requires both manipulation and observation at the same time, which may lead to significant variation of adenomas missing rate between individual endoscopists[19]. CADe with real-time automatic polyp detection or classification powered by AI algorithms has been proposed to help endoscopist to improve the polyp detection rate and adenomas missing rate[20].

A prospective study including 1058 patients was designed as a randomised controlled trial to investigate the impact of an automatic polyp detection CADe acting as an assistant to the endoscopist on PDR and ADR, in Endoscopy Center of the Sichuan Provincial People’s Hospital, China. The colonoscopy with CADe increased ADR (29.1%vs20.3%, p<0.001), the mean number of adenomas per patient (0.53vs0.31, p<0.001), hyperplastic polyps (114vs52, p<0.001)[21].

Over 71000 images from 20 centers was used to train and test a deep learning-based CADe in Changhai Hospital in Shanghai, China. The CADe was able to identify polyps in the test dataset with 95.0% sensitivity and 99.1% specificity. Colonoscopists can detect more polyps (0.90 vs 0.82, P < 0.001) and adenomas (0.32 vs 0.30, P = 0.045) with the aid of CADe, particularly polyps < 5 mm and flat polyps (0.65 vs 0.57, P < 0.001; 0.74 vs 0.67, P = 0.001, respectively)[22].

Besides detection, there has been attempt for classification. A deep convolutional neural network model was trained to predict the histology of polyps using only narrow band imaging. The accuracy of the model was 94% (95% CI 86% to 97%), the sensitivity for identification of adenomas was 98% (95% CI 92% to 100%), specificity was 83% (95% CI 67% to 93%), negative predictive value 97% and positive predictive value 90%[6].

#### Endoscopic Ultrasound

EUS has been proven to be an effective imaging modality for local-reginal staging of gastrointestinal tumors. The diagnostic ability of EUS is higher than that of computed tomography (CT), transabdominal ultrasonography, and magnetic resonance imaging (MRI) [23][24]. It has also proved to be a useful alternative therapeutic modality in surgery. However, EUS may be less accurate for early staging of esophageal cancer. According to a meta-analysis by Puli et al., the diagnostic accuracy of EUS was higher for T3-T4 lesions (>90%) than T1-T2 (65%) [24]. It’s also shown low accuracy of using EUS for differentiating benign and malignant rectal cancer after treatment. Another limitation for EUS (as well as other ultrasonography procedures) is its operator-dependency. The performance of EUS improves with experiences. High inter-observer variability (61%-77%) has been reported and a wide range of overall accuracy for tumor staging could be found between different studies (63% to 95%) [23].

AI is believed to play important role in endoscopic procedures, not only to detect anatomical features, differentiate benign and malignant lesions, delineate lesion contours, but more important to reduce learning time for junior endoscopists, decrease work load and standardize the overall quality of endoscopic procedures.

## Ethical considerations

The rapidly evolving field of AI and digital technology in the fields of medicine and public health raises several ethical, legal, and social concerns. All these concerns are discussed in deliverable [DEL01](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B0505B020-362C-45B2-94BF-215D2EBBD8F5%7D&file=DEL01.docx&action=default) “*AI4H ethics considerations,”* which was developed by the working group on “Ethical considerations on AI4H” (WG-Ethics). This section refers to DEL01 and disccuss the ethical considerations of the TG-Endoscopy.

Collecting massive data is necessary for AI solution development. However, ethical considerations such as patient privacy concerns should be taken into careful consideration and relevant regulations should be followed. Otherwise, the privacy of patients must be protected in the process of data collection, transmission and utility. If the data contains patient private information or identified codes, data desensitization must be performed. Generally, it is better for data sources, such as hospital and other clinical institution, to be responsible for handling the ethical, legal and privacy of the relevant data. And the following procedure is suggested.

1. Patients consent procedure at each individual institution.
2. Review of the data collection plan by a local medical ethics committee or an institutional review board.
3. Anonymization of the video or image frames (including demographic information) by clinical institution prior to sending to AI developer.
4. Anonymization of the video or image frames (including demographic information) by AI developer prior to utility (optional).

## Regulatory considerations

For AI-based technologies in healthcare, regulation is not only crucial to ensure the safety of patients and users, but also to accomplish market acceptance of these devices. This is challenging because there is a lack of universally accepted regulatory policies and guidelines for AI-based medical devices. To ensure that the benchmarking procedures and validation principles of FG-AI4H are secure and relevant for regulators and other stakeholders, the working group on *“*[*Regulatory considerations on AI for health”* *(WG-RC)*](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/wg/SitePages/WG-RC.aspx) compiled the requirements that consider these challenges. The deliverables with relevance for regulatory considerations are [DEL02](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BF2F46A99-7457-4BC8-81A3-0E1E63D6072A%7D&file=DEL02.docx&action=default) *“AI4H regulatory considerations”* (which provides an educational overview of some key regulatory considerations), [DEL02\_1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B6AF7C004-8BCE-4151-9F44-45F041A1EB1D%7D&file=DEL02_1.docx&action=default) *“Mapping of IMDRF essential principles to AI for health software”,* and[DEL02\_2](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B1ED0D4D1-876C-4A0F-AEF7-06D3F445F5E6%7D&file=DEL02_2.docx&action=default) *“Guidelines for AI-based medical device (AI-MD): Regulatory requirements”* (which provides a checklist to understand expectations of regulators, promotes the step-by-step implementation of safety and effectiveness of AI-based medical devices, and compensates for the lack of a harmonized standard).

TG-Endoscopy refers to the deliverables of WG-RC.

# Existing AI solutions

Tencent Healthcare and Changhai Hospital developed a CADe system in 2021 that was built based on the You Only Look Once v2 deep learning framework. The system detects potential polyp and presents an alert rectangle surrounding polyps on a second monitor for colonoscopists. Colonoscopists detect more polyps and adenomas with the aid of CADe system, particularly polyps<5 mm and flat polyps. However, high efficacy is not realized in colonoscopies with inadequate bowel preparation and withdrawal time. [22]

National Cancer Center Hospital and NEC Japan successfully developed a system in 2017 that immediately detects colorectal cancer and ulcerative colon polyps, a precursor to cancer, during an endoscopic examination using artificial intelligence (AI). It automatically detects colorectal cancer and polyps from images and videos taken during an endoscopic examination of the colon and aids in the discovery of lesions by endoscopists. It improves polyp detection, which was an issue during such exams, and increases the detection rate. In this manner, it greatly contributes to the prevention and early detection of colorectal cancer.[25]

Wision AI and Sichuan Provincial People’s Hospital developed a real-time automatic polyp detection system in 2018 that detects colorectal polyps during an endoscopic examination using deep learning. The detection algorithm is a deep CNN based on SegNet architecture. If any polyp is detected by the system, a hollow tracking box around would be shown on the monitor. As a conclusion, in a low prevalent ADR population, an automatic polyp detection system during colonoscopy resulted in a significant increase in the number of diminutive adenomas detected, as well as an increase in the rate of hyperplastic polyps. the cost–benefit ratio of such effects has to be determined further [21][15]. Its product for real-time colorectal polyps detection named as EndoScreener® has been certificated by Chinese NMPA, American FDA and European CE-MDR.

National Chiao Tung University and Tri-Service General Hospital developed a CADx in 2018 with a deep neural network to analyze narrow-band images of diminutive colorectal polyps. The system could classify the polyps in narrow-band images as neoplastic or hyperplastic. [16]

Sun Yat-sen University developed a CADe system in 2018 with deep learning to detect upper gastrointestinal cancers by endoscopy that is named as Gastrointestinal Artificial Intelligence Diagnostic System (GRAIDS). It is the first real-time AI-aided image recognition system that has been implemented in clinical practice for detecting upper gastrointestinal cancers during endoscopy. [26]

Zhongshan Hospital and University of California developed an artificial intelligence–based CNN-CAD system through transfer learning leveraging a state-of-the-art pretrained CNN architecture, ResNet50. The system is used to determine the invasion depth of the gastric cancer and screen patients for endoscopic resection. This system distinguished early gastric cancer from deeper submucosal invasion and minimized overestimation of invasion depth, which could reduce unnecessary gastrectomy.[27]

Cancer Institute Hospital Ariake, AI Medical Service and Tada Tomohiro Institute of Gastroenterology and Proctology developed a CNN-based diagnostic system based on Single Shot MultiBox Detector architecture to detect gastric cancer in endoscopic images. This constructed CNN system for detecting gastric cancer could process numerous stored endoscopic images in a very short time with a clinically relevant diagnostic ability. [28]

Renmin Hospital of Wuhan University developed a system using a novel deep convolution neural network (DCNN) to detect early gastric cancer (EGC) without blind spots during esophagogastroduodenoscopy (EGD). This system could identify EGC from non-malignancy and classify gastric location into 10 or 26 parts with high accuracy. [29]

Kindai University developed a system in 2017 that could diagnose colon polyps as adenomatous or non-adenomatous using a simple CNN. [30]

# Existing work on benchmarking

This section focuses on the existing benchmarking processes in the context of AI and Endoscopy for quality assessment. By the effort of researchers, there have been many AI methods for automatic detection, segmentation and classification of polyps or other lesions in endoscopy.

## EndoCV

The International Workshop and Challenge on Computer Vision in Endoscopy(EndoCV) is held from 2019 to 2022 annually[31][32][33][34][35][36], which aimed to benchmark methods on larger test-set comprising of mostly video sequences as in the real-world clinical scenario for endoscopy artefact detection, endoscopy disease detection and polyp generalization[37][38][39][40]. To achieve high diversity, the dataset was built by data from multiple centers in different countries that includes Egypt, France, Italy, Norway, Sweden and UK. The resolution of data included standard definition, HD and Ultra HD. And the data was collected by different endoscopy manufacturers, includes Olympus (mostly), Fujifilm and Karl Storz.

The latest 4th International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV2022) was held in conjunction with IEEE International Symposium on Biomedical Imaging (ISBI2022)[37]. There were two sub-challenges: Endoscopy artefact detection (EAD 2.0) and polyp generalization (PolypGen 2.0). The aim of the sub-challenge EAD 2.0 was to localize bounding boxes, predict class labels and pixel-wise segmentation of 8 different artefact classes for given clinical endoscopy video clips, including specularity, bubbles, saturation, contrast, blood, instrument, blur and imaging artefacts. PolypGen 2.0 aimed to benchmark methods on the basis of generalization capabilities to unseen colonoscopy video sequence data for both detection and segmentation deep learning methods.

* **AI task:**
* **Detection task**: The aim of this task was to test the performance of participants’ methods for detection and localization task on our comprehensive and sorted multi-center datasets. The participants were tested on both detection-based metric and localization metric. A weighted final metric was used to evaluate for the best performing method.
* **Segmentation task**: Each participants methods was evaluated on multi-center curated and sorted datasets.
* **Dataset:**
* **EAD 2.0:** A total of 280 patient videos from multiple organs and institutions with 45,478 annotations on both single frame and sequence video data. Training data for the detection task consisted of total 2531 frames with 31,069 bounding boxes while 643 frames with 7511 binary masks for the segmentation task[40]. Besides, there was a new set of test data were curated that include unique video sequences consisting of more than 500 frames.
* **PolypGen 2.0:** The dataset was composed of a total of 6282 frames including both single and sequence frames from 6 centers incorporating more than 300 patients. It consisted of 3762 positive sample frames and 2520 negative sample frames with 3446 annotated polyp labels with precise delineation of polyp boundaries (pixel level for segmentation task and bounding boxes for detection task) verified by six senior gastroenterologists
* In addition to this dataset, additional 23 unique patient video clips (> 100 frames per video) making in total of 46 sequences for PoypGen2.0 and 24 sequences were collected for EAD2.0.
* **Annotation:**
* First, a small subset of dataset were annotated by all clinical experts and a joint consensus was made available.
* Then, the remaining subset of dataset was annotated by post-doctoral researchers (working on endoscopy) and validated by clinicians at two different centres (10-fold cross-validation).
* Finally, through a joint conference call all annotation validation was achieved.
* **Metrics:**
* **Detection task:**
* Standard computer vision metric: mean average precision (mAP).
* Standard intersection over union (IoU).
* Final detection score (trade-off between mAP and IoU): 0.6\*mAP + 0.4\*IoU.
* Generalization gap (Gerror): defined as the difference between detection score and the generalization score (on unseen data) .
* Centroid localisation error (Lerror): defined as the distance between centroids positions of detected boxes between the consecutive frames in a video (new).
* Clinical applicability metrics: runtime (to be used post challenge only).
* **Segmentation task:**
* Standard segmentation metrics that include Dice coefficient (DICE), F2-error, positive predictive value (PPV), Hausdorff distance (HD) and sensitivity (recall) were used.
* The ranking on leaderboard were based on the highest mean value between DSC, PPV and sensitivity; and the least HD value.
* Generalizability difference (Gerror): Difference between DSC on mixed sample data and DSC on unseen data will be key in deciding winner of this task.
* Clinical applicability metrics: runtime (to be used post challenge only).

## EndoVis

Endoscopic Vision Challenge(EndoVis)[41] organizes high-profile international challenges for the comparative benchmarking and validation of endoscopic vision algorithms that focus on different problems each year at International Conference on Medical Image Computing and Computer Assisted Intervention(MICCAI) from 2015 till now except 2016, while there were several sub-challenges in each year.

### GIANA

Gastrointestinal Image ANAlysis (GIANA) is one of the sub-challenge in EndoVis, which was held in 2017, 2018[42] and 2021[43]. In general, GIANA includes polyp detection, segmentation and classification in colonoscopy images, and polyp segmentation, angiodysplasia detection and localization in wireless capsule images.

Table 2: Gastrointestinal Image ANAlysis (GIANA)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Year** | **Task** | **Modality** | **Definition** | **Clinical use** | **Database content**  | **Gruoud truth** |
| GIANA 2018 | Polyp detection | Colonoscopy | Ability to detect presence/absence of polyps in each frame AND, in case of polyp presence, locate it within the image | Prevention of colorectal cancer | 18 short videos for training, more than 20 short and long videos for testing  | Polyp masks, Paris classification Clinical partner: Hospital Clinic, Barcelona, Spain |
| Polyp segmentation | Colonoscopy  | Delimit the region the polyp occupies in the image | Preliminary stage for lesion classification through analysis of polyp region content  | Two sets: Standard Definition images (300 images for training, 612 for testing) and High Definition images (more than 150 images)  | Polyp masks, Paris classificationClinical partner: Hospital Clinic, Barcelona, Spain |
| Angiodysplasia detection  | Wireless Capsule Endoscopy | Label each of the frames into angiodysplasia containing or not  | Automatic detection of small bowel lesions related to bleeding  | 600 images for training (same number of positive and negative examples) and 600 images for testing.  | Angiodysplasia maskClinical partner: Saint Antoine Hopital, Paris, France |
| Angiodysplasia localization  | Wireless Capsule Endoscopy | Label each of the frames into angiodysplasia containing or not and in case angiodysplasia is detected, localize the region it occupies in the image  | Automatic detection of small bowel lesions related to bleeding  | 600 images for training (same number of positive and negative examples) and 600 images for testing.  | Angiodysplasia maskClinical partner: Saint Antoine Hopital, Paris, France |
| GIANA 2021 | Polyp detection  | Colonoscopy  |  Ability to detect presence/absence of polyps in each frame AND, in case of polyp presence, locate it within the image  | Prevention of colorectal cancer  | 18 short videos for training, more than 20 short and long videos for testing  | Polyp masks, Paris classification Clinical partner: Hospital Clinic, Barcelona, Spain |
| Polyp segmentation | Colonoscopy | Delimit the region the polyp occupies in the image | Preliminary stage for lesion classification through analysis of polyp region content  | Two sets: Standard Definition images (300 images for training, 612 for testing) and High Definition images (more than 150 images)  | Polyp masks, Paris classificationClinical partner: Hospital Clinic, Barcelona, Spain |
| Polyp Classification (frames)  | Colonoscopy  | Label each of the frames with one of the following categories: 1) Adenoma, 2) Non-adenoma  | In-vivo diagnosis, advance patient treatmen | 1000 images for training and validation and 200 images for testing  | Label of each frame, polyp regionClinical partner: Hospital Clinic, Barcelona, Spain |

### CATARACTS

The Challenge on Automatic Tool Annotation for cataract Surgery (CATARACTS)[44] released 50 cataract surgery videos accompanied by instrument usage annotations including frame-level instrument presence information in 2017. Pixel-wise semantic annotations for anatomy and instruments of 36 classes for 4670 images sampled from 25 videos of the CATARACTS training set was released further in 2020, including 4 anatomical classes, 29 instruments and 3 classes of other objects appearing in the scene. As one of the sub-challenge in EndoVis 2020, there were three sub-tasks to assess participating solutions on anatomical structure and instrument segmentation, including: 1)Anatomy and instrument, 2)Anatomy and grouped instruments, 3)Anatomy, instrument tips and handles[45]. Their performance was assessed on a hidden test set of 531 images from 10 videos of the CATARACTS test set. There were 25 classes in the test set, including 4 anatomical classes, 18 instruments and 3 other objects in the scene in particular. And the mean Intersection over Union (mIoU) was used to assess model performance.

### DAGI

The challenge of Detection of Abnormalities in Gastroscopic Images[46] was one of the sub-challenge in EndoVis 2015, focusing on comparing different abnormal detection methods for recognizing the abnormal regions from gastroscopic images. The abnormal detection for gastroscopic images were addressed with different abnormal patterns, such as gastritis, cancer, ulcer and bleeding.

Totally 800 gastroscopic images from 137 volunteers were involved, while three senior experts were invited to annotate the lesion/abnormal regions independently, and the pixel-level ground truth were the average.

For benchmark and evaluation, the Area Under Curve (AUC) of Receiver Operating Characteristic (ROC) curve was used. The performance was based on the image-level predictions in particular. For positive images, one image can be considered as true positive if at least 40% of the truly abnormal pixels are detected, otherwise it will be considered as false negative. For negative images, one image can be considered as true negative only when no abnormal pixel is detected, otherwise it will be considered as false positive.

### APDCV

The challenge of Automatic Polyp Detection in Colonoscopy Videos[47] was one of the sub-challenge in EndoVis 2015 and the first challenge about polyp detection. This challenge was to automatically detect polyps in colonoscopy videos, thereby reducing polyp miss-rate and the subsequent mortality rate of colon cancer.

There were two tasks, including frame classification of polyp existence and polyp detection in image, which were defined as polyp detection and polyp localization in the challenge respectively.

Three different public databases were used in the context of the benchmark, including CVC-CLINIC[48], ETIS-LARIB and ASU-Mayo Clinic Colonoscopy Video © Database[49].

CVC-CLINIC contains 612 Standard Definition (SD) frames and comprises 31 different polyps from 31 sequences. ETIS-LARIB database contains 196 High Definition (HD) frames and comprises 44 different polyps from 34 sequences. Ground truth of each polyp consists of a polyp mask for both databases, which was generated by the annotated boundary of polyp by expert video endoscopists from the corresponding associated clinical institution. These two datasets were used for scenario of Still Frame Analysis in the challenge.

The ASU-Mayo Clinic Colonoscopy Video © Database comprises a set of short and long colonoscopy videos, collected at the Department of Gastroenterology at Mayo Clinic, Arizona. This database consists of 38 different, fully annotated videos. Ground truth consisting of binary masks (polyp frames) and black frames (non-polyp frames) were created by volunteer students at Arizona State University and have been reviewed and corrected by a trained expert. This dataset was used for scenario of Video Analysis in the challenge.

For benchmarking and evaluation, general metrics of precision, recall, specificity, F1-measure and F2-meature were used. In particular of video analysis, an additional performance metric to assess whether how fast a method detect polyp was introduced. That was Detection Latency representing the delay in frames between the first appearance of the polyp in the video sequence and the first actual detection of the polyp by a method.

## EndoTect

The EndoTect challenge[50] at the International Conference on Pattern Recognition(ICPR) 2020 aims to motivate the development of algorithms that aid medical experts in finding anomalies that commonly occur in the gastrointestinal tract[51].

A large dataset containing images taken from several endoscopies named as HyperKvasir[52] was used. In total, the dataset contains 110,079 images and 374 videos where it captures anatomical landmarks, pathological findings, and normal findings, while the dataset can be split into four distinct parts.

* Labeled Images. In total, the dataset contains 10,662 labeled images stored using the JPEG format. The labeled images represent 23 different classes of findings.
* Unlabeled Images. In total, the dataset contains 99,417 unlabeled images.
* Segmented Images. The original image, a segmentation mask and a bounding box for 1,000 images from the polyp class were provided.
* Annotated Videos. The dataset contains a total of 373 videos containing different findings and landmarks. Each video has been manually assessed by a medical professional working in the field of gastroenterology and resulted in a total of 171 annotated findings.

There were three tasks.

* Detection Task: classifying images from the GI tract into 23 distinct classes. Metrics of precision, recall/sensitivity, specificity, F1-measure and Matthews correlation coefficient (MCC) for multi-classification (also called *Rk* statistic for multiclass classification) were used and the MCC was used for benchmarking to rank the submission.
* Efficient Detection Task: efficient classification measured by the amount of time spent processing each image. Metric of a combination of the MCC classification score and the number of frames processed per second was used for benchmarking to rank the submission.
* Segmentation Task: automatically segmenting polyps. Metrics of precision, recall, the Dice coefficient, and the Intersection over Union (IoU, also known as the Jaccard index) were used and the IoU was used for benchmarking to rank the submission.

# Benchmarking by the topic group

This section describes all technical and operational details regarding the benchmarking process for the Endoscopy AI task including subsections for each version of the benchmarking that is iteratively improved over time.

## Method

The method for AI benchmarking includes data format requirement of input data and output data, training and testing data annotation quality control as well as testing metrologies and scoring matrixes.

Data annotation is currently a spontaneous non-standard process. It is a challenging task to guarantee the accuracy and representativeness of learning materials without the standardized data annotation quality control measures which are widely recognized by the industry. What’s more, this may also bring a greater risk of erroneous judgment for endoscopic assisted diagnosis.

### Input data structure

Ultrasound and endoscopic images or videos captured with gastroscope, colonoscope, laryngoscope, laryngoscope or EUS, should be submitted as separate files in the following format:

1. Endoscopic data
* Image file format: JPEG format, PNG format or BMP format.
* Image file names: be unique in the dataset and anonymize the personal information of the patient.
* Image resolution: original resolution as captured with endoscopic device, including gastroscope, colonoscope, laryngoscope, laryngoscope and EUS.
* Video file format: AVI format or MPEG-4 format.
* Video file names: be unique in the dataset and anonymize the personal information of the patient.
* Video resolution: original resolution as captured with endoscopic device, including gastroscope, colonoscope, laryngoscope, laryngoscope and EUS.
1. Ultrasound data
* Image file format: JPEG format, PNG format, BMP format or DICOM format.
* Image file names: be unique in the dataset and anonymize the personal information of the patient. For the DICOM file, anonymizing should be performed to remove sensitive information in the DICOM tag.
* Image resolution: original resolution as captured with EUS.
* Video file format: AVI format or MPEG-4 format or DICOM format.
* Video file names: be unique in the dataset and anonymize the personal information of the patient. For the DICOM file, anonymizing should be performed to remove sensitive information in the DICOM tag.
* Video resolution: original resolution as captured with EUS.

### Output data structure

The output of the algorithm should be documented in an arranged and clear way, like a CSV, XML or JSON file with the following information.

* Info of data (name, format, etc).
* Result of the data. It would depend upon the specific condition and the type of task that is being benchmarked.

#### Detection

* Data info: data name, data format, etc.
* Result info:
* Category info: the types would depend on the task.
* Location info: coordinates of a specific point (left-top or center of the bounding box) in the image. For video data, the slice index should be recorded.
* Size info: height and width in pixels.
* Task info(optional): task ID, task name, task type, etc.

#### Classification

* Data info: data name, data format, etc.
* Result info:
* Category info: the types would depend on the task.
* Task info(optional): task ID, task name, task type, etc.

#### Segmentation

* Data info: data name, data format, etc.
* Result info:
* Category info: the types would depend on the task.
* Path of segmentation file: the stored path of the segmentation file.
* Segmentation border info(optional): coordinates of points of the segmentation mask.
* Task info(optional): task ID, task name, task type, etc.

### Annotation procedure

To Guaranty the quality of data annotation and reduce individual differences among doctors, it is recommended that the annotation process should include multiple doctors’ independent annotation, cross-annotation, arbitration, and review.

If appropriate, a corresponding clinical diagnosis report or pathological report would be recommended for reference even the gold standard in data annotation.

Before annotation, the data needs preliminary filtering and laundering to eliminate worthless data, such as missing data, image parameter mismatch, non-inspection site data, foreign matter in the data, image artefacts, image quality cannot satisfy the diagnostic requirements.

#### Annotation of Detection

The annotation of detection includes localizing the object inside the data and categorizing it. The bounding box is usually used to localize the object with a rectangular box which is called a bounding box.

##### Independent annotation

Independent annotation by 2 or more doctors to confirm whether the endoscopic image/video contains lesions or intended objects and if so, mark the location and size of the lesion or intended objects with a bounding box. All the marked bounding boxes should be documented in a clear way, like a CSV file. Independent annotation requirements include:

* Non-annotating information(optional):

Image/video name, image/video identification code, collection device model, collection date, image size, collection frame rate, hospital, patient information (age, gender, race).

* Annotating information(mandatory):

Annotated results (bounding box like [*x, y, w, h, s*], where *s* is the slice index in video and equal to 0 in image), annotator information, annotation procedure information, the date, annotation serial number.

##### Cross-annotation

The independent annotations by different annotators are crossed evaluated to identify the relationship between each other by calculating the similarity, like IoU (Intersection over Union)[53]. If the similarity satisfies pre-set requirements, the independent annotations would be merged to the gold standard candidate in a specific manner, like average. Crossed annotation requirements include:

* Non-annotating information (optional):

Image/video name, image/video identification code, collection device model, collection date, image size, collection frame rate, hospital, patient information (age, gender, race).

* Annotating information (mandatory):

Cross-annotated results (bounding box like [*x, y, w, h, s*], if the pre-set requirements are not satisfied, the bounding box should be [0, 0, 0, 0, 0]), annotation serial numbers for merging, merge manner, annotation procedure information, the date, annotation serial number.

##### Arbitration

If the similarity calculated in the cross-annotation step does not satisfy the pre-set requirements, the corresponding data will be transferred to the arbitration expert to review and re-annotate as a gold standard candidate. Otherwise, the gold standard candidate in step Cross-annotation would be transferred to the review expert. Arbitration requirements include:

* Non-annotating information(optional):

Image/video name, image/video identification code, collection device model, collection date, image size, collection frame rate, hospital, patient information (age, gender, race).

* Annotating information(mandatory):

The arbitrated results (bounding box like [*x, y, w, h, s*]), annotation serial numbers for arbitration, arbitration expert information, annotation procedure information, the date, annotation serial number.

##### Review

The gold standard candidates would be confirmed by the review experts one by one. The data approved by the review expert would be marked as the gold standard. Otherwise, the data without review approval would be sent back to the arbitration procedure or modified by the review expert to generate the gold standard. Review requirements include:

* Non-annotating information(optional):

Image/video name, image/video identification code, collection device model, collection date, image size, collection frame rate, hospital, patient information (age, gender, race)

* Annotating information(mandatory):

The review results (gold standard or sent back to arbitration), serial number for review, review expert information, annotation procedure information, the date, annotation serial number.

Figure 1: Illustration of annotation procedure for detection

#### Annotation of classification

Annotation of classification means arranging a category for the data. For example, the decision of the category might be made subjectively, based on the manual observing of features in the entire or part of data. Also, the category might be made objectively, based on the corresponding objective evidence, like pathological results.

In the subjective annotation procedure, the annotation would be made manually without objective evidence.

##### Independent annotation

Independent annotation of classification by 2 or more doctors to confirm which category the data should be arranged. All the annotated results should be documented in a clear way, like a CSV file. Independent annotation requirements include:

* Non-annotating information(optional):

Image/video name, image/video identification code, collection device model, collection date, image size, collection frame rate, hospital, patient information (age, gender, race).

* Annotating information(mandatory):

Annotated results, annotator information, annotation procedure information, the date, annotation serial number.

##### Cross-annotation

The independent annotations by different annotators are crossed evaluated to identify the relationship between each other by calculating the level of consistency. If the level of consistency satisfies pre-set requirements, the independent annotations would be merged to the gold standard candidate in a specific manner, like majority rule. Crossed annotation requirements include:

* Non-annotating information (optional):

Image/video name, image/video identification code, collection device model, collection date, image size, collection frame rate, hospital, patient information (age, gender, race).

* Annotating information (mandatory):

Cross-annotated results, annotation serial numbers for merging, merge manner, annotation procedure information, the date, annotation serial number.

##### Arbitration

If the level of consistency of independent annotations does not satisfy the pre-set requirements, the corresponding data will be transferred to the arbitration expert to review and re-annotate as a gold standard candidate. Otherwise, the gold standard candidate in step Cross-annotation would be transferred to the review expert. Arbitration requirements include:

* Non-annotating information(optional):

Image/video name, image/video identification code, collection device model, collection date, image size, collection frame rate, hospital, patient information (age, gender, race).

* Annotating information(mandatory):

The arbitrated results, annotation serial numbers for arbitration, arbitration expert information, annotation procedure information, the date, annotation serial number.

##### Review

The gold standard candidates would be confirmed by the review experts one by one. The data approved by the review expert would be marked as the gold standard. Otherwise, the data without review approval would be sent back to the arbitration procedure or modified by the review expert to generate the gold standard. Review requirements include:

* Non-annotating information(optional):

Image/video name, image/video identification code, collection device model, collection date, image size, collection frame rate, hospital, patient information (age, gender, race)

* Annotating information(mandatory):

The review results (gold standard or sent back to arbitration), serial number for review, review expert information, annotation procedure information, the date, annotation serial number.



Figure 2: Illustration of annotation procedure for classification

#### Annotation of segmentation

Annotation of segmentation means the annotation of every pixel in an object within a data. Practically, a polygon is used to annotate the contour of the object, which will be filled for a mask.

Segmentation annotations are usually done on the basis of detection annotations. Usually, the segmentation annotation is independently outlined by three doctors, and the overlapping area of more than two doctors is taken as the gold standard segmentation annotations. Because segmentation annotation is very time-consuming, especially if there are multiple images in a subject that need to be annotated. In order to improve the efficiency of segmentation annotation, it is sometimes allowed to sketch the contour of the object by low-level doctors, and confirm or modify by high-level experts to form the gold standard segmentation annotations.

[TBC] [call for contribution]

## Score and metrics

### Evaluation score and metrics

This section focuses on the benchmarking processes in the context of AI for endoscopy for quality assessment. It addresses different aspects of benchmarking in AI systems (e.g., relevant scientific publications, benchmarking frameworks, scores and metrics, and clinical evaluation attempts). The goal is to collect all relevant learnings from previous benchmarking that could help to implement the benchmarking process in this topic group.

Table 3: Benchmarking metrics

| Methodology | Description |
| --- | --- |
| True positive (TP) | The number of correctly identified positive samples. The number of frame with endoscopic finding which correctly is identified as a frame with an endoscopic finding. |
| True negative (TN) | The number of correctly identified negative samples, i.e., the number of frame without endoscopic finding which correctly is identified as a frame without endoscopic finding. |
| False-positive (FP) | The number of wrongly identified positive samples, i.e., a commonly called a "false alarm". The number of frames without endoscopic finding which is erroneously identified as a frame with endoscopic finding. |
| False-negative (FN) | The number of wrongly identified negative samples. The number of frame with an endoscopic finding which erroneously are identified as a frame without an endoscopic finding. |
| Recall (REC) orSensitivity (SENS) | This metric is also frequently called sensitivity, probability of detection and true positive rate, and it is the ratio of samples that are correctly identified as positive among all existing positive samples. |
| Precision (PREC) | This metric is also frequently called the positive predictive value. It shows the ratio of samples that are correctly identified as positive among the returned samples (the fraction of retrieved samples that are relevant). |
| Specificity (SPEC) | This metric is frequently called the true negative rate. It shows the ratio of negatives that are correctly identified as such (e.g., the fraction of frames without an endoscopic finding are correctly identified as a negative result). |
| Accuracy (ACC) | The percentage of correctly identified true and false samples. |
| Matthews correlation coefficient (MCC) | MCC takes into account true and false positives and negatives. It is a balanced measure even if the classes are of very different sizes. |
| F1 score (F1) | A measure of a test's accuracy by calculating the harmonic mean of the precision and recall. |
| DICE coefficient (DICE) | This metric measures the similarity between two sets of data and is most broadly used in the validation of image [segmentation](https://radiopaedia.org/articles/segmentation?lang=us). It equals twice the number of elements common to both sets divided by the sum of the number of elements in each set. |
| Jaccard coefficient orIntersection over Union (IoU) | This metric measures the similarity between two sets of data and is most broadly used in the validation of object detection and image segmentation. It equals the number of elements common to both sets divided by the sum of the number of unique elements in each set. |
| Polyp detection rate (PDR) | This metric is the percentage of patients undergoing screening endoscopy who have one or more polyp detected. |
| Adenoma detection rate (ADR) | This metric is the percentage of patients undergoing screening endoscopy who have one or more conventional adenomas detected. |
| Detection Latency | This metric is the delay in frames between the first appearance of the polyp/lesion/object in the video sequence and the first actual detection of the polyp/lesion/object by a method. |
| Average precision(AP) | This metric is the average precision in the P-R curve. |
| mAP | This metric is the average value of AP of every classe. |
| Runtime | This metric is the cost time of running one image or frame by a method. |

### Other quality score and metrics criteria

#### Available public data and undisclosed test data set collection

Table 4: Available public data

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Findings** | **Size** | **availability** |
| CVC-612(CVC-ClinicDB)[48] | Polyp,with mask | 612 images | open academic |
| ASU-Mayo polyp database[13] | Polyp,with mask | 18,781 images | by request (no available anynore) |
| ETIS-Larib Polyp DB[54] | Polyp,with mask | 196 images | open academic |
| KID[55] | Angiectasia, bleeding, inflammations, polyp | 2371 images,47 videos | open academic (no available anynore) |
| GIANA’17[42] | Angiectasia,with mask | 600 images | by request  |
| GASTROLAB[56] | GI lesions | Some 100s of images + few videos | open academic (video capsule endoscopy) |
| WEO Clinical Endoscopy Atlas[57] | GI lesions | 152 images | by request (video capsule endoscopy) |
| GI Lesions in Regular Colonoscopy Data Set[58] | GI lesions,with mask | 76 images | by request  |
| Atlas of Gastrointestinal Endoscope[59] | GI lesions | 1295 images | unknown (no available anynore) |
| GastroAtlas[60] | GI lesions | 5,071 video clips | open academic (video capsule endoscopy) |
| Kvasir[61] | Polyps, esophagitis, ulcerative colitis, Z-line, pylorus, cecum,dyed polyp, dyed resection margins, stool | 8,000 images | open academic  |
| Nerthus[62] | Stool - categorization of bowel cleanliness | 21 videos | open academic  |
| Kvasir-SEG[63] | Polyps,with mask | 1000 images | open academic |
| HyperKvasir[52] | GI findings including polyps | 110,079 images and 374 videos | open academic |
| Kvasir-Capsule[64] | GI findings including polyps(video capsule endoscopy) | 4,741,504 images | open academic |
| CVC-ColonDB[65] | Polyps,with mask | 380 images | open academic (no available anynore) |
| EDD2020[39] | GI lesions including polyps | 386 images | open academic |

#### Data sharing policies

This section provides details about legalities in the context of benchmarking. Each dataset that is shared should be protected by special agreements or contracts that cover, for instance, the data sharing period, patient consent, and update procedure (see also [DEL05\_5](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B71FE8B9D-ACB3-48CE-AA3F-136409B550A4%7D&file=DEL05_5.docx&action=default) on *data handling* and [DEL05\_6](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B5C95327E-96A5-4175-999E-3EDB3ED147C3%7D&file=DEL05_6.docx&action=default) on *data sharing practices*).

[TBC]

#### Baseline acquisition

The main purpose of benchmarking is to provide stakeholders with the numbers they need to decide whether AI models provide a viable solution for a given health problem in a designated context. To achieve this, the performance of the AI models needs to be compared with available options achieving the same clinically meaningful endpoint. This, in turn, requires data on the performance of the alternatives, ideally using the same benchmarking data. As the current alternatives typically involve doctors, it might make sense to combine the test data acquisition and labelling with additional tasks that allow the performance of the different types of health workers to be assessed.

[TBC]

#### Reporting methodology

This section discusses how the results of the benchmarking runs will be shared with the participants, stakeholders, and general public.

[TBC]

#### Result

This section gives an overview of the results from runs of this benchmarking version of your topic. Even if your topic group prefers an interactive drill-down rather than a leader board, pick some context of common interest to give some examples.

[TBC]

#### Discussion of the benchmarking

This section discusses insights into these benchmarking iterations and provides details about the ‘outcome’ of the benchmarking process (e.g., giving an overview of the benchmark results and process).

[TBC]

#### Retirement

This section addresses what happens to the AI system and data after the benchmarking activity is completed. It might be desirable to keep the database for traceability and future use. Alternatively, there may be security or privacy reasons for deleting the data. Further details can be found in the reference document of this section [DEL04](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BC68833D1-9B31-4E8E-8A4A-3939D7DEA56F%7D&file=DEL04.docx&action=default) “*AI software lifecycle specification”* (identification of standards and best practices that are relevant for the AI for health software life cycle).

[TBC]

# Overall discussion of the benchmarking

This section discusses the overall insights gained from benchmarking work in this topic group. This should not be confused with the discussion of the results of a concrete benchmarking run (e.g., in 4.2.2.6).

# Regulatory considerations

For AI-based technologies in healthcare, regulation is not only crucial to ensure the safety of patients and users, but also to accomplish market acceptance of these devices. This is challenging because there is a lack of universally accepted regulatory policies and guidelines for AI-based medical devices. To ensure that the benchmarking procedures and validation principles of FG-AI4H are secure and relevant for regulators and other stakeholders, the working group on *“*[*Regulatory considerations on AI for health”* *(WG-RC)*](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/wg/SitePages/WG-RC.aspx) compiled the requirements that consider these challenges.

The deliverables with relevance for regulatory considerations are [DEL02](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BF2F46A99-7457-4BC8-81A3-0E1E63D6072A%7D&file=DEL02.docx&action=default) *“AI4H regulatory considerations”* (which provides an educational overview of some key regulatory considerations), [DEL02\_1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B6AF7C004-8BCE-4151-9F44-45F041A1EB1D%7D&file=DEL02_1.docx&action=default) *“Mapping of IMDRF essential principles to AI for health software”,* and[DEL02\_2](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B1ED0D4D1-876C-4A0F-AEF7-06D3F445F5E6%7D&file=DEL02_2.docx&action=default) *“Guidelines for AI-based medical device (AI-MD): Regulatory requirements”* (which provides a checklist to understand expectations of regulators, promotes the step-by-step implementation of safety and effectiveness of AI-based medical devices, and compensates for the lack of a harmonized standard). [DEL04](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BC68833D1-9B31-4E8E-8A4A-3939D7DEA56F%7D&file=DEL04.docx&action=default) identifies standards and best practices that are relevant for the “*AI software lifecycle specification*.*”* The following sections discuss how the different regulatory aspects related to TG-Endoscopy.

## Existing applicable regulatory frameworks

Most of the AI systems that are part of the FG-AI4H benchmarking process can be classified as *software as a medical device* (SaMD) and eligible for a multitude of regulatory frameworks that are already in place. In addition, these AI systems often process sensitive personal health information that is controlled by another set of regulatory frameworks. The following section summarizes the most important aspects that AI manufacturers need to address if they are developing AI systems for endoscopy.

## Regulatory features to be reported by benchmarking participants

In most countries, benchmarked AI solutions can only be used legally if they comply with the respective regulatory frameworks for the application context. This section outlines the compliance features and certifications that the benchmarking participants need to provide as part of the metadata. It facilitates screening of the AI benchmarking results for special requirements (e.g., the prediction of prediabetes in a certain subpopulation in a country compliant to the particular regional regulatory requirements).

## Regulatory requirements for the benchmarking systems

The benchmarking system itself needs to comply with regulatory frameworks (e.g., some regulatory frameworks explicitly require that all tools in the quality management are also implemented with a quality management system in place). This section outlines the regulatory requirements for software used for benchmarking in this topic group.

## Regulatory approach for the topic group

Building on the outlined regulatory requirements, this section describes how the topic group plans to address the relevant points to be compliant. The discussion here focuses on the guidance and best practices provided by the [DEL02](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BF2F46A99-7457-4BC8-81A3-0E1E63D6072A%7D&file=DEL02.docx&action=default) *“AI4H regulatory considerations.”*

# References

1. Reddy, Yogananda, and Robert P. Willert. "Endoscopic ultrasound: what is it and when should it be used?." Clinical medicine 9.6 (2009): 539-543.
2. Mertz, Howard R., et al. "EUS, PET, and CT scanning for evaluation of pancreatic adenocarcinoma." Gastrointestinal endoscopy 52.3 (2000): 367-371.
3. Wallace, Michael B., et al. "Minimally invasive endoscopic staging of suspected lung cancer." Jama 299.5 (2008): 540-546.
4. Kuwahara, Takamichi, et al. "Current status of artificial intelligence analysis for endoscopic ultrasonography." Digestive Endoscopy 33.2 (2021): 298-305.
5. De Groen, Piet C., Yi-Jhen Li, and Sudha Xirasagar. "Long-term colorectal-cancer mortality after adenoma removal." The New England journal of medicine 371.21 (2014): 2035.
6. Byrne, Michael F., et al. "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 (2019): 94-100.
7. Chen, Peng-Jen, et al. "Accurate classification of diminutive colorectal polyps using computer-aided analysis." Gastroenterology 154.3 (2018): 568-575.
8. Kominami, Yoko, et al. "Computer-aided diagnosis of colorectal polyp histology by using a real-time image recognition system and narrow-band imaging magnifying colonoscopy." Gastrointestinal endoscopy 83.3 (2016): 643-649.
9. Mori, Yuichi, and Shin-ei Kudo. "Detecting colorectal polyps via machine learning." Nature biomedical engineering 2.10 (2018): 713-714.
10. Misawa, Masashi, et al. "Artificial intelligence-assisted polyp detection for colonoscopy: initial experience." Gastroenterology 154.8 (2018): 2027-2029.
11. Zhang, Ruikai, et al. "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 (2016): 41-47.
12. Wang, Pu, et al. "Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy." Nature biomedical engineering 2.10 (2018): 741-748.
13. Yu, Lequan, et al. "Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos." IEEE journal of biomedical and health informatics 21.1 (2016): 65-75.
14. Bernal, Jorge, et al. "Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 endoscopic vision challenge." IEEE transactions on medical imaging 36.6 (2017): 1231-1249.
15. Săftoiu, Adrian, et al. "Role of gastrointestinal endoscopy in the screening of digestive tract cancers in Europe: European Society of Gastrointestinal Endoscopy (ESGE) Position Statement." Endoscopy 52.04 (2020): 293-304.
16. Dumonceau, J-M., et al. "Indications, results, and clinical impact of endoscopic ultrasound (EUS)-guided sampling in gastroenterology: European Society of Gastrointestinal Endoscopy (ESGE) Clinical Guideline." Endoscopy 43.10 (2011): 897-912.
17. Van Der Merwe, Schalk W., et al. "Therapeutic endoscopic ultrasound: European Society of Gastrointestinal Endoscopy (ESGE) Guideline." Endoscopy (2021).
18. Forbes, Nauzer, et al. "Adverse events associated with EUS and EUS-guided procedures." Gastrointestinal endoscopy 95.1 (2022): 16-26.
19. Su, Jing-Ran, et al. "Impact of a real-time automatic quality control system on colorectal polyp and adenoma detection: a prospective randomized controlled study (with videos)." Gastrointestinal endoscopy 91.2 (2020): 415-424.
20. Aslanian, Harry R., et al. "Nurse observation during colonoscopy increases polyp detection: a randomized prospective study." Official journal of the American College of Gastroenterology| ACG 108.2 (2013): 166-172.
21. Wang, Pu, et al. "Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study." Gut 68.10 (2019): 1813-1819.
22. Zhao, Sheng-Bing, et al. "Establishment and validation of a computer-assisted colonic polyp localization system based on deep learning." World Journal of Gastroenterology 27.31 (2021): 5232.
23. Valero, Manuel, and Carlos Robles-Medranda. "Endoscopic ultrasound in oncology: An update of clinical applications in the gastrointestinal tract." World journal of gastrointestinal endoscopy 9.6 (2017): 243.
24. Kelly S, Harris KM, Berry E, Hutton J, Roderick P, Cullingworth J, Gathercole L, Smith MA. "A systematic review of the staging performance of endoscopic ultrasound in gastro-oesophageal carcinoma". Gut 2001; 49: 534-539.
25. Japan Science and Technology Agency, Immediate detection of colorectal cancer with AI. https://www.jst.go.jp/EN/achievements/research/bt2019-07.html, 2019.
26. Luo, Huiyan, et al. "Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study." The Lancet Oncology 20.12 (2019): 1645-1654.
27. Zhu, Yan, et al. "Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy." Gastrointestinal endoscopy 89.4 (2019): 806-815.
28. Hirasawa, Toshiaki, et al. "Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images." Gastric Cancer 21.4 (2018): 653-660.
29. Wu, Lianlian, et al. "A deep neural network improves endoscopic detection of early gastric cancer without blind spots." Endoscopy 51.06 (2019): 522-531.
30. Komeda, Yoriaki, et al. "Computer-aided diagnosis based on convolutional neural network system for colorectal polyp classification: preliminary experience." Oncology 93.Suppl. 1 (2017): 30-34.
31. EndoCV2022: Endoscopic computer vision challenges 2.0, <https://endocv2022.grand-challenge.org/EndoCV-Sequence/>, 2022.
32. EndoCV2021: Polyp Detection and Segmentation: Addressing Generalisability, <https://endocv2021.grand-challenge.org/>, 2021.
33. EndoCV2020, <https://endocv.grand-challenge.org/>, 2020.
34. EndoCV2020: EAD2020, <https://ead2020.grand-challenge.org/>, 2020.
35. EndoCV2020: EDD2020, <https://edd2020.grand-challenge.org/>, 2020.
36. EAD2019, <https://ead2019.grand-challenge.org/>, 2019.
37. Ali, Sharib, and Noha Ghatwary. "Endoscopic computer vision challenges 2.0." (2022)
38. Ali, Sharib, et al. "PolypGen: A multi-center polyp detection and segmentation dataset for generalisability assessment." arXiv preprint arXiv:2106.04463 (2021).
39. Ali, Sharib, et al. "Endoscopy disease detection challenge 2020." arXiv preprint arXiv:2003.03376 (2020).
40. Ali, Sharib, et al. "Endoscopy artifact detection (EAD 2019) challenge dataset." arXiv preprint arXiv:1905.03209 (2019)
41. EndoVis2022, <https://endovis.grand-challenge.org/Endoscopic_Vision_Challenge/>, 2022.
42. EndoVis2017: GIANA 2017, <https://endovissub2017-giana.grand-challenge.org/>, 2017.
43. EndoVis2021: GIANA 2017, <https://giana.grand-challenge.org/>, 2021.
44. EndoVis2020: CATARACTS Semantic Segmentation 2020, <https://cataracts-semantic-segmentation2020.grand-challenge.org/Home/>, 2020.
45. Luengo, Imanol, et al. "2020 CATARACTS Semantic Segmentation Challenge." arXiv preprint arXiv:2110.10965, 2021.
46. EndoVis2015: Detection of abnormalities in gastroscopic images, <https://endovissub-abnormal.grand-challenge.org/>, 2015.
47. EndoVis2015: Automatic Polyp Detection in Colonoscopy Videos, <https://polyp.grand-challenge.org/>, 2015.
48. Bernal, Jorge, et al. "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians." Computerized medical imaging and graphics 43 (2015): 99-111.
49. Tajbakhsh, Nima, Suryakanth R. Gurudu, and Jianming Liang. "Automated polyp detection in colonoscopy videos using shape and context information." IEEE transactions on medical imaging 35.2 (2015): 630-644.
50. EndoTech2021, <https://endotect.com/>, 2021.
51. Hicks, Steven A., et al. "The EndoTect 2020 challenge: evaluation and comparison of classification, segmentation and inference time for endoscopy." International Conference on Pattern Recognition. Springer, Cham, 2021.
52. Borgli, Hanna, et al. "HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy." Scientific data 7.1 (2020): 1-14.
53. Intersection over Union (IoU) for object detection, <https://pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/>, 2005.
54. Silva, Juan, et al. "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer." International journal of computer assisted radiology and surgery 9.2 (2014): 283-293.
55. Koulaouzidis, Anastasios, et al. "KID Project: an internet-based digital video atlas of capsule endoscopy for research purposes." Endoscopy international open 5.06 (2017): E477-E483
56. Gastrolab - the gastrointestinal site. http://www.gastrolab.net/index.htm.
57. Weo clinical endoscopy atlas. http://www.endoatlas.org/index.php.
58. Gastrointestinal lesions in regular colonoscopy dataset. http://www.depeca.uah.es/colonoscopy\_dataset/.
59. The atlas of gastrointestinal endoscope. http://www.endoatlas.com/atlas\_1.html.
60. El salvador atlas of gastrointestinal video endoscopy. http://www.gastrointestinalatlas.com/index.html.
61. Pogorelov, Konstantin, et al. "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection." Proceedings of the 8th ACM on Multimedia Systems Conference. 2017
62. Pogorelov, Konstantin, et al. "Nerthus: A bowel preparation quality video dataset." Proceedings of the 8th ACM on Multimedia Systems Conference. 2017
63. Jha, Debesh, et al. "Kvasir-seg: A segmented polyp dataset." International Conference on Multimedia Modeling. Springer, Cham, 2020
64. Smedsrud, Pia H., et al. "Kvasir-Capsule, a video capsule endoscopy dataset." Scientific Data 8.1 (2021): 1-10.
65. Bernal, Jorge, Javier Sánchez, and Fernando Vilarino. "Towards automatic polyp detection with a polyp appearance model." Pattern Recognition 45.9 (2012): 3166-3182

Annex A:
Glossary

This section lists all the relevant abbreviations, acronyms and uncommon terms used in the document.

|  |  |  |
| --- | --- | --- |
| Acronym/Term | Expansion | Comment |
| TDD | Topic Description Document | A document specifying the standardized benchmarking for a topic on which the FG AI4H Topic Group works. This document is the TDD for the Topic Group [YOUR TOPIC GROUP] |
| TG | Topic Group |  |
| WG | Working Group |  |
| FGAI4H | Focus Group on AI for Health |  |
| AI | Artificial intelligence |  |
| ITU | International Telecommunication Union |  |
| WHO | World Health Organization |  |
| DEL | Deliverable  |  |
| CfTGP | Call for topic group participation |  |
| AI4H  | Artificial intelligence for health |  |
| IMDRF | International Medical Device Regulators Forum |  |
| MDR | Medical Device Regulation |  |
| ISO | International Standardization Organization |  |
| GDPR | General Data Protection Regulation |  |
| FDA | Food and Drug administration |  |
| SaMD | Software as a medical device |  |
| AI-MD | AI-based medical device |  |
| LMIC | Low-and middle-income countries |  |
| GDP | Gross domestic product |  |
| API | Application programming interface |  |
| IP | Intellectual property |  |
| PII | Personally identifiable information |  |
| […] |  |  |

Annex B:
Declaration of conflict of interests

In accordance with the ITU rules in this section working on this document should define his conflicts of interest that could potentially bias his point of view and the work on this document.

Tencent Healthcare (Shenzhen) Co., Ltd

Tencent Healthcare is committed to play a good role as a digital assistant in healthcare scenario, by integrating the digital capabilities of Tencent and its partners to build series of practical products, including Tencent AIMIS, medical insurance payment by Wechat and electronic health card. By opening up Tencent's advanced technologies and services such as artificial intelligence, big data, cloud computing, etc., to realize the combination of advanced technology and healthcare.

China Academy of Information and Communications Technology

[TBC]

Olympus Medical Systems Corp.

[TBC]

Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences

[TBC]

China Unicom (Guangdong) Industrial Internet Co., Ltd

[TBC]

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_