International Telecommunication Union

ITU-T FG-AI4H Deliverable

TELECOMMUNICATION
STANDARDIZATIONSECTOR
OFITU

15 September 2023

PRE-PUBLISHED VERSION

DEL10.20

FG-AI4H Topic Description Document for the Topic Group on AI for endoscopy (TG-Endoscopy)

Summary

This document describes the topic of AI for endoscopy, including two subtopics as colonoscopy and endoscopic ultrasound. Endoscopy is the core technical means for early diagnosis and screening of digestive cancer, while AI solution for endoscopy is expected to help clinicians improve their examination quality and reduction of missed diagnoses. The document is committed to give a general description on AI for endoscopy, including AI tasks, existing AI solutions, data annotation process, existing benchmarking and regulatory consideration, etc.

Keywords

Artificial intelligence; health; topic groups; overview; ethics; regulations; data quality; data audit; clinical relevance; topic description; endoscopy; colonoscopy; endoscopic ultrasound; data annotation; benchmarking; gold standard

Change Log

This document contains Version 1 of the Deliverable DEL10.20 on "*FG-AI4H Topic Description Document for the Topic Group on AI for endoscopy (TG-Endoscopy)*" approved on 15 September 2023 via the online approval process for the ITU-T Focus Group on AI for Health (FG-AI4H).

Editors:

Jianrong Wu Tencent Healthcare (Shenzhen), China	Email: edwinjrwu@tencent.com
Shan Xu CAICT, China	Email: xushan@caict.ac.cn

Contributors:

Yajun Zhang Tencent Technology (Shenzhen), China	Email: yajunzhang@tencent.com
Yi Cai Olympus Medical Systems Corp, Japan	Email: i.sai@olympus.com
Junbo Li Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, China	Email: Lijb@sibet.ac.cn
Yanchun Zhu China Unicom (Guangdong) Industrial Internet Co., Ltd, China	Email: zhuyc82@chinaunicom.cn

CONTENTS

Page

List of Tables

List of Figures

ITU-T FG-AI4H Deliverable 10.20

FG-AI4H Topic Description Document for the Topic Group on AI for endoscopy (TG-Endoscopy)

1 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 AI for Endoscopy systems. It serves as deliverable No. DEL 10.20 of the ITU/WHO Focus Group on AI for Health (FG-AI4H).

2 About the FG-AI4H topic group on AI for Endoscopy

The introduction highlights the potential of a standardized benchmarking of AI systems for AI for Endoscopy to help solving 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 FG-AI4H meeting I, e-meeting, 7-8 May 2020, Dr. Jianrong Wu from Tencent Healthcare was nominated as topic driver for the TG-Endoscopy.

2.1 Documentation

This document is the TDD for the TG-Endoscopy. It introduces the health topic including the AI task, outlines its relevance and the potential impact that the benchmarking will have on the health system and patient outcome, and provides an overview of the existing AI solutions for Endoscopy. It describes the existing approaches for assessing the quality of AI for Endoscopy systems and provides the details that are likely relevant for setting up a new standardized benchmarking. It specifies the actual benchmarking methods for all subtopics at a level of detail that includes technological and operational implementation. There are individual subsections for all versions of the benchmarking. Finally, it summarizes the results of the topic group's benchmarking initiative and benchmarking runs. In addition, the TDD addresses ethical and regulatory aspects.

The TDD 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 AI for 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-S-025-A01	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

Select the following link:

- https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/endoscopy/[draft]_FGAI4H-P-025-A01_clean.docx

2.2 Status of this topic group

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

2.2.1 Status update for meeting S

- Rearrange TDD following FG-AI4H-J-105
- Update the TDD for AI for endoscopy

2.2.2 Status update for meeting P

- Update the TDD for AI for endoscopy

2.2.3 Status update for meeting N

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

2.2.4 Status update for meeting M

- Update the TDD for AI for endoscopy
  - Add new subtopic as endoscopic ultrasound
- Invite new participants

2.2.5 Status update for meeting L

- Transform the initial document of TG-Endoscopy into TDD template format
  - Invite new participants

2.2.6 Status update for meeting K

- Updates the initial documents of AI for endoscopy
  - Invite new participants

2.2.7 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)

2.2.8 Status update for meeting I

- Discussed the proposal from Tencent Healthcare
- 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

2.3 Topic Group participation

The 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

3 Topic description

This section contains a detailed description and background information of the specific health topic for the benchmarking of AI in AI for Endoscopy and how this can help to solve a relevant 'real-world' problem.

Topic Groups summarize related benchmarking AI subjects to reduce redundancy, leverage synergies, and streamline FG-AI4H meetings. However, in some cases different subtopic groups can be established within one topic group to pursue different topic-specific fields of expertise. The TG-Endoscopy has two subtopics, including colonoscopy and endoscopic ultrasound.

3.1 Subtopic Colonoscopy

3.1.1 Definition of the AI task

This section provides a detailed description of the specific task. The AI systems of this TG are expected to solve. This section corresponds to DEL03 *"AI requirements specifications*," which describes the functional, behavioural, and operational aspects of an AI system.

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

3.1.1.1 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 clinical applications, it is possible to classify colorectal polyps in endoscopic image from a patient into different categories, such as non-adenomatous polyps, adenomatous polyps and cancerous. Different categories would need specific treatments. And by the assistance of classification result, clinician could make more accurate diagnosis. It is also possible to evaluate the image quality of all endoscopic images of the patient, like categorization of bowel cleanliness, from which the quality of the image should meet the diagnostic quality requirements.

3.1.1.2 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. Generally, bounding boxes are used to distinct objects in video frames or images.

In clinical applications, it is possible to detect different findings in colonoscopy for different purposes, such as polyps, angiectasia, bleeding, inflammations, esophagitis, ulcerative colitis, pylorus, cecum, dyed polyp, dyed resection margins and stool. Specifically, polyp detection is the most usual AI application in colonoscopy. Detection for polyps can effectively reduce the polyp missing rate in colorectal screening, which would further reduce the adenomatous missing rate.

3.1.1.3 Segmentation

Image segmentation separates an image into regions on 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 segmentation is to highlight foreground elements and make it easier to be evaluated. Image segmentation provides pixel-by-pixel details of an object, distinguishing it from classification and object detection.

For example, in endoscopy, the polyp size can be automatically calculated based on the segmentation results, while the polyp size is one of the key factors for polyp diagnosis by clinician.

3.1.2 **Current gold standard**

This section provides a description of the established gold standard of the addressed health topic.

Colorectal cancers (CRC) are the third most prevalent cancer and the second-highest cause of cancer deaths worldwide. Colonoscopy is considered the gold standard for CRC screening to detect and remove the polyps and adenomas in the colorectum [1]. 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.

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 [2][3][4]. By subjective methods, the gold standard would be made by colonoscopists manually [5]. Usually, subjective method is used for polyp detection [6][7] and segmentation tasks [8][9][10], whose annotation results are bounding-box and mask respectively. The subjective method might involve one or multiple colonoscopists in single-step or multi-steps procedures of annotation.

3.1.3 Relevance and impact of an AI solution

This section addresses the relevance and impact of the AI solution and describes how solving the task with AI improves a health issue.

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 [11]. 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 [12].

A prospective study including 1058 patients was designed as a randomized controlled trial (RCT) 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, $p<0.001$) [13].

Over 71000 images from 20 centers were 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) [14].

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% [2].

3.1.4 Existing AI solutions

This section provides an overview of existing AI solutions for colonoscopy that are already in operation. AI solutions for colonoscopy are moving towards commercialization and clinical practice, while there are more and more CADe products approved by Chinese National Medical Products Administration (NMPA), U.S. Food and Drug Administration (FDA) and European CE.

- 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 [15]. The system detects potential polyps 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. [14]. The real-time polyp detection system by Tencent Healthcare has been certificated by Chinese NMPA in June, 2023.
- 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. [16]

- Wision AI and Sichuan Provincial People's Hospital developed a real-time automatic polyp detection system in 2018 [8] 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 [13][17]. 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. [18]
- 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. [19]
- 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 [20].
- 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 [21].
- 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 [22].
- Kindai University developed a system in 2017 that could diagnose colon polyps as adenomatous or non-adenomatous using a simple CNN [23].
- Wuhan ENDOANGEL Medical Technology Co., LTD developed an AI system called EndoAngel®, consisting of polyp detection and quality monitoring functions. The polyp detection function can remind endoscopists of the location of the polyp. The quality monitoring function can monitor the velocity of insertion of the endoscope, record the time of insertion and withdrawal of the endoscope, and remind endoscopists of the blind areas caused by intestinal segment slipping [24]. It has been certificated by Chinese NMPA in May, 2023.

3.2 Subtopic on endoscopic ultrasound

3.2.1 Definition of the AI task

Endoscopic ultrasound (EUS) is a minimally invasive procedure in which endoscopy is combined with ultrasound to obtain images of the internal organs. Comparing with colonoscopy, EUS is a multi-modality procedure capturing ultrasound and image at the same time. In general, the AI task with EUS is mainly divided into classification, detection, and segmentation.

3.2.1.1 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.

As EUS is frequently used in the assessment of digestive disease, the clinical applications of AI for EUS largely involve their use in classifying suspicious lesions in upper and lower digestive tract and surrounding tissues from endoscopic images and ultrasound data (RF, B-mode, color flow, contrast enhance ultrasound, elastography, etc.). In addition, AI assisted EUS would be used in respiratory and urinary system to distinguish malignant from benign lesions. By the assistance of classification result, clinician could make more accurate diagnosis, reduce un-necessary EUS guided biopsy and provide more suitable treatment. It is also possible to evaluate the image quality of EUS images of the patient, like station classification and quality assessment for pancreatis EUS scan.

3.2.1.2 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. Generally, bounding boxes are used to distinct objects in video frames or images.

Unlike conventional endoscopy, where AI assisted detection is possibly used to avoid missing blind spots during the procedure, EUS can hardly be used as the first-line screening choice for digestive or respiratory tract due to its limited image quality for endoscopic imaging. Instead, it is possible to different findings beneath or surrounding digestive and respiratory tract in EUS for different purposes, such as lymph nodes, bleeding and inflammations.

3.2.1.3 Segmentation

Image segmentation separates an image into regions on 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 segmentation is to highlight foreground elements and make it easier to be evaluated. Image segmentation provides pixel-by-pixel details of an object, distinguishing it from classification and object detection.

In EUS, size and extension of the suspicious lesion can be automatically calculated based on the segmentation results, which can help clinicians to provide more suitable treatment.

3.2.2 Current gold standard

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 [25][26]. 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, and it is most commonly used for staging of gastrointestinal (GI) malignancies, evaluating pancreaticobiliary disease, evaluating subepithelial abnormalities, evaluating extraluminal abnormalities, staging of lung cancer and image guidance for therapeutic procedures [27]. European Society of Gastrointestinal Endoscopy (ESGE) has suggested EUS for pancreatic cancer screening in selected high-risk patients, 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., and recommended EUS as therapeutic procedures over various types of diseases, including percutaneous transhepatic biliary drainage (PTBD), pancreatic duct (PD) drainage, etc. Specifically, EUS is capable to identify small pancreatic tumours with a staging sensitivity greater than 90% [28], and endobronchial ultrasound (EBUS) has been used for lung cancer staging with a diagnostic accuracy of 90–100% [29].

Research on artificial intelligence (AI) in EUS is still limited [30][31][32][33][34]. Only a handful of reports were published based on limited clinical data through retrospective or prospective studies, with a main focus on pancreatic diseases. Currently, there's no commercial AI product for EUS on the market.

3.2.3 Relevance and impact of an AI solution

EUS has been proven to be an effective imaging modality for local-reginal staging of gastrointestinal tumours. The diagnostic ability of EUS is higher than that of computed tomography (CT), transabdominal ultrasonography, and magnetic resonance imaging (MRI) [35][36]. It has also proved to be a useful alternative therapeutic modality in surgery. However, EUS may be less accurate for early staging of oesophageal 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%) [36]. 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 tumour staging could be found between different studies (63% to 95%).

AI is believed to play an 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 workload and standardize the overall quality of endoscopic procedures.

3.2.4 Existing AI solutions

This section provides an overview of existing AI solutions for EUS that are already in operation. It should be noted that currently there's no well-accepted AI solution for EUS. The solution listed below are mainly premature prototype system or even models from academic or industrial research.

- Researchers from Pusan National University Hospital, Silla University, Asan Medical Center and Yonsei University College of Medicine developed a CNN-CAD system to analyse gastric mesenchymal tumours on EUS images. The CNN-CAD system can differentiate GISTs from non-GIST tumours within a short amount of time and with high sensitivity and specificity. However, the dataset used in the study were relatively small and only high-quality EUS images were selected for the training and test datasets. [37]
- Researchers from Changhai Hospital reported a single-centre retrospective study in 2010. SVM-based classification was implemented to differentiate pancreatic cancer from normal tissue with high accuracy, sensitivity, and specificity [38]. Further study was reported in 2013 with more data from Changhai Hospital [39]. Results show the superiority of SVM based CAD system for pancreas EUS. However, substantial decrease in classification performance can be found between the two studies using data from the same clinical site and very similar technology.
- Tokyo Medical University developed a CNN-based EUS-CAD system and assessed its ability to detect pancreatic ductal carcinoma (PDAC), using control images from patients with chronic pancreatitis (CP) and those with a normal pancreas (NP). Results indicates EUS-CAD system can work not only in assisting the training of beginners of EUS instead of an instructor, but also in supporting fatigued experts or carelessness caused by performing a large number of screening examinations. [40]
- The European EUS Elastography Multicentric Study Group performed a prospective multicentric study in 2012 and develop an ANN-based CAD to differentiate benign from malignant pancreatic lesions using real-time EUS elastography [41]. In 2015, another prospective multicentric study were conducted to access ANN-based CAD to classify pancreatic cancer using dynamic contrast-enhanced EUS [42]. Results from two studies suggest that integration of clinical data into efficacious ANNs, in concordance with imaging

enhancements (real-time sono-elastography, contrast-enhancement, hybrid imaging, 3-dimensional imaging, and so forth) and cytologic parameters, would certainly be beneficial for improved clinical decision making in patients with focal pancreatic lesions.

- Wuhan EndoAngel Medical Technology Company, Renmin Hospital of Wuhan University, Wuhan Union Hospital and Wuhan Puai Hospital developed a BP MASTER (pancreaticobiliary master) system for training and quality control of pancreatis EUS scan. Results show the BP MASTER system has potential to play an important role in shortening the pancreatic EUS learning curve and improving EUS quality control in the future. [43]
- China Medical University Hospital and National Taiwan University developed a CNN-based CAD to classify lung lesions using endobronchial ultrasound images (EBUS). The results showed that the fusion of the fine-tuned CaffeNet and SVM system have the potential to assist lung cancer detection. [44]
- Shimane University of Japan developed an EBUS-computer-aided diagnosis system using CNN to differentiate benign from malignant lesions based on EBUS findings. The developed EBUS-computer-aided diagnosis system is capable to read EBUS findings that are difficult for clinicians to judge with precision and help differentiate between benign lesions and lung cancers. [45]
- NeuralSeg Ltd, St Joseph's Healthcare Hamilton and McMaster University performed clinical study to access an artificial intelligence algorithm (NeuralSeg) in identifying and predicting LN malignancy based on EBUS images. Results suggest that NeuralSeg is able to accurately rule out nodal metastasis and can possibly be used as an adjunct to EBUS when nodal biopsy is not possible or inconclusive. [46]
- Olympus, Chiba University and Dokkyo Medical University developed CNN-based CAD for the detection and classification of nodal metastasis from EBUS images. The prediction of LN metastasis by CAD using EBUS images showed high diagnostic accuracy with high specificity. CAD during EBUS-TBNA may help improve the diagnostic efficiency and reduce invasiveness of the procedure. [47]

4 Ethical considerations

The rapidly evolving field of AI and digital technology in the fields of medicine and public health raises a number of ethical, legal, and social concerns that have to be considered in this context. They are discussed in deliverable DEL01 "*AI4H ethics considerations,"* which was developed by the working group on "Ethical considerations on AI4H" (WG-Ethics). This section refers to DEL01 and should reflect 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 hospitals and other clinical institutions, to be responsible for handling the ethical, legal and privacy of the relevant data.

The following procedures is executed in our practice and recommended to other practice of AI for endoscopy.

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

5 Existing work on benchmarking

This section focuses on the existing benchmarking processes in the context of AI and AI for Endoscopy for quality assessment. The goal is to collect all relevant learnings from previous benchmarking that could help to implement the benchmarking process in this topic group.

5.1 Subtopic on colonoscopy

5.1.1 Publications on benchmarking systems

Some work has been done in the scientific community assessing the performance of AI for Endoscopy. This section summarizes insights from the most relevant publications on this topic. It covers parts of the deliverable DEL7 *"AI for health evaluation considerations,"* DEL7.1 *"AI4H evaluation process description,"* DEL7.2 *"AI technical test specification,"* DEL7.3 *"Data and artificial intelligence assessment methods (DAISAM),"* and DEL7.4 *"Clinical Evaluation of AI for health"*.

5.1.1.1 EndoCV

The International Workshop and Challenge on Computer Vision in Endoscopy(EndoCV) was held from 2019 to 2022 annually [48][49][50][51][52][53], 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 [54][55][56][57]. 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).

5.1.1.2 EndoVis

Endoscopic Vision Challenge (EndoVis) [58] 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.

5.1.1.2.1 GIANA

Gastrointestinal Image ANAlysis (GIANA) is one of the sub-challenge in EndoVis, which was held in 2017, 2018 [59] and 2021 [60].

5.1.1.2.2 CATARACTS

The Challenge on Automatic Tool Annotation for cataract Surgery (CATARACTS) [61] released 50 cataract surgery videos accompanied by instrument usage annotations including frame-level instrument presence information in 2017.

5.1.1.2.3 DAGI

The challenge of Detection of Abnormalities in Gastroscopic Images (DAGI) [62] was one of the sub-challenges 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.

5.1.1.2.4 APDCV

The challenge of Automatic Polyp Detection in Colonoscopy Videos (APDCV) [63] was one of the sub-challenges 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.

5.1.1.3 EndoTect

The EndoTect challenge [64] 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 [65].

5.1.2 Benchmarking by AI developers

All developers of AI solutions for endoscopy implemented internal benchmarking systems for assessing the performance. This section will outline the insights and learnings from this work of relevance for benchmarking in this topic group.

5.1.2.1 EndoCV

To achieve high diversity, the dataset of EndoCV 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.

There were two sub-challenges in EndoCV2022: 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 multicenter 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 multicenter 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. 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 was 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).

5.1.2.2 EndoVis

5.1.2.2.1 GIANA

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	Ground 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 classification Clinical 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 mask Clinical partner: Saint Antoine Hospital, 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 mask Clinical partner: Saint Antoine Hospital, 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 classification Clinical 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 treatment	1000 images for training and validation and 200 images for testing	Label of each frame, polyp region Clinical partner:

						Hospital Clinic, Barcelona, Spain

5.1.2.2.2 CATARACTS

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 [66].

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.

5.1.2.2.3 DAGI

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.

5.1.2.2.4 APDCV

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 [67], ETIS-LARIB and ASU-Mayo Clinic Colonoscopy Video Database [68].

- 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.

5.1.2.3 EndoTect

A large dataset containing images taken from several endoscopies named as HyperKvasir [69] 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 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.

5.1.3 Relevant existing benchmarking frameworks

Triggered by the hype around AI, recent years have seen the development of a variety of benchmarking platforms where AIs can compete for the best performance on a determined dataset.

- **EvalAI**: EvalAI is an open source platform for evaluating and comparing machine learning (ML) and artificial intelligence (AI) algorithms at scale [70].
- **AIcrowd**: AIcrowd enables data science experts and enthusiasts to collaboratively solve real-world problems, through challenges [71].
- **Kaggle**: Kaggle offers a no-setup, customizable, Jupyter Notebooks environment. Access GPUs at no cost to you and a huge repository of community published data & code [72].
- **CodaLab**: CodaLab is an open source platform to learn, create, collaborate through challenges [73].
- **Grand challenge**: A platform for end-to-end development of machine learning solutions in biomedical imaging.

Table 3: Relevant existing benchmarking frameworks

Benchmarking frameworks	Challenges	Users	Submissions	Organization	Paper published	Awarded in prizes	Datasets / stored data	Notebooks/ algorithms
EvalAI	200+	18k+	180k+	30+				
AIcrowd	246+	59k+			60+	823+ USD	13 TB+	
Kaggle		13.6m +					50k+	400k+

Benchmarking frameworks	Challenges	Users	Submissions	Organization	Paper published	Awarded in prizes	Datasets / stored data	Notebooks/ algorithms
CodaLab	1946	36k+	409k+					
Grand challenge	356	82k+	88k+					27k+

There are several public datasets related with AI for endoscopy, see Table 4.

Table 4: Available public dataset

Dataset	Findings	Size	Availability
CVC-612(CVC-ClinicDB) [67]	Polyp, with mask	612 images	open academic
ASU-Mayo polyp database [9]	Polyp, with mask	18,781 images	by request (no available anymore)
ETIS-Larib Polyp DB [74]	Polyp, with mask	196 images	open academic
KID [75]	Angiectasia, bleeding, inflammations, polyp	2371 images, 47 videos	open academic (no available anymore)
GIANA'17 [59]	Angiectasia, with mask	600 images	by request
GASTROLAB [76]	GI lesions	Some 100s of images + few videos	open academic (video capsule endoscopy)
WEO Clinical Endoscopy Atlas [77]	GI lesions	152 images	by request (video capsule endoscopy)
GI Lesions in Regular Colonoscopy Data Set [78]	GI lesions, with mask	76 images	by request
Atlas of Gastrointestinal Endoscope [79]	GI lesions	1295 images	unknown (no available anymore)
GastroAtlas [80]	GI lesions	5,071 video clips	open academic (video capsule endoscopy)
Kvasir [81]	Polyps, esophagitis, ulcerative colitis, Z-line, pylorus, cecum, dyed polyp, dyed resection margins, stool	8,000 images	open academic
Nerthus [82]	Stool - categorization of bowel cleanliness	21 videos	open academic
Kvasir-SEG [83]	Polyps, with mask	1000 images	open academic
HyperKvasir [69]	GI findings including polyps	110,079 images and 374 videos	open academic

Dataset	Findings	Size	Availability
Kvasir-Capsule [84]	GI findings including polyps (video capsule endoscopy)	4,741,504 images	open academic
CVC-ColonDB [85]	Polyps, with mask	380 images	open academic (no available anymore)
EDD2020 [56]	GI lesions including polyps	386 images	open academic

5.2 Subtopic Endoscopic Ultrasound

5.2.1 Publications on benchmarking systems

Although research on AI for EUS application is increasing rapidly during the last few years, public accessible dataset and benchmarking system does not exist. Several review papers have been published to summarize latest research in the field, but none of those can provide comparable benchmarking for different studies [86][87][88][89][90]. It's extremely important for the scientific community to establish a public accessible EUS image database and benchmarking system to push forward AI-assisted EUS research.

5.2.2 Benchmarking by AI developers

All developers of AI solutions for EUS implemented internal benchmarking systems for assessing the performance. Depending on the tasks of the AI solutions (detection, classification, segmentation etc.), different metrics will be used in order to enable performance comparison. These metrics are not much different from those used in medical image analysis and computer vision, specifically colonoscopy, such as mean average precision (mAP), intersection over union (IoU), Dice coefficient (DICE), positive predictive value (PPV), precision, recall, specificity, F1-measure, Matthews correlation coefficient and Area Under Curve (AUC) of Receiver Operating Characteristic (ROC) curve.

5.2.3 Relevant existing benchmarking frameworks

Relevant existing benchmarking frameworks of AI EUS are the same as colonoscopy, including EvalAI, AIcrowd, Kaggle, CodaLab and Grand challenge. For more details, chapter 5.1.3 could be referred to. And to the best of our knowledge, currently there's no public available dataset for AI EUS.

6 Benchmarking by the topic group "For further study."

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

It reflects the considerations of various deliverables: DEL5 *"Data specification"* (introduction to deliverables 5.1-5.6), DEL5.1*"Data requirements"* (which lists acceptance criteria for data submitted to FG-AI4H and states the governing principles and rules), DEL5.2 *"Data acquisition"*, DEL5.3 *"Data annotation specification"*, DEL5.4 *"Training and test data specification"* (which provides a systematic way of preparing technical requirement specifications for datasets used in training and testing of AI models), DEL5.5 *"Data handling"* (which outlines how data will be handled once they are accepted), DEL5.6 *"Data sharing practices"* (which provides an overview of the existing best practices for sharing health-related data based on distributed and federated environments, including the requirement to enable secure data sharing and addressing issues of data governance), DEL06 *"AI training best practices specification"* (which reviews best practices for proper AI model training and guidelines for model reporting), DEL7*"AI for health evaluation considerations"* (which discusses the validation and evaluation of AI for health models, and considers requirements for a benchmarking

platform), DEL7.1 *"AI4H evaluation process description"* (which provides an overview of the state of the art of AI evaluation principles and methods and serves as an initiator for the evaluation process of AI for health), DEL7.2 *"AI technical test specification"* (which specifies how an AI can and should be tested *in silico*), DEL7.3 *"Data and artificial intelligence assessment methods (DAISAM)"* (which provides the reference collection of WG-DAISAM on assessment methods of data and AI quality evaluation), DEL7.4*"Clinical Evaluation of AI for health"* (which outlines the current best practices and outstanding issues related to clinical evaluation of AI models for health), DEL7.5 *"FG-AI4H assessment platform"* (which explores assessment platform options that can be used to evaluate AI for health for the different topic groups), DEL9 *"AI for health applications and platforms"* (which introduces specific considerations of the benchmarking of mobile- and cloud-based AI applications in health), DEL9.1 *"Mobile based AI applications,"* and DEL9.2 *"Cloud-based AI applications"* (which describe specific requirements for the development, testing and benchmarking of mobile- and cloud-based AI applications).

6.1 Subtopic Colonoscopy

The benchmarking of AI for Endoscopy is being developed and improved continuously to reflect new features of AI systems or changed requirements for benchmarking. It should be noted that, the benchmarking by the TG-Endoscopy is not so ready and need further study.

6.1.1 Benchmarking version V1.0

This section includes all technological and operational details of the benchmarking process for the benchmarking version V1.0.

6.1.1.1 Overview

This section provides an overview of the key aspects of this benchmarking iteration, version V1.0. Besides recommended testing metrologies and scoring matrixes, data format requirement of input data and output data, training and testing data annotation quality control are involved in the method for AI benchmarking.

6.1.1.2 Benchmarking methods

This section provides details about the methods of the benchmarking version V1.0. All developers of AI solutions for endoscopy implemented internal benchmarking systems for assessing the performance.

6.1.1.2.1 Benchmarking system architecture

Referring to the benchmarking and evaluation of the challenge of Automatic Polyp Detection in Colonoscopy Videos (APDCV), the benchmarking system of colonoscopy should consist of AI tasks, Benchmarking metrics and Task based metrics calculation. Benchmarking version V1.0 is being built following this structure.

While the selection and calculation of metrics differs from AI tasks and applications, task based metrics calculation is decided by the type of AI tasks and benchmarking metrics used. For example, PDR is applicable for polyp detection but not polyp classification.

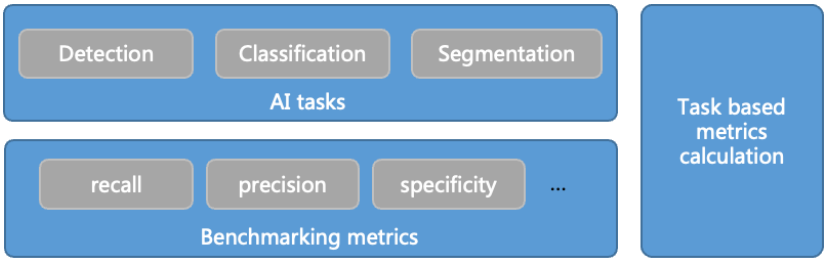

Figure 1: Architecture of benchmarking version

6.1.1.2.2 Benchmarking system dataflow

Initially, benchmarking version V1.0 will be a standalone and internal system for assessing the performance. The type of AI tasks and benchmarking metrics are defined manually, and the prediction by AI system is needed to be done before benchmarking.

Firstly, the test dataset was predicted by AI system to generate the prediction of test dataset, whose data structure need to applicable to the benchmarking system. Then, the AI task and benchmarking metrics are needed to be set based on the feature of algorithm, so as to calculate task based metrics. Finally, the prediction of test dataset by AI system was evaluated with ground truth of test dataset by the task based metrics.

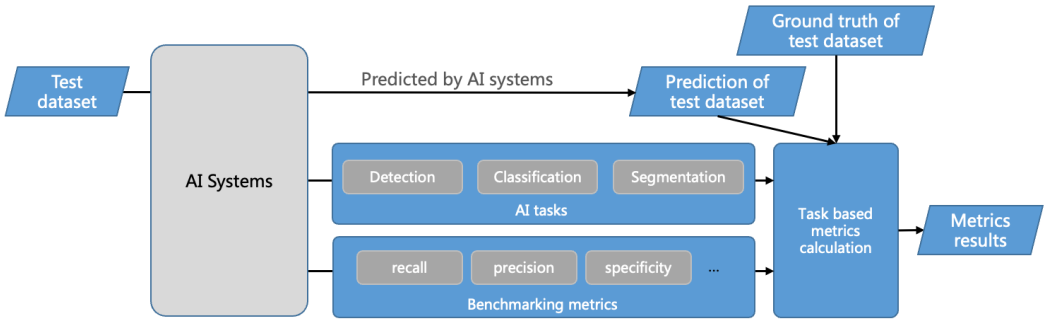

Figure 2: Dataflow of benchmarking version

6.1.1.2.3 Safe and secure system operation and hosting

The access to the system will be only authorized inside the corporation. More details of safety and security will be considered in the following version.

6.1.1.2.4 Benchmarking process

The current version of the benchmarking system will be a standalone system and not for open access. The prediction of test dataset by AI systems, definition of AI tasks and benchmarking metrics in benchmarking, and execution of benchmarking calculation will be handled and done by authorized AI developer.

6.1.1.3 AI input data structure for the benchmarking

This section describes the recommended structure of input data provided to the AI solutions as part of the benchmarking of AI for Endoscopy.

Endoscopic images or videos captured with colonoscope should be submitted as separate files in the following format:

- 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.
- 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.

6.1.1.4 AI output data structure

Similar to the input data structure for the benchmarking, this section describes the recommended structure of output data the AI systems are expected to generate in response to the input data.

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

- Information 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.

6.1.1.4.1 Detection

- Data Information: data name, data format, etc.
- Result Information:
  - Category Information: the types would depend on the task.
  - Location Information: 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 Information: height and width in pixels.
- Task info(optional): task ID, task name, task type, etc.

6.1.1.4.2 Classification

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

6.1.1.4.3 Segmentation

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

6.1.1.5 Test data label/annotation structure

While the AI systems can only receive the input data described in the previous sections, the benchmarking system needs to know the expected correct answer (sometimes called ‘labels’, ‘ground truth’ or ‘gold standard’) for each element of the input data so that it can compare the expected AI output with the actual one. Since this is only needed for benchmarking, it is encoded separately.

To guaranty the quality of data annotation and reduce individual differences among doctors, it is recommended that the annotation process should involve multiple steps by multiple doctors, such as independent annotation, cross-annotation, arbitration, and review. Specially, arbitration and review may be combined as one step by one doctor.

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.

6.1.1.5.1 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 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) [91]. 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 doctor 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 doctor. 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 doctor information, annotation procedure information, the date, annotation serial number.
- Review: The gold standard candidates would be confirmed by the review doctor one by one. The data approved by the review doctor 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 doctor 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 doctor information, annotation procedure information, the date, annotation serial number.

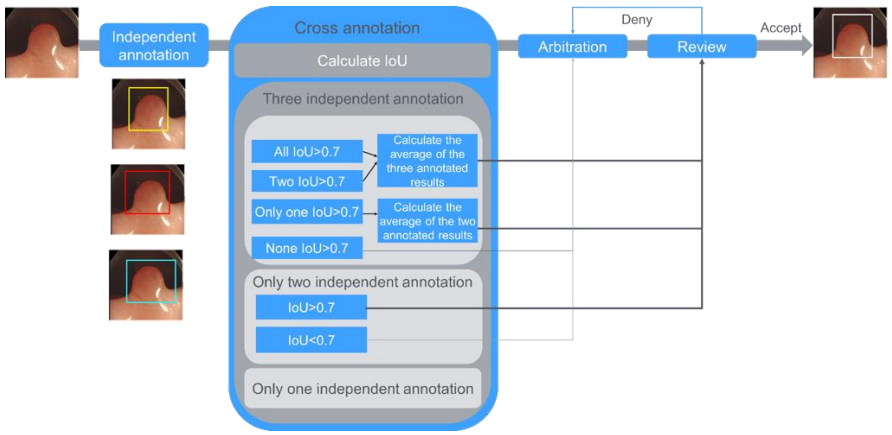

Figure 3: Illustration of annotation procedure for detection

6.1.1.5.2 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 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 doctor 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 doctor. 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 doctor information, annotation procedure information, the date, annotation serial number.

- Review: The gold standard candidates would be confirmed by the review doctor one by one. The data approved by the review doctor 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 doctor 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 doctor information, annotation procedure information, the date, annotation serial number.

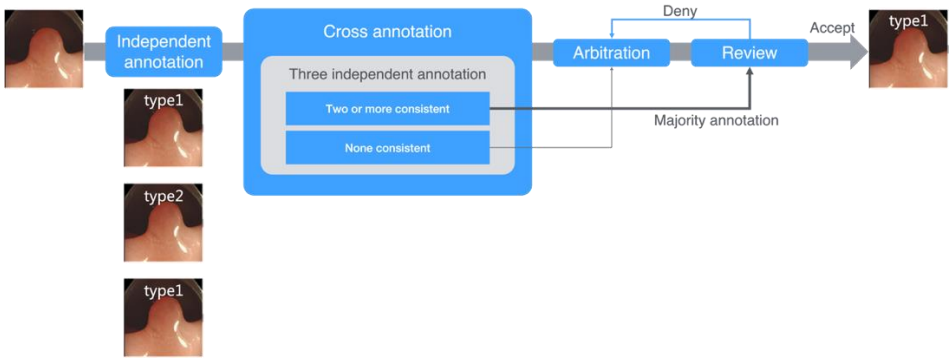

Figure 4: Illustration of annotation procedure for classification

6.1.1.5.3 Annotation of segmentation

Annotation of segmentation means the annotation of every pixel in an object within a data. Practically, there are two methods for annotation of segmentation, including annotating the contour of the object with a polygon and annotating the region of the object with a mask.

- Initial annotation: Initial annotation to sketch the contour or mask of the object by one doctor. All the annotated results should be well recorded and linked to corresponding images in a clear way. Initial annotation requirements include:
  - Non-annotating information(optional): Image name, image 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.
- Review: The initial annotation would be confirmed and modified by the review doctor. The data approved by the review doctor would be marked as the gold standard. Review annotation requirements include:
  - Non-annotating information(optional): Image name, image identification code, collection device model, collection date, image size, collection frame rate, hospital, patient information (age, gender, race)
  - Annotating information(mandatory): The gold standard, serial number for review, review doctor information, annotation procedure information, the date, annotation serial number.

6.1.1.6 Scores and metrics

Scores and metrics are at the core of the benchmarking. This section describes the scores and metrics applicable to measure the performance, robustness, and general characteristics of AI systems. The following table is a list of applicable scores and metrics, which could be used in demands.

Table 5: Benchmarking metrics

Methodology	Description	AI Task
True positive (TP)	The number of correctly identified positive samples. The number of frames with endoscopic findings which correctly is identified as a frame with an endoscopic finding.	Detection Classification
True negative (TN)	The number of correctly identified negative samples, i.e., the number of frames without an endoscopic finding which correctly is identified as a frame without endoscopic finding.	Detection Classification
False-positive (FP)	The number of wrongly identified positive samples, i.e., a commonly called a "false alarm". The number of frames without an endoscopic finding which is erroneously identified as a frame with an endoscopic finding.	Detection Classification
False-negative (FN)	The number of wrongly identified negative samples. The number of frames with an endoscopic finding which erroneously are identified as a frame without an endoscopic finding.	Detection Classification
Recall (REC) or Sensitivity (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.	Detection Classification
Precision (PREC) or Positive predictive value (PPV)	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 positive samples (the fraction of retrieved samples that are relevant).	Detection Classification
Negative predictive value (NPV)	It shows the ratio of samples that are correctly identified as negative among the predicted negative samples (the fraction of retrieved samples that are relevant).	Detection Classificaiton
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).	Detection Classification
Accuracy (ACC)	The percentage of correctly identified true and false samples.	Detection Classification
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.	Classification
F1 score (F1)	A measure of a test's accuracy by calculating the harmonic mean of the precision and recall.	Classification
DICE coefficient (DICE)	This metric measures the similarity between two sets of data and is most broadly used in the validation of image segmentation. It equals twice the number of elements common to both sets divided by the sum of the number of elements in each set.	Segmentation
Jaccard coefficient or Intersection 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.	Segmentation
Polyp detection rate (PDR)	This metric is the percentage of patients undergoing screening endoscopy who have one or more polyp detected.	Detection
Adenoma detection rate (ADR)	This metric is the percentage of patients undergoing screening endoscopy who have one or more conventional adenomas detected.	Detection
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.	Detection
Average precision (AP)	This metric is the average precision in the P-R curve.	Detection
Mean average precision (mAP)	This metric is the average value of AP of every class.	Detection
Runtime	This metric is the cost time of running one image or frame by a method.	Detection Classification Segmentation

6.1.1.7 Test dataset acquisition

The test dataset acquisition is in progress.

6.1.1.8 Data sharing policies

After finishing the test dataset acquisition, the sharing of dataset should be protected by special agreements or contracts that cover, for instance, the data sharing period, patient consent, and update procedure (see also DEL5.5 on *data handling* and DEL5.6 on *data sharing practices*).

6.1.1.9 Baseline acquisition

The baseline will be acquired after finishing the test dataset acquisition.

6.1.1.10 Reporting methodology

The results of benchmarking runs will be shared for AI developers internally. There is no public reporting methodology, except publication of technical paper.

6.1.1.11 Result

Results will be not available before finishing the benchmarking version 1.0 and test dataset acquisition.

6.1.1.12 Discussion of the benchmarking

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

In benchmarking of subtopic colonoscopy, recommended requirements for benchmarking methods, data structure of input and output, annotation structure and information, score and metrics, test dataset and result are described. With regards to the difference of AI tasks, there are individual requirements for detection, segmentation, and classification in corresponding chapters.

Referring to the benchmarking and evaluation of APDCV, the benchmarking in this subtopic is being built as a standalone system initially, consisting of AI tasks, Benchmarking metrics and Task based metrics calculation. As there are several existing benchmarking systems. Data structure of input/output and annotation structure and information are considered first, so the progresses of data structure of input/output, annotation structure and information are more advanced than benchmarking system and test dataset acquisition. 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. The benchmarking version 1.0 tries to propose a general standardized solution of requirements for data structure and annotation.

6.1.1.13 Retirement

Generally, the retirement of the AI system and dataset should follow the policy agreed with providers and users before the benchmarking activity. 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 "*AI software lifecycle specification"* (identification of standards and best practices that are relevant for the AI for health software life cycle).

6.2 Subtopic Endoscopic Ultrasound

6.2.1 Benchmarking version V1.0

This section includes all technological and operational details of the benchmarking process for the benchmarking version V1.0. It should be noted that, the benchmarking by the TG-Endoscopy is not so ready and need further study.

6.2.1.1 Overview

This section provides an overview of the key aspects of this benchmarking iteration, version V1.0.

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

6.2.1.2 Benchmarking methods

6.2.1.2.1 Benchmarking system architecture

Comparing with colonoscopy, EUS is a multi-modality procedure capturing ultrasound and image at the same time, but the requirements of benchmarking are similar. EUS and colonoscopy share the same benchmarking system architecture described in chapter ‘6.1.1.2.1’.

6.2.1.2.2 Benchmarking system dataflow

EUS and colonoscopy share the same benchmarking system dataflow described in chapter ‘6.1.1.2.2’.

6.2.1.2.3 Safe and secure system operation and hosting

EUS and colonoscopy share the same safe and secure system operation and hosting described in chapter ‘6.1.1.2.3’.

6.2.1.2.4 Benchmarking process

EUS and colonoscopy share the same benchmarking process described in chapter ‘6.1.1.2.4’.

6.2.1.3 AI input data structure for the benchmarking

Ultrasound images or videos captured with EUS should be submitted as separate files in the following format:

- 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.

6.2.1.4 AI output data structure

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

- Information 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.

6.2.1.4.1 Detection

- Data Information: data name, data format, etc.
- Result Information:
  - Category Information: the types would depend on the task.
  - Location Information: 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 Information: height and width in pixels.
- Task info(optional): task ID, task name, task type, etc.

6.2.1.4.2 Classification

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

6.2.1.4.3 Segmentation

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

6.2.1.5 Test data label/annotation structure

While the AI systems can only receive the input data described in the previous sections, the benchmarking system needs to know the expected correct answer (sometimes called ‘labels’, ‘ground truth’ or ‘gold standard’) for each element of the input data so that it can compare the expected AI output with the actual one. Since this is only needed for benchmarking, it is encoded separately.

Comparing with colonoscopy, EUS is a multi-modality procedure capturing ultrasound and image at the same time. The test data label/annotation structure needs to consider of two modalities as ultrasound data and image/video.

Referring to chapter of ‘Test data label/annotation structure’ of colonoscopy, it is recommended that the annotation process should involve multiple steps by multiple doctors, such as independent annotation, cross-annotation, arbitration, and review. Specially, arbitration and review may be combined as one step by one doctor.

6.2.1.5.1 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 doctors to confirm whether the image/video/ultrasound data 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/ultrasound data name, image/video/ultrasound data 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). 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/ultrasound data name, image/video/ultrasound data 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 doctor 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 doctor. Arbitration requirements include:
    - Non-annotating information(optional): Image/video/ultrasound data name, image/video/ultrasound data 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 doctor information, annotation procedure information, the date, annotation serial number.
- Review: The gold standard candidates would be confirmed by the review doctor one by one. The data approved by the review doctor 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 doctor to generate the gold standard. Review requirements include:
    - Non-annotating information(optional): Image/video/ultrasound data name, image/video/ultrasound data 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 doctor information, annotation procedure information, the date, annotation serial number.

6.2.1.5.2 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 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/ultrasound data name, image/video/ultrasound data 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/ultrasound data name, image/video/ultrasound data 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 doctor 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 doctor. Arbitration requirements include:
    - Non-annotating information(optional): Image/video/ultrasound data name, image/video/ultrasound data 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 doctor information, annotation procedure information, the date, annotation serial number.
- Review: The gold standard candidates would be confirmed by the review doctor one by one. The data approved by the review doctor 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 doctor to generate the gold standard. Review requirements include:
    - Non-annotating information(optional): Image/video/ultrasound data name, image/video/ultrasound data 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 doctor information, annotation procedure information, the date, annotation serial number.

6.2.1.5.3 Annotation of segmentation

Annotation of segmentation means the annotation of every pixel in an object within a data. Practically, there are two methods for annotation of segmentation, including annotating the contour of the object with a polygon and annotating the region of the object with a mask.

- Initial annotation: Initial annotation to sketch the contour or mask of the object by one doctor. All the annotated results should be well recorded and linked to corresponding images in a clear way. Initial annotation requirements include:
    - Non-annotating information(optional): Image/ultrasound data name, image/ultrasound data 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.

- Review: The initial annotation would be confirmed and modified by the review doctor. The data approved by the review doctor would be marked as the gold standard. Review annotation requirements include:
  - Non-annotating information(optional): Image/ultrasound data name, image/ultrasound data identification code, collection device model, collection date, image size, collection frame rate, hospital, patient information (age, gender, race)
  - Annotating information(mandatory): The gold standard, serial number for review, review doctor information, annotation procedure information, the date, annotation serial number.

6.2.1.6 Scores and metrics

EUS and colonoscopy share the same scores and metrics described in chapter '6.1.1.6'.

6.2.1.7 Test dataset acquisition

The test dataset acquisition is in progress.

6.2.1.8 Data sharing policies

After finishing the test dataset acquisition, the sharing of dataset should be protected by special agreements or contracts that cover, for instance, the data sharing period, patient consent, and update procedure (see also DEL5.5 on *data handling* and DEL5.6 on *data sharing practices*).

6.2.1.9 Baseline acquisition

The baseline will be acquired after finishing the test dataset acquisition.

6.2.1.10 Reporting methodology

EUS and colonoscopy will share the same reporting methodology described in chapter '6.1.1.10'.

6.2.1.11 Result

Currently, there's no public available EUS dataset and benchmarking system. It's impossible to perform comparable benchmarking for different AI solutions. Several review papers have been published to summarize latest research in AI-assisted EUS in different clinical fields [86][87][88][89][90]. For instance, Dumitrescu. et. al. conducted meta-analysis for the diagnostic value of AI -assisted EUS for pancreatic cancer with ten clinical studies and 1871 patients [88]. The overall diagnostic accuracy showed 0.92 (95% CI, 0.89–0.95) sensitivity and 0.9 (95% CI, 0.83–0.94) specificity.

6.2.1.12 Discussion of the benchmarking

In general, EUS produces ultrasound images of different kinds (B mode, contrast enhanced ultrasound, elastography), just like conventional endoscopy dose (white light, narrow band imaging, dye-spray chromoendoscopy). So the benchmarking methods, general requirements for data structure of input and output, annotation structure and information, score and metrics, test dataset and result are no big different with other endoscopic subgroup. General descriptions are given in above sections. With regards to the difference of different AI tasks, there are individual requirements for detection, segmentation, and classification in corresponding chapters.

It should be noted that one data type, radiofrequency (RF) data, and its related AI task (for instance, beamforming, data compression, denoising, reconstruction, etc) are discarded from current version of benchmarking process for mainly two reasons. (1) Only a few EUS manufacturers and research facilities have the ability to access EUS RF data from EUS system. Research on RF data based AI-EUS is rare at the moment; (2) Currently there's no standard for storing RF data for different EUS manufacturers. It can be added in the future if needed.

6.2.1.13 Retirement

EUS and colonoscopy will share the same reporting methodology described in chapter ‘6.1.1.13’.

7 Overall discussion of the benchmarking

Endoscopy is the core technical means for early diagnosis and screening of digestive cancer, which can drastically reduce the incidence and mortality caused by digestive cancer. Furthermore, with the breakthrough of the new generation of artificial intelligence technology represented by deep learning, revolutionary progress has been made, and 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 AI for endoscopy systems in two subtopics, including colonoscopy and endoscopic ultrasound.

Colonoscopy is considered the gold standard for CRC screening to detect and remove the polyps and adenomas in the colorectum. By the effort of researchers, some work has been done in the scientific community assessing the performance of such application, such as challenges and datasets. In each challenge, general elements were involved, including task, data, annotation, metrics, while there might be different definition and selection of these elements. The fundamental element would be a determined dataset. A variety of datasets have been released and open accessed. We can see dataset annotated with single type of lesion or multiple types of lesions. We can also see dataset of images or videos. Based on these challenges and dataset, researchers have published a variety of high-quality publications. And there are already commercial AI products on the market. In this TDD, recommended requirements for data structure of input and output, annotation structure and information, score and metrics, test dataset and result are described. There is more specific description about recommended requirements for data structure of input and output, annotation structure and information, which aims to guarantee the accuracy and representativeness of dataset and annotation. Referring to the benchmarking and evaluation of APDCV, the benchmarking in this subtopic is being built as a standalone system initially. The access to the system will be only authorized inside the corporation.

As a fresh technology, clinical evidence has shown the benefits of endoscopic ultrasound over the potential adverse events and clinical guidelines have been published and continuously updated to ensure the safely use of the procedures. In general, EUS produces ultrasound images of different kinds (B mode, contrast enhanced ultrasound, elastography), just like conventional endoscopy dose (white light, narrow band imaging, dye-spray chromoendoscopy). So the benchmarking methods, general requirements for data structure of input and output, annotation structure and information, score and metrics, test dataset and result are no big different with other endoscopic subgroup. Although research on artificial intelligence in EUS is still limited, we believed AI would play important role in endoscopic ultrasound 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 workload and standardize the overall quality of endoscopic procedures.

Generally, it should be noted that, the benchmarking by the TG-Endoscopy is not so ready and need further study.

8 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)* compiled the requirements that consider these challenges.

The deliverables with relevance for regulatory considerations are DEL2 *"AI4H regulatory considerations"* (which provides an educational overview of some key regulatory considerations), DEL2.1 *"Mapping of IMDRF essential principles to AI for health software",* and DEL2.2 *"Guidelines for AI based medical device (AI-MD): Regulatory requirements"* (which provides a checklist to understand expectations of regulators, promotes step-by-step implementation of safety and effectiveness of AI-based medical devices, and compensates for the lack of a harmonized standard). DEL04 identifies standards and best practices that are relevant for the "*AI software lifecycle specification."* The following sections discuss how the different regulatory aspects relate to the TG-Endoscopy.

8.1 Existing applicable regulatory frameworks

Most of the AI systems that are part of the FG-AI4H benchmarking process can be classified as *software as 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.

If the AI systems for endoscopy is for clinical purpose and classified as *software as medical device* (SaMD), it would be covered by existing regulatory frameworks, such as NMPA, MDR, FDA, GDPR, and ISO. And the AI manufacturers need to address all the requirements of those regulatory frameworks.

8.2 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.

The benchmarking participants need to provide compliance features and certifications as part of the metadata following the regulatory requirements in DEL2 *"AI4H regulatory considerations"*.

8.3 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.

Referring to the regulatory requirements in DEL2 "AI4H regulatory considerations", if the benchmarking system is built for evaluation of medical device, it might need to comply with the following requirements.

Table 6: Regulatory requirements for the benchmarking systems

Requirement(s)	**Checklist item(s)**	**Applicable standards and regulations**
The manufacturer should plan the model evaluation.	− There is an evaluation plan. − The plan specifies the evaluation activities, the roles involved and the milestones at which these activities have to be performed. − The plan foresees the evaluation with clinically relevant data sets independent from training datasets.	[ISO 13485] clauses 7.3.2, 7.3.6 and 7.3.7 [ISO 14971] clause 10. GMLP guiding principle (8) (by FDA et al.)
The manufacturer should gain an understanding on how the machine makes a decision to evaluate the correctness and robustness of the model.	− There is a validation specification and validation results for the evaluation of the model with validation data set. − There is a test specification and test results for the final evaluation of the model with new test data. − There are documented values for specified quality metrics. − There may be an analysis of datasets that have exhibited good model performance versus datasets that have performed badly. − For individual data sets there may be an evaluation of the feature that the model particularly determined in the decision. − There may be an analysis/visualization of the dependency	[EU-MDR (2017/745)] Annex I (17), Annex II (6.1). [IEC 62304] clauses 5.5 *ff*. [ISO 13485] clause 7.3.4 *ff*. [b-XAVIER] "Perspectives and good practices for AI and continuously learning systems in healthcare" [b-XAVIER University] "Building explainability and trust

	(strength, direction) of the prediction of the feature values. − There may be a synthetization of data sets that activate the model particularly strong. − There may be an approximation of the model using a simplified surrogate model.	for AI in healthcare" DIN SPEC 2 [b-ISO/IEC TR 24028] clauses 10.2 and 10.3 GMLP guiding principles (6) (e.g., overfitting) and (8) (confounding factors) (by FDA et al.)

8.4 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 in order to be compliant. The discussion here focuses on the guidance and best practice provided by the DEL2 *"AI4H regulatory considerations."*

To comply with applicable regulatory requirements, TG-Endoscopy will refer to the guidance and best practice provided by DEL2 *"AI4H regulatory considerations."*.

References

[1] 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.

[2] 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.

[3] Chen, Peng-Jen, et al. "Accurate classification of diminutive colorectal polyps using computer-aided analysis." Gastroenterology 154.3 (2018): 568-575.

[4] 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.

[5] Mori, Yuichi, and Shin-ei Kudo. "Detecting colorectal polyps via machine learning." Nature biomedical engineering 2.10 (2018): 713-714.

[6] Misawa, Masashi, et al. "Artificial intelligence-assisted polyp detection for colonoscopy: initial experience." Gastroenterology 154.8 (2018): 2027-2029.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.
[15] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.
[16] Japan Science and Technology Agency, Immediate detection of colorectal cancer with AI. https://www.jst.go.jp/EN/achievements/research/bt2019-07.html, 2019.
[17] 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.
[18] 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.
[19] 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.
[20] 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.
[21] 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.
[22] Wu, Lianlian, et al. “A deep neural network improves endoscopic detection of early gastric cancer without blind spots.” Endoscopy 51.06 (2019): 522-531.
[23] 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.
[24] Gong D, Wu L, Zhang J, et al. Detection of colorectal adenomas with a real-time computer-aided system (ENDOANGEL): a randomised controlled study[J]. The lancet Gastroenterology & hepatology, 2020, 5(4): 352-361.
[25] Van Der Merwe, Schalk W., et al. “Therapeutic endoscopic ultrasound: European Society of Gastrointestinal Endoscopy (ESGE) Guideline.” Endoscopy (2021).
[26] Forbes, Nauzer, et al. “Adverse events associated with EUS and EUS-guided procedures.” Gastrointestinal endoscopy 95.1 (2022): 16-26.
[27] Reddy, Yogananda, and Robert P. Willert. “Endoscopic ultrasound: what is it and when should it be used?.” Clinical medicine 9.6 (2009): 539-543.
[28] Mertz, Howard R., et al. “EUS, PET, and CT scanning for evaluation of pancreatic adenocarcinoma.” Gastrointestinal endoscopy 52.3 (2000): 367-371.
[29] Wallace, Michael B., et al. “Minimally invasive endoscopic staging of suspected lung cancer.” Jama 299.5 (2008): 540-546.
[30] Kuwahara, Takamichi, et al. “Current status of artificial intelligence analysis for endoscopic ultrasonography.” Digestive Endoscopy 33.2 (2021): 298-305.
[31] Liu JQ, Ren JY, Xu XL, Xiong LY, Peng YX, Pan XF, Dietrich CF, Cui XW. Ultrasound-based artificial intelligence in gastroenterology and hepatology. World J Gastroenterol. 2022 Oct 14;28(38):5530-5546. Doi: 10.3748/wjg.v28.i38.5530. PMID: 36304086; PMCID: PMC9594013.
[32] Liu E, Bhutani MS, Sun S. Artificial intelligence: The new wave of innovation in EUS. Endosc Ultrasound. 2021 Mar-Apr;10(2):79-83. Doi: 10.4103/EUS-D-21-00052. PMID: 33885005; PMCID: PMC8098839.
[33] Dumitrescu EA, Ungureanu BS, Cazacu IM, Florescu LM, Streba L, Croitoru VM, Sur D, Croitoru A, Turcu-Stiolica A, Lungulescu CV. Diagnostic Value of Artificial Intelligence-Assisted Endoscopic Ultrasound for Pancreatic Cancer: A Systematic Review and Meta-

Analysis. Diagnostics (Basel). 2022 Jan 25;12(2):309. Doi: 10.3390/diagnostics12020309. PMID: 35204400; PMCID: PMC8870917.
[34] Wu J, Wu C, Zhou C, Zheng W, Li P. Recent advances in convex probe endobronchial ultrasound: a narrative review. Ann Transl Med. 2021 Mar;9(5):419. Doi: 10.21037/atm-21-225. PMID: 33842640; PMCID: PMC8033319.
[35] 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.
[36] 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.
[37] Kim YH, Kim GH, Kim KB, Lee MW, Lee BE, Baek DH, Kim DH, Park JC. Application of A Convolutional Neural Network in The Diagnosis of Gastric Mesenchymal Tumours on Endoscopic Ultrasonography Images. J Clin Med. 2020 Sep 29;9(10):3162. Doi: 10.3390/jcm9103162. PMID: 33003602; PMCID: PMC7600226.
[38] Zhang MM, Yang H, Jin ZD, Yu JG, Cai ZY, Li ZS. Differential diagnosis of pancreatic cancer from normal tissue with digital imaging processing and pattern recognition based on a support vector machine of EUS images. Gastrointest Endosc. 2010 Nov;72(5):978-85. Doi: 10.1016/j.gie.2010.06.042. Epub 2010 Sep 19. PMID: 20855062.
[39] Zhu M, Xu C, Yu J, Wu Y, Li C, Zhang M, Jin Z, Li Z. Differentiation of pancreatic cancer and chronic pancreatitis using computer-aided diagnosis of endoscopic ultrasound (EUS) images: a diagnostic test. PloS One. 2013 May 21;8(5):e63820. Doi: 10.1371/journal.pone.0063820. PMID: 23704940; PMCID: PMC3660382.
[40] Tonozuka R, Itoi T, Nagata N, Kojima H, Sofuni A, Tsuchiya T, Ishii K, Tanaka R, Nagakawa Y, Mukai S. Deep learning analysis for the detection of pancreatic cancer on endosonographic images: a pilot study. J Hepatobiliary Pancreat Sci. 2021 Jan;28(1):95-104. Doi: 10.1002/jhbp.825. Epub 2020 Oct 15. PMID: 32910528.
[41] Săftoiu A, Vilmann P, Gorunescu F, Janssen J, Hocke M, Larsen M, Iglesias-Garcia J, Arcidiacono P, Will U, Giovannini M, Dietrich CF, Havre R, Gheorghe C, McKay C, Gheonea DI, Ciurea T; European EUS Elastography Multicentric Study Group. Efficacy of an artificial neural network-based approach to endoscopic ultrasound elastography in diagnosis of focal pancreatic masses. Clin Gastroenterol Hepatol. 2012 Jan;10(1):84-90.e1. doi: 10.1016/j.cgh.2011.09.014. Epub 2011 Oct 1. PMID: 21963957.
[42] Săftoiu A, Vilmann P, Dietrich CF, Iglesias-Garcia J, Hocke M, Seicean A, Ignee A, Hassan H, Streba CT, Ioncică AM, Gheonea DI, Ciurea T. Quantitative contrast-enhanced harmonic EUS in differential diagnosis of focal pancreatic masses (with videos). Gastrointest Endosc. 2015 Jul;82(1):59-69. Doi: 10.1016/j.gie.2014.11.040. Epub 2015 Mar 16. PMID: 25792386.
[43] Zhang J, Zhu L, Yao L, Ding X, Chen D, Wu H, Lu Z, Zhou W, Zhang L, An P, Xu B, Tan W, Hu S, Cheng F, Yu H. Deep learning-based pancreas segmentation and station recognition system in EUS: development and validation of a useful training tool (with video). Gastrointest Endosc. 2020 Oct;92(4):874-885.e3. doi: 10.1016/j.gie.2020.04.071. Epub 2020 May 6. Erratum in: Gastrointest Endosc. 2021 Mar;93(3):781. PMID: 32387499.
[44] Chen CH, Lee YW, Huang YS, Lan WR, Chang RF, Tu CY, Chen CY, Liao WC. Computer-aided diagnosis of endobronchial ultrasound images using convolutional neural network. Comput Methods Programs Biomed. 2019 Aug;177:175-182. Doi: 10.1016/j.cmpb.2019.05.020. Epub 2019 May 22. PMID: 31319946.
[45] Hotta T, Kurimoto N, Shiratsuki Y, Amano Y, Hamaguchi M, Tanino A, Tsubata Y, Isobe T. Deep learning-based diagnosis from endobronchial ultrasonography images of pulmonary lesions. Sci Rep. 2022 Aug 12;12(1):13710. Doi: 10.1038/s41598-022-17976-5. PMID: 35962181; PMCID: PMC9374687.
[46] Churchill IF, Gatti AA, Hylton DA, Sullivan KA, Patel YS, Leontiadis GI, Farrokhyar F, Hanna WC. An Artificial Intelligence Algorithm to Predict Nodal Metastasis in Lung Cancer. Ann

Thorac Surg. 2022 Jul;114(1):248-256. Doi: 10.1016/j.athoracsur.2021.06.082. Epub 2021 Aug 8. PMID: 34370986.
[47] Ito Y, Nakajima T, Inage T, Otsuka T, Sata Y, Tanaka K, Sakairi Y, Suzuki H, Yoshino I. Prediction of Nodal Metastasis in Lung Cancer Using Deep Learning of Endobronchial Ultrasound Images. Cancers (Basel). 2022 Jul 8;14(14):3334. doi: 10.3390/cancers14143334. PMID: 35884395; PMCID: PMC9321716.
[48] EndoCV2022: Endoscopic computer vision challenges 2.0, https://endocv2022.grand-challenge.org/EndoCV-Sequence/, 2022.
[49] EndoCV2021: Polyp Detection and Segmentation: Addressing Generalisability, https://endocv2021.grand-challenge.org/, 2021.
[50] EndoCV2020, https://endocv.grand-challenge.org/, 2020.
[51] EndoCV2020: EAD2020, https://ead2020.grand-challenge.org/, 2020.
[52] EndoCV2020: EDD2020, https://edd2020.grand-challenge.org/, 2020.
[53] EAD2019, https://ead2019.grand-challenge.org/, 2019.
[54] Ali, Sharib, and Noha Ghatwary. "Endoscopic computer vision challenges 2.0." (2022)
[55] Ali, Sharib, et al. "PolypGen: A multi-center polyp detection and segmentation dataset for generalisability assessment." arXiv preprint arXiv:2106.04463 (2021).
[56] Ali, Sharib, et al. "Endoscopy disease detection challenge 2020." arXiv preprint arXiv:2003.03376 (2020).
[57] Ali, Sharib, et al. "Endoscopy artifact detection (EAD 2019) challenge dataset." arXiv preprint arXiv:1905.03209 (2019)
[58] EndoVis2022, https://endovis.grand-challenge.org/Endoscopic_Vision_Challenge/, 2022.
[59] EndoVis2017: GIANA 2017, https://endovissub2017-giana.grand-challenge.org/, 2017.
[60] EndoVis2021: GIANA 2017, https://giana.grand-challenge.org/, 2021.
[61] EndoVis2020: CATARACTS Semantic Segmentation 2020, https://cataracts-semantic-segmentation2020.grand-challenge.org/Home/, 2020.
[62] EndoVis2015: Detection of abnormalities in gastroscopic images, https://endovissub-abnormal.grand-challenge.org/, 2015.
[63] EndoVis2015: Automatic Polyp Detection in Colonoscopy Videos, https://polyp.grand-challenge.org/, 2015.
[64] EndoTech2021, https://endotect.com/, 2021.
[65] 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.
[66] Luengo, Imanol, et al. "2020 CATARACTS Semantic Segmentation Challenge." arXiv preprint arXiv:2110.10965, 2021.
[67] 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.
[68] 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.
[69] Borgli, Hanna, et al. "HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy." Scientific data 7.1 (2020): 1-14.
[70] EvalAI, https://eval.ai/, 2023.
[71] AIcrowd, https://www.aicrowd.com/, 2023.
[72] Kaggle, https://www.kaggle.com/, 2023.
[73] CodaLab, https://codalab.org/, 2023
[74] 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.

[75] 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
[76] Gastrolab – the gastrointestinal site. http://www.gastrolab.net/index.htm.
[77] Weo clinical endoscopy atlas. http://www.endoatlas.org/index.php.
[78] Gastrointestinal lesions in regular colonoscopy dataset. http://www.depeca.uah.es/colonoscopy_dataset/.
[79] The atlas of gastrointestinal endoscope. http://www.endoatlas.com/atlas_1.html.
[80] El alvador atlas of gastrointestinal video endoscopy. http://www.gastrointestinalatlas.com/index.html.
[81] 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
[82] Pogorelov, Konstantin, et al. “Nerthus: A bowel preparation quality video dataset.” Proceedings of the 8th ACM on Multimedia Systems Conference. 2017
[83] Jha, Debesh, et al. “Kvasir-seg: A segmented polyp dataset.” International Conference on Multimedia Modeling. Springer, Cham, 2020
[84] Smedsrud, Pia H., et al. “Kvasir-Capsule, a video capsule endoscopy dataset.” Scientific Data 8.1 (2021): 1-10.
[85] Bernal, Jorge, Javier Sánchez, and Fernando Vilarino. “Towards automatic polyp detection with a polyp appearance model.” Pattern Recognition 45.9 (2012): 3166-3182
[86] Dahiya DS, Al-Haddad M, Chandan S, Gangwani MK, Aziz M, Mohan BP, Ramai D, Canakis A, Bapaye J, Sharma N. Artificial Intelligence in Endoscopic Ultrasound for Pancreatic Cancer: Where Are We Now and What Does the Future Entail? J Clin Med. 2022 Dec 16;11(24):7476. doi: 10.3390/jcm11247476. PMID: 36556092; PMCID: PMC9786876.
[87] Liu E, Bhutani MS, Sun S. Artificial intelligence: The new wave of innovation in EUS. Endosc Ultrasound. 2021 Mar-Apr;10(2):79-83. doi: 10.4103/EUS-D-21-00052. PMID: 33885005; PMCID: PMC8098839.
[88] Dumitrescu EA, Ungureanu BS, Cazacu IM, Florescu LM, Streba L, Croitoru VM, Sur D, Croitoru A, Turcu-Stiolica A, Lungulescu CV. Diagnostic Value of Artificial Intelligence-Assisted Endoscopic Ultrasound for Pancreatic Cancer: A Systematic Review and Meta-Analysis. Diagnostics (Basel). 2022 Jan 25;12(2):309. doi: 10.3390/diagnostics12020309. PMID: 35204400; PMCID: PMC8870917.
[89] Wu J, Wu C, Zhou C, Zheng W, Li P. Recent advances in convex probe endobronchial ultrasound: a narrative review. Ann Transl Med. 2021 Mar;9(5):419. doi: 10.21037/atm-21-225. PMID: 33842640; PMCID: PMC8033319.
[90] Liu JQ, Ren JY, Xu XL, Xiong LY, Peng YX, Pan XF, Dietrich CF, Cui XW. Ultrasound-based artificial intelligence in gastroenterology and hepatology. World J Gastroenterol. 2022 Oct 14;28(38):5530-5546. doi: 10.3748/wjg.v28.i38.5530. PMID: 36304086; PMCID: PMC9594013.
[91] Intersection over Union (IoU) for object detection, https://pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/, 2005.

Annex A:
Glossary

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

Acronym/Term	Expansion	Comment
AEs	Adverse events	
AI	Artificial intelligence	
AI4H	Artificial intelligence for health	
AI-MD	AI based medical device	
AMR	Adenoma miss rates	
API	Application programming interface	
AUC	Area Under Curve	
CADe	Computer assisted-detection systems	
CADx	Computer-aided diagnosis system	
CATARACTS	The Challenge on Automatic Tool Annotation for cataract Surgery	
CfTGP	Call for topic group participation	
CRC	Colorectal cancers	
CT	Computed tomography	
DAGI	The challenge of Detection of Abnormalities in Gastroscopic Images	
DCNN	Deep convolution neural network	
DEL	Deliverable	
EBUS	Endobronchial ultrasound	
EGC	Early gastric cancer	
EUS	Endoscopic ultrasound	
FDA	Food and Drug administration	
FGAI4H	Focus Group on AI for Health	
GDP	Gross domestic product	
GDPR	General Data Protection Regulation	
GIANA	Gastrointestinal Image ANAlysis	
HD	Hausdorff distance	
IMDRF	International Medical Device Regulators Forum	
IoU	Intersection over union	
IP	Intellectual property	
ISBI	IEEE International Symposium on Biomedical Imaging	
ISO	International Standardization Organization	
ITU	International Telecommunication Union	
LMIC	Low-and middle-income countries	
mAP	Mean average precision	
MCC	Matthews correlation coefficient	
MDR	Medical Device Regulation	

MICCAI	International CPonference on Medical Image Computing and Computer Assisted Intervention	
MRI	Magnetic resonance imaging	
NPV	Negative predictive value	
PD	Pancreatic duct drainage	
PII	Personal identifiable information	
PPV	Positive predictive value	
PTBD	Percutaneous transhepatic biliary drainage	
ROC	Receiver Operating Characteristic curve	
SaMD	Software as a medical device	
TDD	Topic Description Document	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 AI for Endoscopy
TG	Topic Group	
WG	Working Group	
WHO	World Health Organization	

Annex B:
Declaration of conflict of interests

The contributors declare that they have no conflicts of interest.

Tencent Healthcare (Shenzhen) Co., Ltd

Harnessing the technical capabilities, Tencent Healthcare aims to promote innovation in technologies, applications and cooperation models in the healthcare sector. Through upstream and downstream partnerships, Tencent strives to strengthen the digital capabilities for the industry, resulting in improved medical services, enhanced diagnostic efficiency, and ultimately leading to a new digital healthcare ecosystem. Tencent Healthcare encompasses Medical AI diagnosis, Smart Hospital, and Tencent Medipedia, offering comprehensive, convenient, precise and efficient medical and healthcare services to the public.

The China Academy of Information and Communications Technology

Founded in 1957, the China Academy of Information and Communications Technology (hereinafter referred to as CAICT) is a scientific research institute directly under the Ministry of Industry and Information Technology (MIIT) of China. Committed to "the think-tank and enabler for innovation and development in an information society", CAICT has provided strong support for major strategy, plan, policy, test, and certification for the development of the national ICT sector and the IT application, thus proving itself an important facilitator in the leapfrog development and innovation of China's information and communications sector, playing an important role in international cooperation related to the ICT sector and the integration of industrialization and informatization.

Olympus Medical Systems Corp.

At Olympus Medical Systems, we focus on improving patient care quality every day. We do this through developing and designing world-leading, clinically-advanced, precision technologies and services. Our products enable healthcare professionals, from a broad range of specialties, to ‘peer’ inside the body, using endoscopic procedures. This allows them to see more and do more. By focusing on early detection and minimally invasive treatment of a broad range of diseases, our mutual mission is to improve patient outcomes, minimize discomfort, and accelerate the recovery process. Our innovative technologies and services can also optimize workflow and maximize operational efficiency.

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

Suzhou Institute of Biomedical Engineering and Technology (SIBET), Chinese Academy of Sciences (CAS) is the only institute for research and development of biomedical instruments in CAS. To meet the significant needs in biomedical products, we focused on the basic, strategic, prospective researches in advanced biomedical instruments, reagents and biomedical materials, stimulated the development of biomedical engineering technology, established a platform for the innovation and transformation of medical instruments. Its main research fields cover medical optics, biomedical diagnostics, and rehabilitation technology.

China Unicom (Guangdong) Industrial Internet Co., Ltd

China Unicom (Guangdong) Industrial Internet Co., Ltd. is the first subsidiary with independent legal personality established by China Unicom in Guangdong Province. The company is positioned as an industrial Internet innovation service provider, with the mission of "industrial Internet expert", integrates innovative techniques such as big data, cloud computing, Internet of Things, artificial intelligence, data security, etc., and empowers thousands of industries. Up to now, the company has served more than 1,000 enterprises and more than 200 government units, promoting regional economic development.