Medical Image Analysis 99 (2025) 103307

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Validating polyp and instrument segmentation methods in colonoscopy through Medico 2020 and MedAI 2021 Challenges

Debesh Jha [a,*], Vanshali Sharma [e], Debapriya Banik [d], Debayan Bhattacharya [f], Kaushiki Roy [d], Steven A. Hicks [b], Nikhil Kumar Tomar [a], Vajira Thambawita [b], Adrian Krenzer [h], Ge-Peng Ji [l], Sahadev Poudel [j], George Batchkala [m], Saruar Alam [s], Awadelrahman M.A. Ahmed [g], Quoc-Huy Trinh [i], Zeshan Khan [n], Tien-Phat Nguyen [i], Shruti Shrestha [p], Sabari Nathan [q], Jeonghwan Gwak [r], Ritika K. Jha [a], Zheyuan Zhang [a], Alexander Schlaefer [f], Debotosh Bhattacharjee [d], M.K. Bhuyan [e], Pradip K. Das [e], Deng-Ping Fan [v], Sravanthi Parasa [o], Sharib Ali [k], Michael A. Riegler [b,c,*], Pål Halvorsen [b,c], Thomas de Lange [t,u], Ulas Bagci [a]

[a] *Machine & Hybrid Intelligence Lab, Department of Radiology, Northwestern University, Chicago, USA*
[b] *SimulaMet, Oslo, Norway*
[c] *Oslo Metropolitan University, Oslo, Norway*
[d] *Jadavpur University, Kolkata, India*
[e] *Indian Institute of Technology, Guwahati, India*
[f] *Institute of Medical Technology and Intelligent Systems, Technische Universität Hamburg, Germany*
[g] *University of Oslo, Norway*
[h] *Julius-Maximilian University of Würzburg, Germany*
[i] *Faculty of Information Technology, University of Science, VNU-HCM, Viet Nam*
[j] *Department of IT Convergence Engineering, Gachon University, Seongnam 13120, South Korea*
[k] *School of Computing, University of Leeds, LS2 9JT, Leeds, United Kingdom*
[l] *College of Engineering, Australian National University, Canberra, Australia*
[m] *Department of Engineering Science, University of Oxford, Oxford, UK*
[n] *National University of Computer and Emerging Sciences, Karachi Campus, Pakistan*
[o] *Swedish Medical Center, Seattle, USA*
[p] *NepAL Applied Mathematics and Informatics Institute for Research (NAAMII), Kathmandu, Nepal*
[q] *Couger Inc, Tokyo, Japan*
[r] *Department of Software, Korea National University of Transportation, Chungju-si, South Korea*
[s] *University of Bergen, Bergen, Norway*
[t] *Department of Medicine and Emergencies - Mölndal Sahlgrenska University Hospital, Region Västra Götaland, Sweden*
[u] *Department of Molecular and Clinical Medicin, Sahlgrenska Academy, University of Gothenburg, Sweden*
[v] *Computer Vision Lab (CVL), ETH Zurich, Zurich, Switzerland*

ARTICLE INFO

Keywords:
Colonoscopy
Polyp segmentation
Transparency
Polyp challenge
Computer-aided diagnosis
Medicine

ABSTRACT

Automatic analysis of colonoscopy images has been an active field of research motivated by the importance of early detection of precancerous polyps. However, detecting polyps during the live examination can be challenging due to various factors such as variation of skills and experience among the endoscopists, lack of attentiveness, and fatigue leading to a high polyp miss-rate. Therefore, there is a need for an automated system that can flag missed polyps during the examination and improve patient care. Deep learning has emerged as a promising solution to this challenge as it can assist endoscopists in detecting and classifying overlooked polyps and abnormalities in real time, improving the accuracy of diagnosis and enhancing treatment. In addition to the algorithm's accuracy, transparency and interpretability are crucial to explaining the whys and hows of the algorithm's prediction. Further, conclusions based on incorrect decisions may be fatal, especially in medicine. Despite these pitfalls, most algorithms are developed in private data, closed source, or proprietary software, and methods lack reproducibility. Therefore, to promote the development of efficient and transparent methods, we have organized the *"Medico automatic polyp segmentation (Medico 2020)"* and *"MedAI: Transparency in Medical Image Segmentation (MedAI 2021)"* competitions. The Medico 2020 challenge received submissions from 17

* Corresponding authors.
E-mail addresses: debesh.jha@northwestern.edu (D. Jha), michael@simula.no (M.A. Riegler).

https://doi.org/10.1016/j.media.2024.103307
Received 27 September 2023; Received in revised form 11 August 2024; Accepted 12 August 2024
Available online 5 September 2024

teams, while the MedAI 2021 challenge also gathered submissions from another 17 distinct teams in the following year. We present a comprehensive summary and analyze each contribution, highlight the strength of the best-performing methods, and discuss the possibility of clinical translations of such methods into the clinic. Our analysis revealed that the participants improved dice coefficient metrics from 0.8607 in 2020 to 0.8993 in 2021 despite adding diverse and challenging frames (containing irregular, smaller, sessile, or flat polyps), which are frequently missed during a routine clinical examination. For the instrument segmentation task, the best team obtained a mean Intersection over union metric of 0.9364. For the transparency task, a multi-disciplinary team, including expert gastroenterologists, accessed each submission and evaluated the team based on open-source practices, failure case analysis, ablation studies, usability and understandability of evaluations to gain a deeper understanding of the models' credibility for clinical deployment. The best team obtained a final transparency score of 21 out of 25. Through the comprehensive analysis of the challenge, we not only highlight the advancements in polyp and surgical instrument segmentation but also encourage subjective evaluation for building more transparent and understandable AI-based colonoscopy systems. Moreover, we discuss the need for multi-center and out-of-distribution testing to address the current limitations of the methods to reduce the cancer burden and improve patient care.

1. Introduction

Gastrointestinal (GI) cancer is a very important global health problem and the second most common cause of mortality in the United States. According to the recent 2023 estimates, there will be approximately 1,958,310 new cancer incidences and 609,820 cancer deaths in the United States (Siegel et al., 2023). Among various types of cancer, the highest number of deaths occur from lung, prostate, and colorectum in men and lung, breast, and colorectum cancer in women. As colorectal cancer is prevalent among both men and women, it is the second leading cause of cancer related death overall. One of the key indicators of colon cancer is the development of polyps in the colon and rectum. The 5-year survival rate for colon cancer is 68%, and 44% for stomach cancer (Asplund et al., 2018). If colorectal polyps are detected and removed early, the survival is close to 100. Levin et al. (2008). Thus, regular screening is crucial for early detection of these polyps, as it allows for earlier diagnosis and prompt treatment.

Endoscopic procedures, such as colonoscopy, are considered the gold standard for detecting and treating mucosal abnormalities in the GI tract (such as polyps) and cancer (Moriyama et al., 2015). However, manual screening for polyps is susceptible to error and is also time-consuming. This has driven the development of Computer Aided Detection (CADe) and Computer-Aided Diagnosis (CADx) systems that can be integrated into the clinical workflow (Riegler et al., 2016) and potentially contribute to the prevention of colorectal cancer. In the past, traditional machine learning-based CADx systems (Ballesteros et al., 2017; Hwang et al., 2007b) were popular. With the recent advancement in the hardware capabilities, such as powerful GPUs and the emergence of deep learning (LeCun et al., 2015), the research has shifted towards deep learning-based CADx systems (Fan et al., 2020; Jha et al., 2019). These algorithms have shown superior performance compared to traditional CADx solutions.

However, despite their superior performance, deep learning-based CADx systems are still considered a "black box", meaning their inner workings are not fully understood or there is a lack of transparency in understanding the predictions made by the model. Because of the complexity of multiple layers and interconnected nodes in the Convolutional Neural Network (CNN), it is challenging to interpret the decision or understand the features contributing to the outcome. For these systems to be widely adopted in clinical settings, they must be rigorously evaluated on benchmark datasets. They must demonstrate the ability to handle patient and recording device variability, provide explainability and robustness and process data in real-time. Only by carefully evaluating these systems, we can ensure the reliability and effectiveness of detecting and diagnosing cancer and its precursors (such as polyps) in a clinical setting.

In this paper, we present a comprehensive analysis of the results of the two prominent challenges in the field of automatic polyp segmentation, namely, *"Medico automatic polyp segmentation (Medico 2020)*[1]*"* challenge and the *"MedAI: Transparency in Medical Image Segmentation (MedAI 2021)"*[2] challenge. These challenges aimed to explore the potential of CADx solutions on the same shared datasets, focusing on developing novel state-of-the-art (SOTA) methods in terms of high-performance metrics, efficiency, transparency, and explainability, aiming to evaluate the relevance of such algorithms in clinical workflows. The challenges posed four distinct tasks:

- **Accurate polyp segmentation task** to develop novel algorithms to enhance the early detection and treatment of colon cancer (Medico 2020, MedAI 2021).
- **Algorithm efficiency task** to develop efficient methods with the highest frames-per-second (FPS) on predetermined hardware (Medico 2020).
- **Surgical instruments segmentation task** to enable tracking and localization of essential tools in endoscopy and help to improve targeted biopsies and surgeries in complex GI tract organs (MedAI 2021).
- **Transparency task** to evaluate different models from a transparency perspective, focusing on explanations of the training procedure, failure analysis, and (model's predictions interpretation by interdisciplinary team (MedAI 2021).

These tasks were focused on the development of SOTA algorithms for polyp and instrument segmentation in a variety of settings, including performance evaluation, resource utilization (efficiency), and transparency. By analyzing the results of these challenges, we can better understand the field's current state, identify the strength and weaknesses of different methods and find the most effective method for our problem. It is also useful to identify the research gap and areas for future innovation in the field of polyp, instrument and medical image segmentation. Fig. 1 provides an overview of both challenges along with the total number of images used for training and testing in each task. Ground truth samples with their corresponding original images are also presented for the segmentation tasks. In addition, task-specific metrics are presented (for example, FPS for "Algorithm efficiency").

In short, the main contributions are the following: (i) We present a comprehensive and detailed analysis of all participant results; (ii) we provide an overview and comparative analysis of the developed methods; (iii) we obtain and discuss new insights into the current state of AI in the field of GI endoscopy including open challenges and future directions; and (iv) finally, we provide a detailed discussion of issues such as trust, safety, interpretability, transparency, generalizability issues and multi-center in context to current limitations of CADx systems.

[1] https://multimediaeval.github.io/editions/2020/tasks/medico/

[2] https://github.com/Nordic-Machine-Intelligence/MedAI-Transparency-in-Medical-Image-Segmentation

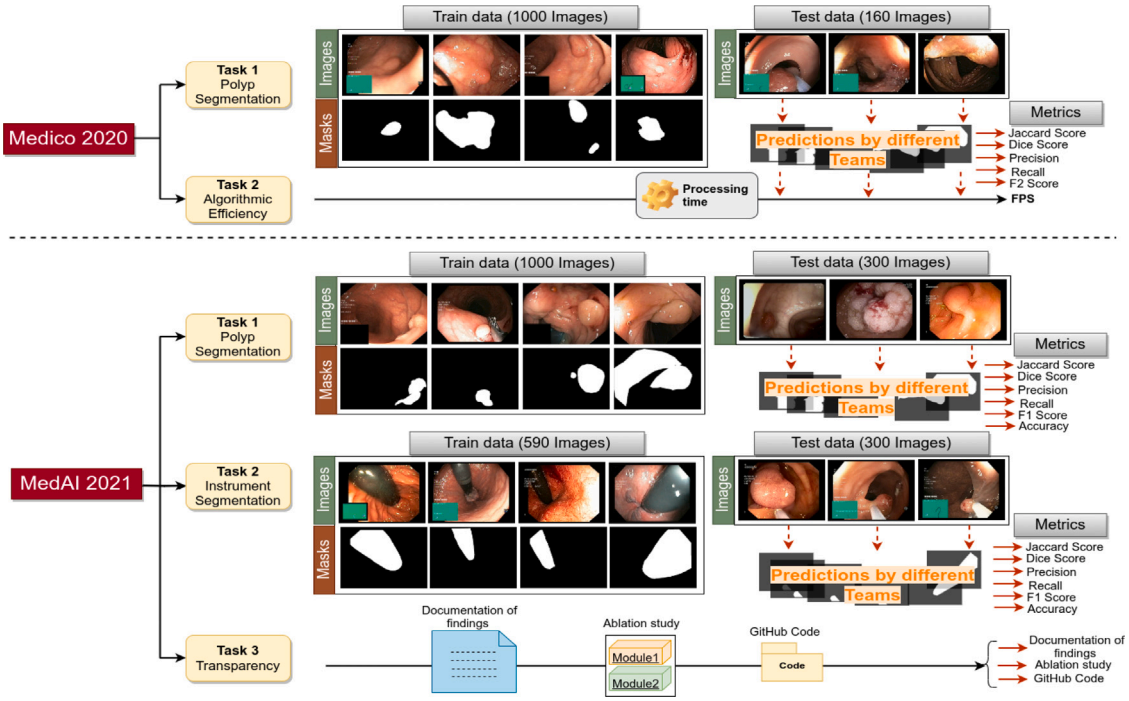


Fig. 1. The overview of the "***Medico 2020 Polyp***" and "***MedAI 2021 Transparency*** " challenges. We describe each task along with the number of training and testing datasets and the evaluation metrics used in the tasks.

2. Challenge description

2.1. Medico 2020 automatic polyp segmentation challenge

The "Medico Automatic Polyp Segmentation Challenge" was an international benchmarking challenge hosted through the MediaEval platform (Multimedia Evaluation Workshop). Medico 2020 is the fourth iteration of the Medico Multimedia Tasks series, following the pattern established in previous years. This challenge aimed to benchmark automated polyp segmentation algorithms using the same dataset and develop methods to detect difficult-to-detect polyps (such as flat, sessile, and small or diminutive polyps). Researchers from medical image analysis, machine learning, multimedia, and computer vision were invited to submit their results for this challenge, which included two tasks. The members from the organizer's institute were allowed to participate but were ineligible for receiving any recognition certificates. Participants could use any method, focusing on creating automated solutions. Below, we provide the task description of each sub-task.

(a) **Automatic Polyp Segmentation Task**: In this task, the participants were asked to develop innovative algorithms for segmenting polyps in colonoscopic images. The focus was on developing efficient systems that could accurately segment the maximum polyp area in a frame while being fast enough for practical use in a clinical setting. This task addresses the need for robust CADx solutions for colonoscopy.

To participate in the challenge, participants were required to train their segmentation models on an available training dataset. Once the test dataset was released, participants could test their models and submit their predicted segmentation maps to the organizers in a zip file with the name of each segmentation map image matching the colonoscopy image in the test dataset.

(b) **Algorithmic Efficiency Task**: CADx systems for polyp segmentation that operate in real-time can provide valuable feedback to clinicians during colonoscopy examinations, potentially reducing the risk of missing polyps and incomplete removal. However, real-time deep learning-based CADx solutions often have fewer parameters and may therefore have lower segmentation accuracy compared to more computationally intensive CADx solutions. In order to address this trade-off between accuracy and speed, the efficiency task of the challenge was designed to encourage the development of lightweight segmentation models that are both accurate and fast.

To participate in this task, participants were asked to submit docker images of their proposed algorithms. These algorithms were then evaluated on a dedicated Nvidia GeForce GTX 1080 graphics card, and the results were used to rank the teams. A mean Intersection over union (mIoU) threshold was set for considering a solution to be a valid efficient segmentation solution, and teams were ranked according to their Frames per second (FPS). By focusing on developing efficient CAD solutions, this task aimed to foster the creation of real-time systems that can provide valuable feedback to clinicians while maintaining high accuracy. A detailed description of the challenge, tasks, and evaluation metrics can be found in Jha et al. (2020a). In the supplementary material, we have provided information about the organizers for both challenges, as well as the schedule, award criteria, and publication policy.

2.2. MedAI: Transparency in medical image segmentation challenge

MedAI: Transparency in Medical Image Segmentation challenge (MedAI 2021) was held for the first time at the Nordic AI Meet[3] 2021 (Nordic young researchers symposium) that focused on medical image segmentation and transparency in Machine Learning (ML) based CADx systems. This challenge proposed three tasks to address specific endoscopic GI image segmentation challenges, including two separate segmentation scenarios and one scenario on transparent ML systems. The latter task emphasized the need for explainable and interpretable ML algorithms in the field of medical image analysis. Similar to the other challenge, participants were granted the flexibility to use any method, focusing on developing automated solutions. The members from the organizer's institute were permitted to participate but were not considered for awards.

To participate in this challenge, participants were given a training dataset to use for their ML models. These models were then tested

[3] https://nordicaimeet.com

Table 1
Overview of GI image analysis challenges with a specific focus on polyp detection, segmentation, localization, and WCE lesion detection and segmentation between 2015 and 2021. Here, WL = White Light Endoscopy, NBI = Narrow Band Imaging & WCE = Wireless Capsule Endoscopy. The total number of images and videos offered at different tasks are summed and presented in the 'Size' class.

Challenge name	Organ	Modality	Findings	Size	Dataset Availability
Automatic Polyp Detection in Colonoscopy videos 2015 (Bernal et al., 2017)	Colon	WL	Polyps	808 images & 38 videos	By request
GIANA 2017 (Bernal and Aymeric, 2017)	Colon	WL	Polyps & angiodysplasia	3,462 images & 38 videos	By request
GIANA 2018 (Angermann et al., 2017; Bernal et al., 2018)	Colon	WL, WCE	Polyps & small bowel lesions	8,262 images & 38 videos	By request
EndoCV 2021 (Ali et al., 2022a,b)	Colon	NBI, WL	Polyps	3,446 images	Open academic
Medico 2020 (Jha et al., 2020a) (Ours)	Colon	WL	Polyps	160 images (test) & 1000 images (train)	Open academic
MedAI Transparency challenge 2021 (Hicks et al., 2021) (Ours)	Colon, bladder	WL	Polyps, Instrument, Normal frames	600 images (test) & (1000 +590) images (train)	Open academic

on a concealed test dataset, allowing participants to evaluate their performance. The focus on transparency underscores the importance of developing ML algorithms that provide not only accurate and efficient results but also provide interpretable and explainable predictions. By addressing these specific challenges, this challenge aimed to foster the development of innovative and effective CADx solutions for GI endoscopy. Below, we present each challenge sub-task.

(a) Automatic Polyp Segmentation Task: In this task, participants were invited to submit segmentation masks of polyps from colonoscopic images of the large bowel. They were provided with a training dataset to develop their models, and a hidden test dataset was later released to them without the ground truth segmentation masks. Participants were required to submit a zip file containing their predicted masks in the same resolution as the input images, with the filenames of each mask matching the corresponding input image and using the ".png" file format. The objective of this task was similar to Medico 2020. By using a hidden test dataset, the results of this task were reliable and provided a valuable benchmark for the field.

(b) Automatic Instrument Segmentation Task: The instrument segmentation task required the development of algorithms that could generate segmentation masks for GI accessory instruments such as biopsy forceps or polyp snares used during live endoscopy procedures. This task aimed to create segmentation models that enable tracking and localization of essential tools in endoscopy that could aid endoscopists during interventions (such as polypectomies) by providing a precise and dense map of the instrument. Like the polyp segmentation task, participants were given a training dataset to develop their models. The submission procedure for this task was similar to that of the polyp segmentation task, with participants required to submit zip files containing their predicted masks in the same resolution as the input images and with filenames matching the corresponding input images.

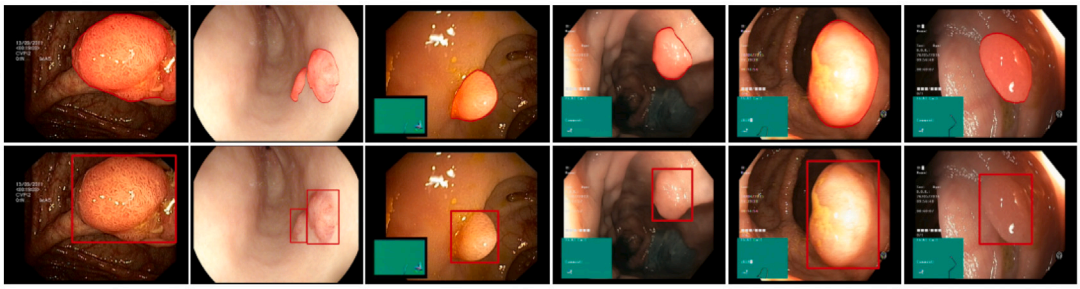
(c) Transparency Task: The transparency task focused on the importance of transparent research in medical artificial intelligence (AI). The main goal of this task was to evaluate systems from a transparency perspective, which included detailing the training procedure of the algorithms, the dataset used for training, the interpretation of the model's predictions, the use of explainable AI methods, etc. To participate in this task, researchers were encouraged to perform ablation studies, conduct a thorough failure analysis, and share their code in a GitHub repository with clear steps for reproducing the results. We allowed the participants to submit, considering the transparency and left them to decide what to deliver for the task.

In addition, participants were required to submit a one-page document summarizing their findings from the transparency task. We encouraged the participants to list package dependencies and architecture code (with instruction for building, compiling, and training) and share trained model weights in a standardized format. Additionally, we encouraged participants to include the code for model evaluation and provide repository licensing information to enable others to use the code and the trained model responsibly. Moreover, we suggested that the participants explain model predictions using intermediate heatmaps, statistical analysis and alternatives, such as SHapley Additive exPlanations (Lundberg and Lee, 2017). By promoting transparency in AI research, this task aimed to foster the development of reliable, interpretable, and trustworthy algorithms for use in medical image segmentation. A detailed description of the challenge can also be found in Hicks et al. (2021).

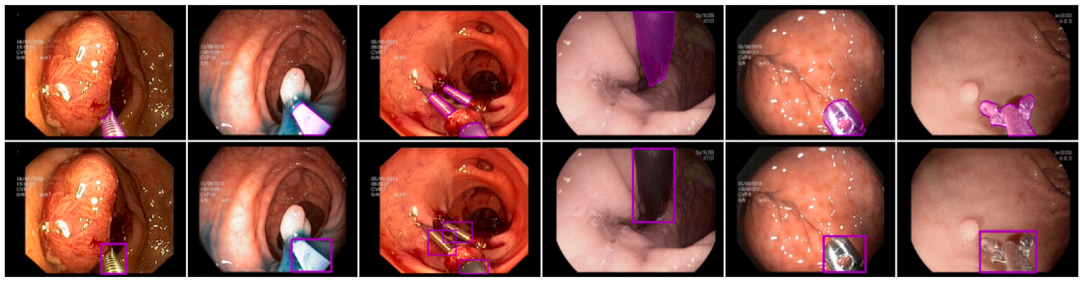
3. Related work

Polyp detection and segmentation using ML has been an active field of research for over a decade but have been previously limited by hand-crafted features (Bernal et al., 2012; Hwang et al., 2007a). Previous methods had limitations in sub-optimal performance, poor generalization to unseen images, and complexity that limited real-world applicability. However, in the recent 5–6 years, with the success of CNNs, the polyp segmentation task has seen a tremendous performance boost, including the winning model in the MICCAI challenge (Bernal et al., 2017). The widespread use of CNNs, particularly the U-Net (Ronneberger et al., 2015) and its variants, have been successfully applied on several polyp segmentation datasets and discussed in challenge reports. In addition, recent advances in CNN architectures for polyp segmentation have focused on improving convolution operations (Alam et al., 2020), adding attention blocks (Jha et al., 2019; Oktay et al., 2018), incorporating feature aggregation blocks(Mahmud et al., 2021) and using self-supervised learning techniques (Bhattacharya et al., 2021b). These modifications and learning strategies have proven effective in improving the accuracy and reliability of polyp segmentation using CNNs. Apart from the contributions of individual research groups, several challenges (Bernal et al., 2017; Ali et al., 2021) have been organized to improve the detection and classification of mucosal abnormalities in the GI tract from either single image frames or videos. However, the dataset provided in the challenge and the details of the proposed algorithms are often not publicly available, making it difficult to reproduce and build upon them. Hence, there is a need for open-access benchmarking datasets and reproducible algorithms to facilitate progress in this field.

Table 1 provides an overview of GI image analysis challenges held in the past eight years. The challenge was conducted using images



(a) Examples samples from the test data of Medico 2020 (first three columns) and MedAI 2021 (last three columns) for the polyp segmentation task.



(b) Example samples from the MedAI 2021 Instrument segmentation task.

Fig. 2. Example of the test datasets from the Medico 2020 and MedAI 2021 datasets.

from different modalities with a specific focus on polyp segmentation, detection, localization and wireless capsule endoscopy lesion detection and localization. In 2015, Bernal et al. (2017) organized the "Automatic Polyp Detection in colonoscopy videos" challenge. Likewise, they organized the GIANA challenge in 2017 and 2018[4] focused on colonoscopy data and included tasks such as detection of lesions in Video capsule Endoscopy (VCE), polyp detection, and polyp segmentation. Recent challenges attempted to address generalizability in polyp detection and segmentation (Ali et al., 2022a) with single frames and sequence colonoscopy datasets. They demonstrated how variability in images can affect algorithm performances. Altogether, these challenges have led to many algorithmic innovations in detecting and classifying GI abnormalities (especially polyp segmentation and detection).

Additionally, past challenges have not emphasized on the explainability and reliability of deep learning model predictions. Most challenges also do not focus on open source codes for research and development, making it difficult for proposed algorithms to be adopted in clinical settings due to a lack of transparency. Moreover, the reported methods are not reproducible, which hinders further algorithmic advancement. Thus, we lose track of what are best practices and where we are heading in this field. Through our challenges in Medico 2020 and MedAI 2021, we address reproducibility and open science which are the two most important aspects that can enable experienced and new ML scientists to build upon and advance the field.

4. Challenge datasets and evaluation metrics

4.1. Medico 2020 dataset

The training dataset contains 1000 polyp images and their corresponding ground truth mask taken from Kvasir-SEG (Jha et al., 2020b). Kvasir-SEG consists of diverse images varying in appearance, such as sizes (for example, diminutive, regular or large), colors (same color as mucosa, or different colors such as reddish), textures (smooth or granular), locations (anywhere in large intestine such as left colon, sigmoid colon or rectum), numbers of polyp per images (for example, one or many), image quality (illumination, artifacts) and shapes (flat, pedunculated, and sessile). The variation ensures that the algorithms trained on this dataset can handle real-world variations in clinical settings. Some samples are shown in Fig. 2(a).

The datasets were acquired from real routine clinical examinations at Vestre Viken Health Trust (VV) in Norway by a team of expert gastroenterologists. The VV is the collaboration of the four hospitals that provide healthcare services to 470,000 people. The resolution of images varies from 332 × 487 to 1920 × 1072 pixels. Some images contain green thumbnails in the lower-left corner of the images showing the position marking from the ScopeGuide (Olympus). After data acquisition, our team categorized the dataset into a polyp class. To extend the dataset to the segmentation class, a team of one experienced engineer, a medical doctor, and an expert gastroenterologist annotated the polyp images using the label box tool. After annotation, we extract the corresponding ground truth and bounding box information. Once the ground truth was created, the images and ground truths were combined to facilitate the review process. These images were sent to a team of expert gastroenterologists for validation through a web-based interface. After validation, we compiled them into training and test datasets. The data proportion for each set followed the general split ratio used in the literature.

The training dataset has been made publicly available as open access and is widely available at.[5] The test dataset contains unique polyp images encompassing a wide range of diverse clinical scenarios with different polyp characteristics, varying lighting conditions and image resolution, low-quality images, as well as complex polyp images (for example, with instruments and residual stool) that the model has never encountered before. Only the organizers had access to the test case labels. Currently, the test data can be downloaded from.[6]

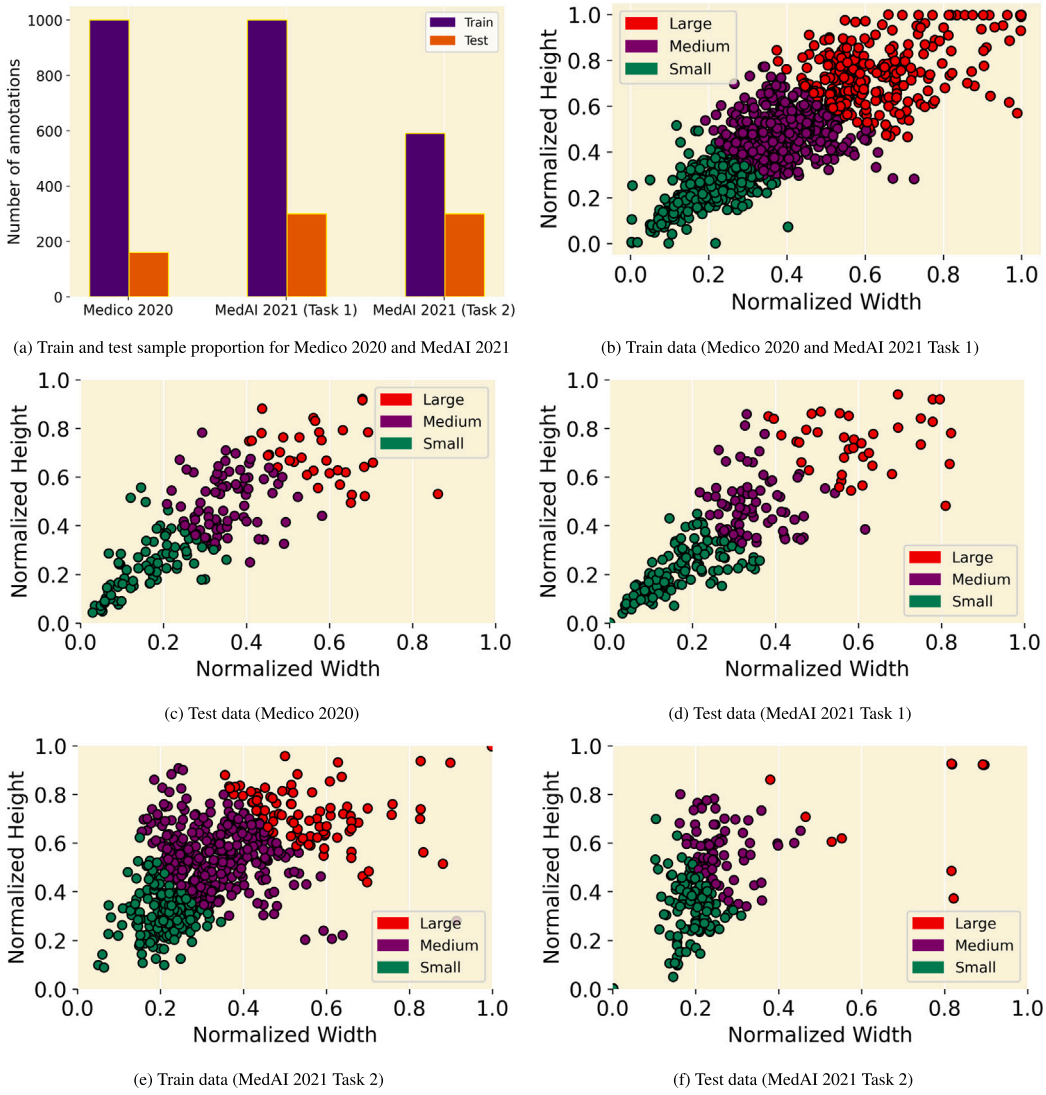
4.2. MedAI transparency challenge 2021 dataset

We utilize our Kvasir-SEG (Jha et al., 2020b) as the development dataset for the polyp segmentation task. Similarly, Kvasir-Instrument

[4] https://giana.grand-challenge.org/

[5] https://datasets.simula.no/kvasir-seg/

[6] https://drive.google.com/file/d/1uP2W2g0iCCS3T6Cf7TPmNdSX4gayOrv2



(a) Train and test sample proportion for Medico 2020 and MedAI 2021

(b) Train data (Medico 2020 and MedAI 2021 Task 1)

(c) Test data (Medico 2020)

(d) Test data (MedAI 2021 Task 1)

(e) Train data (MedAI 2021 Task 2)

(f) Test data (MedAI 2021 Task 2)

Fig. 3. Data distribution details of train and test sets used in Medico 2020 and MedAI 2021 challenges. Large, medium, and small represent the distribution information of regions of interest in the data samples.

(Jha et al., 2021) was used as the training dataset for the instrument segmentation task. It can be downloaded from.[7] We followed the same data acquisition and annotation protocol for test dataset creation as the Medico 2020 challenge. Some sample images for polyp segmentation and instrument segmentation tasks are presented in Fig. 2(a) and Fig. 2(b). Fig. 3 shows the data distribution of the train and test datasets used in Medico 2020 and MedAI 2021. We have categorized the images into "small", "medium" and "large" according to the size of regions of interest using a randomly selected threshold of 0.3 and 0.1 and plotted the normalized height versus normalized width of each data point. This is to visualize the dimension of each data point and observe the diversity and complexity of the dataset used in the study. The information about the size categories and the dataset's dimensions is crucial for assessing the performance and robustness of the proposed algorithms.

4.3. Metrics for polyp and instrument segmentation tasks

We used mean Intersection over Union (mIoU) as a primary evaluation metric for the polyp and instrument segmentation tasks. If the teams achieved the same mIoU values, their ranking was further evaluated based on the higher value of the Dice coefficient (DSC). We also recommend calculating other important standard evaluation metrics that hold significant relevance in clinical settings such as Accuracy (Acc), Recall (Rec), Precision (Pre), F-2 score, and Frames per second (FPS) to ensure a detailed evaluation.

4.4. Metrics for efficiency tasks

Efficiency is crucial in colonoscopy as it directly impacts the models' feasibility and practicality in real-world scenarios. Endoscopists often need to analyze numerous frames in real-time during routine colonoscopy, and lag (latency) in the analysis could lead to suboptimal results. Our approach to FPS calculation was based on the time taken to process a single image, averaged over the entire dataset, and then extrapolated to a per-second rate. Therefore, we strongly recommend calculating processing speed in terms of FPS.

4.5. Metrics for transparency tasks

The transparency task aimed to assess the transparency and understandability of algorithms for medical AI by utilizing a qualitative

[7] https://datasets.simula.no/kvasir-instrument/

Table 2
Summary information of participating teams in Medico 2020. Here, '√' = Team participated, '–' = No participation, Task 1 = Polyp segmentation task and Task 2 = Algorithm efficiency task.

Chal.	Team name	Task 1	Task 2
Medico 2020	FAST-NU-DS	√	√
	AI-TCE	√	–
	ML-MMIVSARUAR	√	–
	UiO-Zero	√	–
	HBKU_UNITN_SIMULA	√	–
	AI-JMU	√	√
	SBS	√	√
	AMI Lab	√	√
	UNITRK	√	√
	MedSeg_JU	√	–
	IIAI-Med	√	–
	HGV-HCMUS	√	√
	GeorgeBatch	√	√
	PRML2020GU	√	√
	VT	√	–
	IRIS-NSYSU	√	–
	NKT	√	√

approach in the evaluation metrics. We evaluated transparency tasks using a more quantitative approach than polyp and instrument segmentation. A multi-disciplinary team assessed each submission and evaluated the transparency and understandability of the proposed solutions. Each team was scored based on the three criteria: open source code, model evaluation and clinical evaluation. The open source code was evaluated based on the presence of a publicly available repository, code quality and quality of the readme file. The model evaluation included failure analysis, ablation study, explainability of the method, and metrics used. Evaluation by clinical experts considered the usefulness of the technique and its understandability. With these three criteria, we aimed to measure the transparency of the provided solutions. A detailed score distribution under different criteria is shown in Table 10, which was part of our Task 3. Ultimately, this task aimed to promote the development of more transparent and interpretable AI systems.

5. Participating research teams

5.1. Methods used in Medico 2020

Table 2 summarizes all the teams participating in the "Medico 2020" challenge. It can be seen from Table 2 that all 17 teams participated in Task 1, whereas only 9 teams participated in Task 2.

FAST-NU-DS: Team FAST-NU-DS (Ali et al., 2020a) explored the advantage of using depth-wise separable convolution in the atrous convolution of the ResUNet++(Jha et al., 2019) architecture. Modifications were made to get the lightweight image segmentation. Deep atrous spatial pyramid pooling was also implemented on the ResUNet++ architecture. The purpose of this architectural design was to provide good performance on the image segmentation evaluation metrics and inference time. To get the lightweight model architecture, changes were made to the atrous bridge in ResUNet++ architecture. The convolution layer in the atrous bridge was replaced with depthwise separable convolution. Depth-wise separable convolution first applies channel-wise filters, followed by a 1×1 pointwise convolution, to maintain performance while streamlining computations. The implementation of depth-wise separable convolution resulted in less number of parameters and giga-floating point operations (GFLOPs).

AI-TCE: Team AI-TCE (Nathan and Ramamoorthy, 2020) proposed an efficient supervision network that uses EfficientNet (Tan and Le, 2019a) and an attention Unit. The proposed network had the properties of an encoder–decoder structure with supervision layers. An EfficientNet-B4 was used as a pre-trained architecture in the encoder block. The decoder block combined dense block and Concurrent Spatial and Channel Attention block. Both the encoder and decoder were connected by Convolution Block Attention Module (CBAM). All the outputs of the decoder layer were supervised, i.e., individual decoder output was taken and upsampled with the output layer and supervised by the loss function. Also, all upsampled outputs were concatenated and fed into CBAM. In the upsampling, the convolution transpose layer was used.

ML-MMIV SARUAR: Team ML-MMIV SARUAR (Alam et al., 2020) used the U-Net with pre-trained ResNet50 on the ImageNet dataset as the encoder for the polyp segmentation task. The use of a pre-trained encoder helped the model to converge easily. The input image was fed into the pre-trained ResNet50 encoder, consisting of a series of residual blocks as their main component. These residual blocks helped the encoder extract the important features from the input image, which were then passed to the decoder. Skip connections between the encoder and decoder branch help the model to get all the low-level semantic information from the encoder, which allowed the decoder to generate the desired feature maps.

UiO-Zero: Team UiO-Zero (Ahmed and Ali, 2020) used the generative adversarial networks (GAN) framework for solving the automatic segmentation problem. Perceiving the problem as an image-to-image translation task, conditional GANs were utilized to generate masks conditioned by the images as inputs. The polyp segmentation GAN-based model consists of two networks, namely a generator and discriminator, that were based on convolution neural networks. A generator takes the images as input and tries to produce realistic-looking masks conditioned by this input and a discriminator, which was basically a classifier that had access to the ground truth masks and tried to classify whether the generated masks was real or not. To stabilize the training, the images were concatenated with the masks (generated or real) before being fed to the discriminator.

HGV-HCMUS: The HGV-HCMUS (Trinh et al., 2020) team proposed methods combining the Residual module, Inception module, Adaptive CNN with U-Net (Ronneberger et al., 2015) model, and PraNet (Fan et al., 2020) for semantic segmentation of various types of polyps in endoscopic images. The team submitted five different runs considering five different solutions. In the first approach, a simple U-Net architecture was adopted to parse masks of polyps. Second, the regular ReLU was replaced with Leaky ReLU to deal with dead neurons. Third, to further boost the result, an Inception module was introduced to extract better features. Fourth, a pre-trained model with the ResNet-50 backbone was used to build ResUNet, yielding better obtained results. Last, PraNet was employed for polyp segmentation in colonoscopy images. This solution provided the best outcome and was used to generate the results.

AI-JMU: Team AI-JMU (Krenzer and Puppe, 2020) explored various image segmentation models, specifically the Cascade Mask R-CNN (Cai and Vasconcelos, 2019) and Mask R-CNN (He et al., 2017) with ResNet (He et al., 2016) as well as the ResNeSt (Zhang et al., 2022) architectures was used as the backbone. Additionally, the team investigated the effect of varying the depth of both the ResNet and ResNeSt architectures. Depths of 50, 101, and 200 were evaluated for the ResNeSt model, and depths of 50 and 101 for the ResNet model. They reported that the best outcome was obtained using ResNeSt-101 when combined with Cascade Mask R-CNN.

SBS: Team SBS (Shrestha et al., 2020) exploited ResNet 34 (He et al., 2016) and EfficientNet-B2 (Tan and Le, 2019a) backbones in the U-Net. The team introduced two different models: Single Model and Ensemble Model. The ResNet-34 was used in the single model. The weights saved after the training phase was loaded in the network, and test data were fed to get the predicted polyp masks. However, in the case of the ensemble model, both ResNet-34 and EfficientNetB2 were used to predict the masks. Then the individual prediction was ensembled using bitwise multiplication between the two predicted masks. The ensemble model provided better evaluation results as compared to the single model, as when multiple algorithms were ensembled predictive power

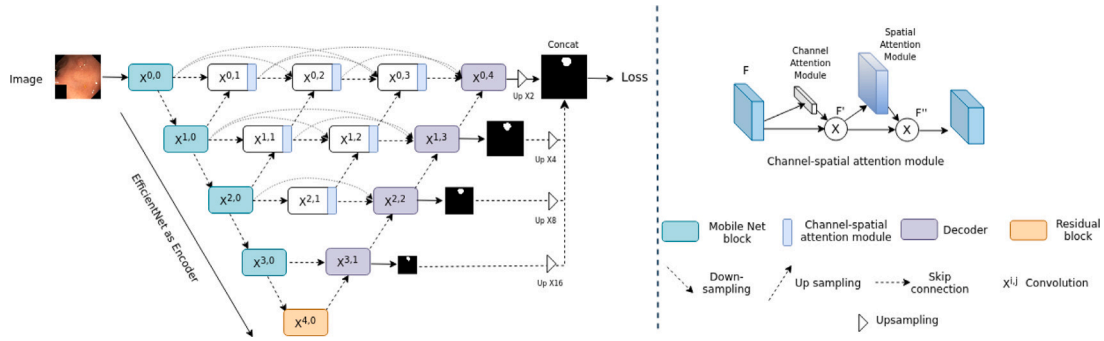


Fig. 4. Overview of the winning solution for the Polyp segmentation task (Task 1) from Team ***PRML2020GU***. The architecture utilizes pre-trained weights from EfficientNet in the encoder. Additionally, it uses dense skip connections, deep supervision and channel-spatial attention for fast convergence and better performance.

increases and error rate decreases. Hence, the final results are reported using the ensemble model using ResNet-34 and EffiecientNetB2 as backbones in the U-Net architecture.

AMI Lab: Team AMI Lab (Kang and Gwak, 2020) utilized the knowledge distillation technique to improve ResUNet++ (Jha et al., 2019), which performs well on automatic polyp segmentation. First, the data augmentation module was used to generate augmented images for the input. Second, the obtained augmented images were fed to both the student model and the teacher model. Third, the distillation loss between the outputs of student and teacher models was calculated. Similarly, the loss between the output of the student model and the ground truth label was computed to train the student model.

UNITRK: Team UNITRK (Khadka, 2020) employed the UNet model pre-trained on the brain MRI dataset. The notion of knowledge transfer has been the key motivating factor to choose a simple pre-trained model. The model was fine-tuned with the polyp dataset. The fine-tuning of the pre-trained model helped to converge faster without the requirement of a large number of training examples. The additive soft attention mechanism was integrated with the pre-trained UNet architecture. The key benefit of this attention UNet structure in comparison to multi-stage CNNs was that it does not require training of multiple models to deal with object localization and thus reduces the number of model parameters. It helps to focus on relevant regions in the input images.

MedSeg_JU: Team MedSeg_JU (Banik and Bhattacharjee, 2020) proposed an approach for polyp segmentation based on deep conditional adversarial learning. The proposed framework consists of two interdependent modules: a generator network and a discriminator network. The generator was an encoder–decoder network responsible to predict the polyp mask while the discriminator enforces the segmentation to be as similar to the ground truth segmented mask. The training process of the network alternates between training the generator and the discriminator, with the generator trained to produce a predicted synthetic mask by freezing the discriminator and the discriminator trained while freezing the generator.

IIAI-Med: Team IIAI-Med team (Ji et al., 2020) presented a novel deep neural network, called the Parallel Reverse Attention Network (PraNet) (Fan et al., 2020), for the task of automatic polyp segmentation at MediaEval 2020. The network first aggregated features in high-level layers using a parallel partial decoder (PPD). This combined feature was then used to generate a global map as the initial guidance area for the following components. Additionally, the network mines boundary cues using a reverse attention (RA) module which establishes the relationship between areas and boundary cues. Thanks to the recurrent cooperation mechanism between areas and boundaries, the PraNet was able to calibrate misaligned predictions, improving segmentation accuracy and achieving real-time efficiency (nearly 30fps). The code and results are available at https://github.com/GewelsJI/MediaEval2020-IIAI-Med.

HBKU_UNITN_SIMULA Team HBKU_UNITN_SIMULA (Nguyen et al., 2020) proposed two different approaches leveraging the advantages of either ResUNet++ or PraNet model to efficiently segment polyps in colonoscopy images, with modifications on the network structure, parameters, and training strategies to tackle various observed characteristics of the given dataset. For the first approach, PraNet was used, which is a parallel reverse attention network that helps to analyze and use the relationship between areas and boundary cues for accurate polyp segmentation. The PraNet with Training Signal Annealing strategy was used to improve segmentation accuracy and effectively train from scratch on the given small dataset. For the second approach, ResUNet++ was used, which takes advantage of residual blocks, squeeze and excitation blocks, atrous spatial pyramid pooling, and attention blocks. The input path was modified and integrates a guided mask layer to the original structure for better segmentation accuracy. They used the two approaches to experiment with different runs. The best polyp segmentation outcome was achieved when the results from three PraNet and five ResUNet++ models, trained on different train-val dataset splits, were averaged.

GeorgeBatch: Team GeorgeBatch (Batchkala and Ali, 2020) used the standard U-Net architecture for the binary segmentation task, and experiments were conducted using the intersection-over-union loss (IoU loss) instead of the commonly used binary cross-entropy (BCE) loss. They also experiment with a combination of both losses in the training process. The motivation behind this approach was to strike a balance between accuracy and speed for using automated systems during colon cancer surveillance and surgical removal of polyps. This balance is considered while experimenting with other parameters like loss function and data augmentation to boost performance. The reported outcomes show that using IoU loss results in enhanced segmentation performance, with a nearly 3% improvement on the DSC metric while maintaining real-time performance (close to 200 FPS). The code and results are available at https://github.com/GeorgeBatch/kvasir-seg.

PRML2020GU: An overview of the approach proposed by team PRML2020GU (Poudel and Lee, 2020) is shown in Fig. 4. The team employed an EfficientNetB3 as an encoder backbone with a U-Net decoder and leveraged the concept of U-Net++ of redesigning the skip connections to use multi-scale semantic details. The densely connected skip connections to the decoder side enable flexible multi-scale feature fusion both horizontally and vertically at the same resolution. Besides, the proposed method is powered by deep supervision, where all the outputs after deep supervision is averaged, and the final mask is generated. Further, channel-spatial attention enables significantly better performance and fast convergence. Moreover, integrating the channel and spatial attention modules restrains irrelevant features and allows only useful spatial details.

VT: Team VT (Thambawita et al., 2020) proposed a simple but efficient idea of using an augmentation method called pyramid focus-augmentation (PYRA) that uses grids in a pyramid-like manner (large to small) for polyp segmentation. The method has two main steps:

Table 3
Summary of the participating teams algorithm for Medico 2020. Here, "Aug." = augmentation used, "SGD" = Stochastic gradient descent, "GAN" = generative adversarial network, "ASPP" = Atrous Spatial Pyramid Pooling, and "AP" = Average precision.

Team name	Algorithm	Backbone	Nature	Choice basis	Aug.	Loss	Optimizer
FAST-NU-DS (Ali et al., 2020a)	Depth-wise separable convolution and ASPP	ResUNet++	Cascade of depth-wise separable convolutions	mIoU and DSC	Yes	IoU	Adam
AI-TCE (Nathan and Ramamoorthy, 2020)	Multi-Supervision Net	EfficientNetB4	Encoder-Multi Supervision Decoder	Acc and DSC	Yes	Categorical cross-entropy + Dice loss	Adam
ML-MMIV SARUAR (Alam et al., 2020)	Encoder–decoder based architecture based on ResNet50	ResNet50	Cascade of residual blocks	mIoU and DSC	Yes	Cross-entropy	Adam
UiO-Zero (Ahmed and Ali, 2020)	GAN	None	GAN with CNN based generator and discriminator	Image-to-image translation	No	Standard conditional GAN adversarial loss	Adam
HBKU UNITN SIMULA (Nguyen et al., 2020)	Residual module, Inception module, Adaptive CNN with U-Net and PraNet	U-Net and ResNet-50	Cascade of residual blocks and inception module	mIoU and DSC	Yes	Bce + Dsc loss	Adam
AI-JMU (Krenzer and Puppe, 2020)	Cascade Mask R-CNN	ResNeSt backbone, Cascade Architecture	Deep CNN	DSC and mIoU	Yes	Binary cross-entropy	SGD
SSB (Shrestha et al., 2020)	U-Net	ResNet-34, EfficientNet-B2	Ensemble	DSC and mIoU	Yes	Tversky loss	Adam
AMI LAB (Kang and Gwak, 2020)	Knowledge distillation on ResUNet++	ResUNet++	Ensemble	mIoU and DSC	Yes	Distillation loss	Adam
UNITRK (Khadka, 2020)	Knowledge transfer using UNet	Pre-trained U-Net model	Encoder–decoder	mIoU and DSC	Yes	Compound loss of DSC and BCE	Adam
MedSeg_JU (Banik and Bhattacharjee, 2020)	Conditional GAN (cGAN)	None	Encoder–decoder	mIoU and DSC	Yes	Weighted loss of MSE and BCE	Adam
IIAI-Med (Ji et al., 2020)	PraNet	Res2Net	Encoder–decoder	mIoU, DSC and FPS	No	Weighted IoU loss + BCE loss	Adam
HGV-HCMUS (Trinh et al., 2020)	PraNet and ResUNet++ with triple path	ResUNet++	Encoder–decoder	mIoU	Yes	Categorical crossentropy	Adam
GeorgeBatch (Batchkala and Ali, 2020)	U-Net	None	Encoder–decoder	Acc and Speed	Yes	Non-Binarized IoU	Adam
PRML20202GU (Poudel and Lee, 2020)	Efficient-UNet +Channel-Spatial Attention + Deep Supervision	Variants of EfficientNet	Encoder–decoder	mIoU and DSC	Yes	BCE + DSC loss	Adam
VT (Thambawita et al., 2020)	U-Net coupled with PYRA	None	Encoder–decoder	mIoU and DSC	Yes	BCEWithLogits Loss	RMSprop
IRISNSYSU (Maxwell Hwang et al., 2020)	Temporal–Spatial Attention Model	Faster-RCNN	Hybrid attention interface	AP	Yes	Cross entropy	Adam
NTK (Tomar, 2021)	Residual blocks combined with SE network	None	Encoder–decoder	DSC, mIoU and FPS	No	BCE + DSC loss	Adam

data augmentation with PYRA using pre-defined grid sizes followed by training of a DL model with the resulting augmented data. PYRA can be used to improve the performance of segmentation tasks when there is a small dataset to train the DL models or if the number of positive findings is small. The method shows a large benefit in the medical diagnosis use case by focusing the clinician's attention on regions with findings step-by-step.

IRISNSYSU: Team IRISNSYSU (Maxwell Hwang et al., 2020) proposed a local region model with attentive temporal–spatial pathways for automatically learning various target structures. The attentive spatial pathway highlights the salient region to generate bounding boxes and ignores irrelevant regions in an input image. The proposed attention mechanism allows efficient object localization, and the overall predictive performance is increased because there are fewer false positives for the object detection task for medical images with manual annotations.

NKT: Team NKT (Tomar, 2021) proposed a full convolution network following an encoder–decoder approach. It combines the strength of residual learning and the attention mechanism of the squeeze and excitation (SE) network. The encoding network consists of 4 encoder blocks with 32, 64, 128, and 256 filters. The decoding network also consists of 4 decoder blocks with 128, 64, 32, and 16 filters. Both the encoder and decoder block consist of a residual block as their core component. The residual block helps in building deep neural networks by solving the vanishing gradient and exploding gradient problem.

Additionally, in Table 3, we provide an elaborate summary of all the research teams who participated in the "Medico 2020" challenge. It gives a detailed overview of the algorithms, backbone, nature, choice basis, data augmentation used, loss function, and optimizer used by each participating teams.

5.2. Methods used in MedAI 2021

In this subsection, we briefly summarize the methods used by the participating teams in the MedAI 2021 challenge. In Table 4, we present the research teams who have participated in each of these three tasks. It can be seen from this table that most of the teams participated in all three tasks except for three teams, which participated in either one

Table 4
Summary information of participating teams in MedAI 2021. Here, '√' = Team participated, '–' = No participation, Task 1 = Polyp segmentation task, Task 2 = Instrument segmentation task, and Task 3 = Transparency task. A total of 16 teams participated in polyp segmentation and instrument segmentation and 14 teams participated in the Transparency tasks in the challenge.

Chal.	Team name	Task 1	Task 2	Task 3
MedAI 2021	The Segmentors	√	√	√
	The Arctic	√	√	√
	mTEC	√	√	√
	MedSeg_JU	–	√	–
	MAHUNM	√	√	√
	IIAI-CV&Med	√	√	√
	NYCity	√	√	√
	PRML	√	√	√
	leen	√	√	√
	CV&Med IIAI	√	√	√
	Polypixel	√	√	√
	agaldran	√	√	√
	TeamAIKitchen	√	√	√
	CamAI	√	√	√
	OXGastroVision	√	√	√
	Vyobotics	√	–	–
	NAAMII	√	√	–

or two of the sub-tasks. All participating teams have used the same architecture in their submission for polyp segmentation and instrument segmentation subtasks. However, two teams, namely *Vyobotics* (Rauniyar et al., 2021) and *MedSeg_JU* (Banik et al., 2021) have participated in only one of the subtasks. The team *Vyobotics* (Rauniyar et al., 2021) has participated in the polyp segmentation task whereas the team *MedSeg_JU* (Banik et al., 2021) has participated in the surgical instrument segmentation task.

The Segmentors: Team Segmentors (Mirza and Rajak, 2021) proposed solution is a UNet-based algorithm designed for segmenting polyps in images taken from endoscopies. The primary focus of this approach was to achieve high segmentation metrics on the supplied test dataset, which was a crucial requirement for accurate and reliable polyp segmentation. To this end, they experimented with data augmentation and model tuning to achieve satisfactory results on the test sets.

The Arctic: Team Arctic (Somani et al., 2021) utilized a unique hybrid optimization technique that combined the power of DeepLabV3+ (Chen et al., 2018) and ResNet101 (He et al., 2016) to address the specific challenges of GI image segmentation effectively. In order to ensure the accuracy of their results, the team employed a 5-fold cross-validation approach, with a learning rate of 0.0001 and a batch size of 12. Additionally, towards transparency, they proposed a method of rendering feature attention maps to visualize the attention of the network on individual pixels within the image.

mTEC: Team mTEC (Bhattacharya et al., 2021a) introduced a new architecture called Dual Parallel Reverse Attention Edge Network (DPRA-EdgeNet) for joint segmentation of polyp masks and polyp edge masks. This architecture utilizes the reverse attention module from PraNet (Fan et al., 2020) to perform the segmentation tasks. The team implemented two parallel decoder blocks, with one focused on extracting features for polyp segmentation and the other focused on extracting features for polyp edge segmentation. The polyp mask decoder leverages the features from the edge decoder block to improve the accuracy of the segmentation. Additionally, the team employed deep supervision of both edge and polyp features to stabilize the optimization process of the model.

MedSeg_JU: Team MedSeg_JU (Banik et al., 2021) proposed EM-Net, encoder–decoder-based architecture inspired by the M-Net (Mehta and Sivaswamy, 2017) architecture. In their approach, the encoder branch of the network utilized EfficientNet-B3 (Tan and Le, 2019b) as its backbone. The network also employed a multi-scale input method, where the input image was downsampled at rates of 2, 4, and 8 at each level of the encoder branch, providing a multi-level receptive field. The decoder branch was a mirror structure of the encoder, where upsampling was used to increase the size of the feature maps at each level. Skip connections were used to enhance the flow of spatial information lost during downsampling. The final feature maps underwent pointwise convolution and sigmoid activation and were then upsampled to provide deep supervision and a local pixel-level prediction map for each scale of the input image. These maps were then fused to generate the final segmentation mask.

MAHUNM: Team MAHUNM (Haithami et al., 2021) presented an approach for enhancing the segmentation capabilities of DeeplabV3 by incorporating Gated Recurrent Neural Network (GRU). In their approach, the team replaced the 1-by-1 convolution in DeeplabV3 with GRU after the ASSP layer to combine input feature maps. While the convolution and GRU had sharable parameters, the latter had gates that enabled or disabled the contribution of each input feature map. The experimental evaluation conducted on unseen test sets demonstrated that using GRU instead of convolution produced better segmentation results.

IIAI-CV&Med: Team IIAI-CV&Med (Dong et al., 2021b) developed an ensemble of three sub-models, namely Polyp-PVT (Dong et al., 2021a), Sinv2-PVT, and Transfuse-PVT. The official Polyp-PVT, as designed for polyp segmentation, was adopted without modification and achieved state-of-the-art segmentation capability and generalization performance. Transfuse, also designed for polyp segmentation, was improved by replacing the transformer part with the pyramid vision transformer (PVT) (Wang et al., 2022) to enhance its performance. The official Sinv2 (Fan et al., 2021), which proposes an end-to-end network for searching and recognizing concealed objects, was employed and its original backbone of Res2Net was replaced with a stronger PVT transformer (Wang et al., 2022) to extract more meaningful features.

NYCity: Team NYCity (Chen et al., 2021) presented a novel multi-model ensemble framework. The team first collected a set of SOTA models in this field and further improved them through a series of refinements. These models include TransFuse (Zhang et al., 2021) and HarDNet-MSEG (Huang et al., 2021). They improvised TransFuse by replacing its backbone with HarDNet-85 (Chao et al., 2019) and placing an additional BiFuse layer. They further modified HarDNet-MSEG by using HarDNet-85 and ResNet-101 (He et al., 2016) as the backbone. Additionally, they made modifications to the decoder and adopted different receptive fields. By integrating those fine-tuned models into a more powerful ensemble framework, they were able to achieve improved performance.

PRML: Team PRML (Poudel and Lee, 2021) introduced Ef-UNet, a segmentation model that is composed of two main components. First, a U-Net encoder that utilizes EfficientNet (Tan and Le, 2019b) as a backbone, which allows the generation of different semantic details in multiple stages. Second, a decoder integrates spatial information from different stages to generate a final precise segmentation mask. Using EfficientNet as the encoder backbone provides Ef-UNet with the ability to efficiently extract high-level features from the input images while the decoder component effectively integrates these features to produce accurate segmentation results.

leen: Team leen (Ahmed and Ali, 2021) utilized the GANs framework to produce corresponding masks that locate the polyps or instruments on GI polyp images. To ensure transparency and explainability of their models, the team leen adopted the layer-wise relevance propagation (LRP) approach (Bach et al., 2015), which is one of the most widely used methods in explainable artificial intelligence. This approach generated relevant maps that display the contribution of each pixel of the input image in the final decision of the model.

CV&Med IIAI: Team CV&Med IIAI (Chou, 2021) proposed a novel dual model filtering (DMF) strategy, which effectively removed false positive predictions in negative samples through the use of a metrics-based threshold setting. To better adapt to high-resolution input with various distributions, the PVTv2 (Wang et al., 2022) backbone was

Table 5
Summary of the participating teams algorithm for MedAI 2021.

Team name	Segmentation task	Algorithm	Backbone	Nature	Choice basis	Augmentation	Loss	Optimizer
The Segmentors (Mirza and Rajak, 2021)	Polyp, Instrument	U-Net	None	Encoder–decoder	DSC and mIoU	Yes	DSC	Adam
The Arctic (Somani et al., 2021)	Polyp, Instrument	DeeplabV3plus + ResNet101	None	Hybrid	DSC	Yes	Cross-entropy	Adam
mTEC (Bhattacharya et al., 2021a)	Polyp, Instrument	DPRA-EdgeNet	HarDNet	Cascade	DSC and mIoU	No	(Dice + BCE) loss	Adam
MedSeg_JU (Banik et al., 2021)	Instrument	EM-Net	EfficientNet-B3	Encoder–decoder	DSC	Yes	DSC	Adam
MAHUNM (Haithami et al., 2021)	Polyp, Instrument	DeeplabV3 with GRU	ResNet-50/ResNet-101	Sequential	DSC and mIoU	No	BCE With Logits Loss	Adam
IIAI-CV&Med (Dong et al., 2021b)	Polyp, Instrument	Polyp-PVT, Sinv2-PVT and Transfuse-PVT	Transformer	Ensemble	Majority voting	No	IoU	Adam
NYCity (Chen et al., 2021)	Polyp, Instrument	HarDNet-85, ResNet-101	Transformer	Ensemble	Accuracy	Yes	IoU	Gradient centralization
PRML (Poudel and Lee, 2021)	Polyp, Instrument	Ef-UNet	EfficientNet	Encoder–decoder	DSC and mIoU	No	DSC Loss	Adam
leen (Ahmed and Ali, 2021)	Polyp, Instrument	GAN	None	Encoder–decoder	DSC and mIoU	No	BCE and L1 loss	Adam
CV&Med IIAI (Chou, 2021)	Polyp, Instrument	SINetv2	PVT v2	Encoder–decoder	mIoU	No	Pixel position-aware loss	Adam
Polypixel (Tzavara and Singstad, 2021)	Polyp, Instrument	Transfer learning using EfficientNet B1	None	CNN	DSC and mIoU	Yes	IoU	Adam
agaldran (Galdran, 2021)	Polyp, Instrument	Double Encoder–Decoder with temperature scaling Feature	Pyramid Network as Decoder and Resnext101 as pretrained decoder	Sequential	DSC	Yes	DSC	Sharpness-aware minimization (SAM) + Adam
TeamAIKitchen (Keprate and Pandey, 2021)	Polyp, Instrument	U-Net	None	Encoder–decoder	DSC	Yes	DSC	Adam
CamAI (Yeung, 2021)	Polyp, Instrument	Transfer learning (Attention U-Net)	ResNet-152	Ensemble	Accuracy	Yes	Unified focal loss	SGD
OXGastroVision (Ali and Tomar, 2021)	Polyp, Instrument	DDANet + FANet	None	Encoder–decoder	DSC	No	BCE and DSC loss	Adam
Vyobotics (Rauniyar et al., 2021)	Polyp	DDANet	None	Encoder–decoder	DSC and mIoU	Yes	BCE and DSC loss	Adam
NAAMII (Rauniyar et al., 2021)	Polyp, Instrument	U2Net	None	Encoder–decoder	mIoU	Yes	Mean Squared Error, Cross-entropy	Adam

embedded into the SINetV2 (Fan et al., 2021) framework. The SINetV2 framework with camouflaged object detection was used for better identification ability, as polyp segmentation is a downstream task. Additionally, extensive experiments have been conducted to study the effectiveness of DMF, and it was found that the method performs well under different data distributions, making it a favorable solution for problems where the training dataset had a different distribution of negative samples compared to the testing dataset.

Polypixel: Team Polypixel (Tzavara and Singstad, 2021) presented a study in which they used both pretrained and non-pretrained segmentation models for the polyp and instrument segmentation task. The team trained and validated both models on the dataset. The model architectures were retrieved from a Python library, "Segmentation Models" https://github.com/qubvel/segmentation_models, that contained different CNN architectures. This library offered models with both untrained and pre-trained weights, which were trained on the ImageNet dataset. To find the optimal fit for their datasets, they experimented and tested their results using EfficientNet, MobileNet, SE-ResNet, Inception, ResNet, and VGG. They achieved the best results with EfficientNetB1 for the polyp segmentation task.

agaldran: Team agaldran (Galdran, 2021) utilized a double encoder–decoder structure for polyp and instrument segmentation, which consists of two U-Net like structures arranged sequentially as shown in Fig. 5. The first encoder–decoder network processes the original image and produces output fed into the second encoder–decoder network. According to the authors, this setup allows the first network to highlight the important features of the image for segmentation, while the second network further improves the predictions of the first network. For the architectural design of a double encoder–decoder network, they incorporate Feature Pyramid Network (FPN) (Lin et al., 2017) architecture as the decoder mechanism, along with Resnext101 that serves as the pretrained decoder (Kolesnikov et al., 2020). This is done to optimize the feature extraction. To further refine the model's optimization process, they used Sharpness-Aware Minimization (SAM) along with the ADAM optimizer (Foret et al., 2020). The team employed a 4-fold cross-validation approach to train their models, training with four separate models and using temperature sharpening across the ensemble model to produce the final segmentation maps.

TeamAIKitchen: Team TeamAIKitchen (Keprate and Pandey, 2021) presented a methodology for developing, fine-tuning, and analyzing a U-Net-based model for generating segmentation masks for the polyp segmentation task. They modified the original U-Net architecture to extend it to work with less training samples and to generate the output mask of the same size as the input. ReLU activation function was used in the hidden layers. They further experimented with different batch sizes and selected 8 as the best. Same architecture was used for polyp and instrument segmentation with early stopping criteria.

CamAI: Team CamAI (Yeung, 2021) presented a deep learning pipeline that is specifically developed to accurately segment colorectal polyps and various instruments used during endoscopic procedures. To

Table 6
Performance comparison on Polyp segmentation task (Medico 2020). 'Bold' refers to the best score and 'red' color refers to the second best score. We follow this consistently in all the Tables. ↑ indicates a higher value is better.

Team name	mIoU ↑	DSC ↑	Recall ↑	Precision ↑	F2 ↑
PRML2020GU	**0.78975**	**0.86076**	0.90312	0.86731	0.87481
HBKU_UNITN_SIMULA	0.77736	0.84768	0.85034	0.88971	0.84483
AI-TCE	0.77733	0.85030	**0.91646**	0.83897	**0.87901**
HGV-HCMUS	0.76597	0.84050	0.89439	0.84455	0.85768
IIAI-Med	0.76195	0.83854	0.83049	**0.90121**	0.82837
SBS	0.75503	0.83162	0.83168	0.88513	0.82490
ML-MMIVSaruar	0.75168	0.82289	0.83908	0.88228	0.82492
AI-JMU	0.73742	0.81437	0.82661	0.87432	0.81038
MedSeg_JU	0.71330	0.80195	0.83542	0.82864	0.81240
VT	0.70578	0.79264	0.88353	0.78784	0.82368
NKT	0.68473	0.78012	0.80771	0.81264	0.78546
UNITRK	0.64379	0.72878	0.70989	0.85726	0.71312
GeorgeBatch	0.63511	0.73276	0.75003	0.82294	0.73615
AMI Lab	0.61958	0.70889	0.72865	0.79140	0.71226
IRIS-NSYSU	0.50353	0.64173	0.87915	0.58498	0.75089
UiO-Zero	0.43814	0.56185	0.69721	0.55587	0.61102
FAST-NU-DS	0.18344	0.26691	0.27447	0.29184	0.26762

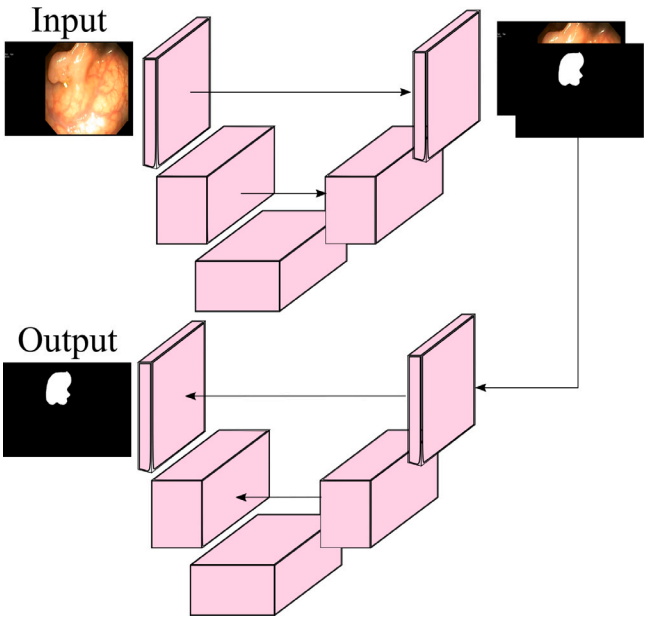


Fig. 5. Overview of winning solution of MedAI 2021 proposed by Team ***agaldran***. A double encoder–decoder network was used to segment polyps and surgical instruments.

improve transparency and interpretability, the pipeline leveraged the Attention U-Net architecture, which enables visualization of the attention coefficients to identify the most salient regions of the input images. This allowed for a better understanding of the model's decision-making process and facilitated the identification of potential errors. To further improve performance, the pipeline incorporated transfer learning using a pre-trained encoder. Additionally, test-time augmentation, softmax averaging, softmax thresholding and connected component labeling were used to further refine predictions and boost performance.

OXGastroVision: Team OXGastroVision (Ali and Tomar, 2021) presented a novel solution that utilizes two state-of-the-art deep learning models, namely the iterative FANet (Tomar et al., 2022) architecture and DDANet (Tomar et al., 2021). The FANet is based on a feedback attention network that allows rectifying predictions iteratively. It consists of four encoder and four decoder layers. Similarly, DDANet is based on a dual decoder attention network with one shared encoder at each layer. While the iterative mechanism in the full FANet architecture can lead to larger computational time, DDANet has real-time performance (70 FPS) but sub-optimal output. To overcome these limitations, the team proposes to use the segmentation maps from the DDANet output as input for the FANet iterative network for pruning. This approach aims to achieve a balance between computational efficiency and segmentation accuracy.

Vyobotics: Team Vyobotics (Rauniyar et al., 2021) presented a solution based on dual decoder attention network (DDANet) (Tomar et al., 2021), a deep learning model that has been specifically designed to achieve decent performance and real-time speed. The team performed data augmentation and trained a smaller network. This smaller network has a lower number of trainable parameters, which resulted in lower GPU training time. The ultimate goal of this approach was to achieve decent evaluation metrics while maintaining a decent FPS speed, which is crucial for real-time applications.

NAAMII: The team participated in polyp and instrument segmentation tasks. They employed U^2Net (Qin et al., 2020) as the base network. They added a separate learnable CNN network on the decoder part of the U2Net to regress the HoG features of the input images. The output from each decoder block was fed into the HoG regressor and learned the parameters to predict the HoG correctly. They jointly minimized Mean Squared Error (MSE) loss for HoG features and CrossEntropy loss for Segmentation. However, they only submitted their method description to the organizer and did not publish it as a research paper.

6. Results

In this section, we present a summary of the evaluated results obtained on the test dataset by all the participating teams in the two challenges: "Medico 2020" and "MedAI 2021". Each challenge consists of tasks with a specific focus and evaluation metrics. There were two tasks for the Medico 2020 challenge, namely *polyp segmentation* and *algorithm efficiency* tasks. In the MedAI 2021, there were three tasks, namely *polyp segmentation*, *endoscopic accessory instrument segmentation* and *transparency task*. The teams were evaluated based on standard evaluation metrics such as mIoU, DSC, Rec, Pre, Acc, F1, F2, and FPS. We emphasized mIoU, DSC, and FPS more, whereas we also acknowledge the importance of recall and precision as they are useful metrics in clinical settings. We have highlighted the best and the second-best scores in boldface and red color, respectively, for all the tasks in the two challenges.

6.1. Medico 2020 results

6.1.1. Polyp segmentation task

In Table 6, we provide the results for the *polyp segmentation* task. It can be observed that Team "PRML2020GU" outperforms other participating teams in the polyp segmentation task. It achieves a mIoU

Table 7
Algorithm efficiency task for polyp segmentation (Medico 2020). Note that some teams provided the same solution for this task as used in Task 1, whereas others designed different architecture specifically for the efficiency task (Task 2). ↑ indicates a higher value is better.

Team name	mIoU ↑	DSC ↑	Recall ↑	Precision ↑	F2 ↑	FPS ↑
GeorgeBatch	0.6351	0.7327	0.7500	0.8229	0.7361	**196.79**
UNITRK	0.6437	0.7287	0.7098	0.8572	0.7131	116.79
NKT	0.6847	0.7801	0.8077	0.8126	0.7854	80.60
HBKU_UNITN_SIMULA	**0.7364**	0.8074	0.8164	**0.8646**	0.8067	33.27
SBS	0.7341	**0.8148**	0.8764	0.8145	**0.8354**	26.66
AMI Lab	0.6195	0.7088	0.7286	0.7914	0.7122	107.87
FAST-NU-DS	0.6582	0.7556	**0.8982**	0.7171	0.8109	67.51
AI-JMU	0.7213	0.8017	0.8359	0.8495	0.8056	3.36
PRML2020GU	0.5083	0.6265	0.6003	0.7870	0.6029	2.25

of 0.7897, DSC of 0.8607, recall of 0.9031, precision of 0.8673, and F2 of 0.8748. Team "HBKU_UNITN_SIMULA" was the second best performing team with mIoU of 0.7773. Similarly, "AI-TCE" was the third best performing team with mIoU of 0.7773. The best-performing team, "PRML2020GU", used an encoder–decoder structure with EfficientNet as the backbone and a U-Net decoder with channel-spatial attention and deep supervision. This architecture had an improvement of 1.23% and 1.30% over the mIoU and DSC achieved by the Team "HBKU_UNITN_SIMULA", which used an average of three PraNet and five ResUNet++ trained on different training and validation datasets.

6.1.2. Algorithm efficiency task

To compute the ranking in the efficiency task, we used both mIoU and FPS, as the aim of the task was to develop lightweight models that are both accurate and fast. We calculated the average scores for normalized FPS and mIoU. Then, we calculated the difference between each team's normalized FPS score and mIoU with their corresponding average scores. Lastly, we added the two differences, and the team with the lowest difference was declared first, and in a similar way, other ranks were calculated. As in Table 7, team "PRML2020GU" has poor speed performance with a processing speed of only 2.25 fps, which is not desirable for a real-time efficient model. An interesting observation is that Team "GeorgeBatch" outperforms other participating teams in the algorithm efficiency task with a processing speed of 196.79 fps, as seen from Table 7. However, it is worth noting that the team obtained a low mIoU of 0.6351 for the polyp segmentation task, even though we are considering it as the winner in this task. Team "UNITRK" obtained a second-best fps of 116.79 and a decent mIoU of 0.6437. Similarly, team "NKT" obtained a balanced mIoU of 0.6847 and a high speed of 80.60 fps, and was ranked third for this task. Despite the two teams, "UNITRK" and "GeorgeBatch", achieving the highest evaluation fps values, there is a trade-off between speed and mIoU. Low FPS cannot be used for real-time medical processing applications, and low overlap evaluation metrics cannot generate precise segmentation masks. To provide further insight, we have included the qualitative results of all the teams participating in the Medico 2020 challenge in Fig. 6. We can see that none of the teams came close to the ground truth mask. Achieving a balance between these metrics is crucial for developing an efficient polyp segmentation model.

6.2. MedAI 2021 challenge results

6.2.1. Polyp segmentation task

In Table 8, we tabulated the evaluation results of all the participating teams in MedAI 2021 for polyp segmentation task. It can be observed that team "agaldran" outperforms other teams in the polyp segmentation task with mIoU of 0.8522, and DSC of 0.8965. Team "CV&Med IIAI" also showed good performance and was ranked 2nd in the polyp segmentation task with a mIoU of 0.8484, a very small difference from the best-performing team. In Fig. 7, we present the qualitative results of the participating teams for the polyp segmentation task of MedAI 2021. None of the methods performed well on this challenging image, emphasizing the need for more robust polyp segmentation methods. However, in the overall test set, the predicted segmentation masks from most of the team performed well on regular polyps (see Supplementary materials Figure). Overall, the qualitative masks produced by teams "agaldran" and "CV&Med IIAI" were better than the other teams.

Table 8
Performance evaluation for the participating teams for the polyp segmentation task in MedAI 2021 Challenge. ↑ indicates a higher value is better..

Team name	mIoU ↑	DSC ↑	Recall ↑	Precision ↑
agaldran	**0.8522**	0.8965	0.9009	0.9242
CV&Med IIAI	0.8484	**0.8993**	0.9186	0.9100
NYCity	0.8418	0.8885	0.8794	**0.9319**
IIAI-CV&Med	0.8361	0.8927	**0.9195**	0.8963
mTEC	0.8334	0.8892	0.9010	0.9096
PRML	0.8116	0.8669	0.8852	0.8922
CamAI	0.8083	0.8701	0.8702	0.9052
The Arctic	0.8022	0.8533	0.8604	0.8821
Polypixel	0.7997	0.8567	0.8868	0.8659
MAHUNM	0.7495	0.8189	0.8397	0.8568
OXGastroVision	0.7334	0.7966	0.8158	0.8374
Vyobotics	0.7220	0.7967	0.8214	0.8359
NAAMII	0.6041	0.6940	0.7499	0.7334
leen	0.4595	0.5531	0.6389	0.5860
The Segmentors	0.3789	0.4205	0.4178	0.4640
TeamAIKitchen	0.2904	0.4100	0.7152	0.4910

Table 9
Performance of participating teams for instrument segmentation task of MedAI 2021 Challenge. ↑ indicates a higher value is better..

Team name	mIoU ↑	DSC ↑	Recall ↑	Precision ↑
agaldran	**0.9364**	**0.9635**	0.9692	**0.9632**
NYCity	0.9326	0.9586	**0.9712**	0.9516
mTEC	0.9245	0.9553	0.9687	0.9490
PRML	0.9178	0.9528	0.9687	0.9441
IIAI-CV&Med	0.9148	0.9490	0.9612	0.9473
CV&Med IIAI	0.9136	0.9512	0.9605	0.9500
Polypixel	0.9114	0.9478	0.9591	0.9438
CamAI	0.9085	0.9437	0.9454	0.9514
The Arctic	0.9078	0.9448	0.9735	0.9231
OXGastroVision	0.8692	0.9073	0.9236	0.9096
MAHUNM	0.8523	0.9080	0.9535	0.8864
MedSeg_JU	0.8205	0.8632	0.9005	0.8464
TeamAIKitchen	0.7257	0.7980	0.7955	0.8510
leen	0.6991	0.7845	0.7963	0.8232
NAAMII	0.6857	0.7741	0.8321	0.7669
The Segmentors	0.3668	0.3971	0.3985	0.4040

6.2.2. Instrument segmentation task

From Table 9, it can be observed that the same team, "agaldran" also outperforms other participating teams in the instrument segmentation task with a high mIoU of 0.9364 and DSC of 0.9635. Team "NYCity" was ranked 2nd in this task with a mIoU of 0.9326 and DSC of 0.9586. However, Team "NYCity" obtained the highest recall of 0.9712, which signifies it has low false negative (FN) regions in the

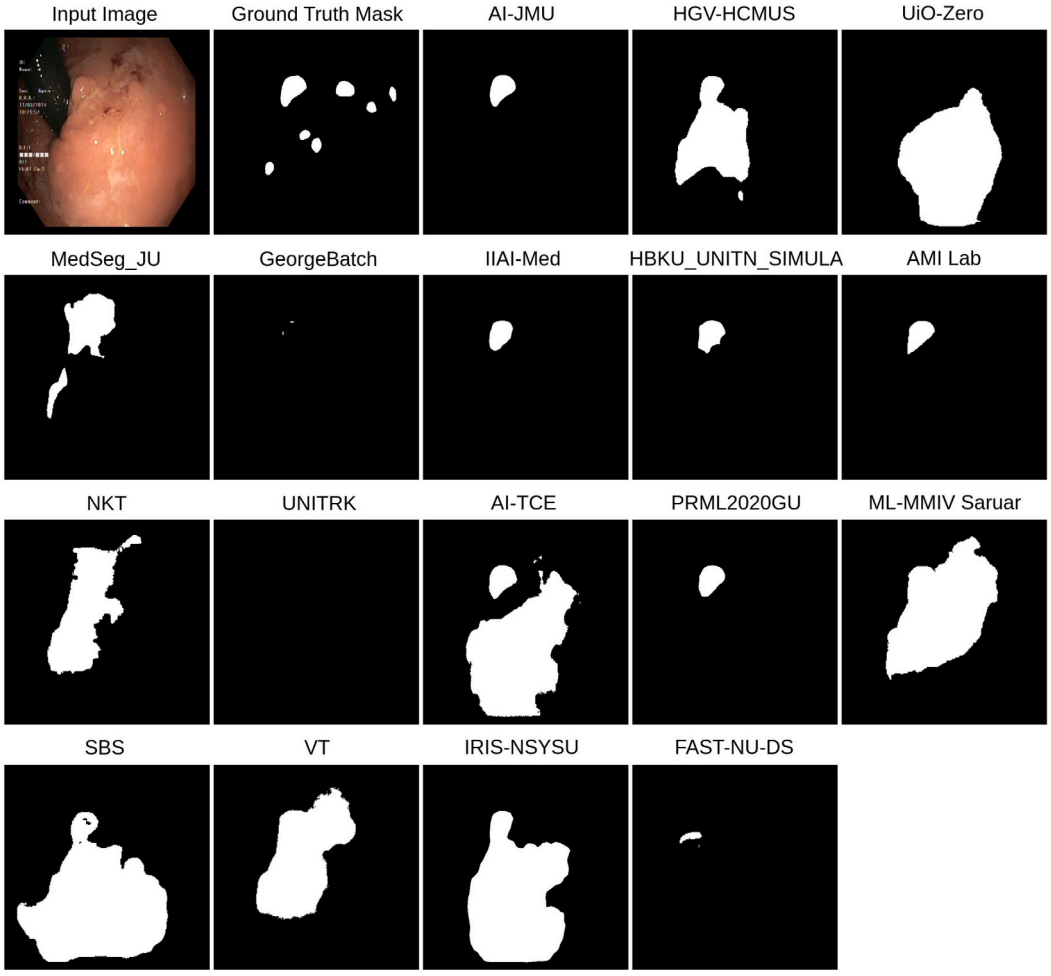


Fig. 6. The figure shows the qualitative results of participating teams for the polyp segmentation task in the Medico 2020 Challenge on challenging scenarios. When each team's predicted mask is compared with its corresponding ground truth, we observe that none of the teams obtained results that fit well with the ground truth.

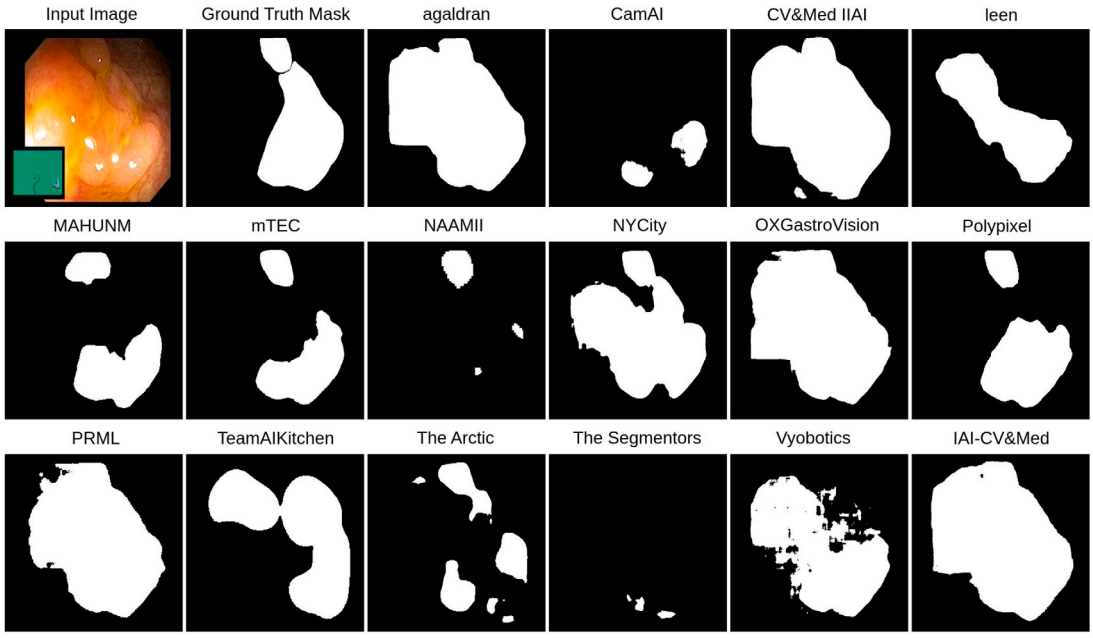


Fig. 7. Qualitative results of all the methods participating in polyps segmentation challenge in MedAI2021.

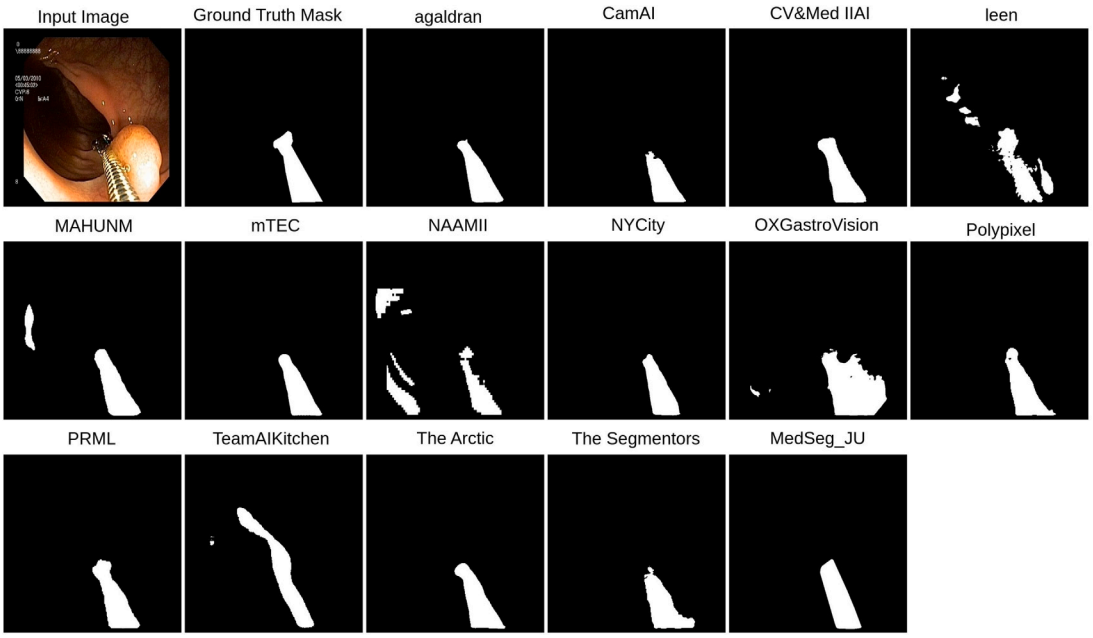


Fig. 8. Qualitative results of all the methods participating in surgical instrument segmentation challenge in MedAI2021.

Table 10
Evaluation of the 'Transparency tasks' for MedAI 2021 Challenge. For this task, a team of experts accessed the submission based on several criteria and provided a score based on the availability and quality of the source code (for e.g., open access, public availability, and documentation for reproducibility), model evaluation (for e.g., failure analysis, ablation study, explainability, and metrics used) and qualitative evaluation from clinical experts (e.g., usefulness and understandability of the results). Here, '0' refers to no submissions for the transparency task. Doctor evaluation was only calculated for the team which manuscript were accepted.

Team name	Open source			Model evaluation				Doctor evaluation		Final Score
	Publicly available (0 or 1)	Code quality (0-3)	Readme (0-3)	Failure Analysis (0-3)	Ablation Study (0-3)	Explainability (0-3)	Metrics Used (0 or 1)	Usefulness (0-3)	Understandable (0-5)	
agaldran	1	2	3	3	3	3	1	2	3	21
mTEC	1	1	3	3	1	0	1	3	4	17
CamAI	1	1	1	2	1	2	1	2	5	16
The Arctic	1	2	1	1	0	3	1	1	3	13
IIAI-CV&Med	1	1	2	0	0	0	1	1	4	10
Polypixel	1	1	2	0	0	0	1	0	0	5
leen	0	1	0	0	0	2	1	0	0	4
MAHUNM	1	1	0	0	0	0	1	0	0	3
OXGastroVision	0	2	0	0	0	0	1	0	0	3
CV&Med IIAI	0	1	0	1	0	0	1	0	0	3
PRML	0	1	0	0	0	0	1	0	0	2
TeamAIKitchen	0	1	0	0	0	0	1	0	0	2
The Segmentors	0	0	0	0	0	0	1	0	0	1
NYCity	0	0	0	0	0	0	1	0	0	1

predicted segmentation mask compared to team "agaldran". Another interesting observation is the team "agaldran" also achieved higher metric values for the instrument segmentation task as compared to the polyp segmentation task, as instrument segmentation is relatively easier than polyp extraction due to the greater variability of the latter regarding color and appearance. In Fig. 8, we also present the qualitative results of the research teams who participated in the instrument segmentation challenge of MedAI2021. From the qualitative results, it can be observed that the ground truth prediction made by team "agaldran" is also superior to the other team. Therefore, it can be concluded from the obtained evaluation metrics for the two tasks that team "agaldran" proposed a more robust algorithm and was accurately able to segment polyp and instrument at high accuracy.

6.2.3. *Transparency task*

We present the transparency results in Table 10. Team "agaldran" outperformed other competitors with a final score of 21 out of 25. Similarly, "mTEC" obtained a score of 17 out of 25 and was ranked 2nd. Likewise, team "CamAI" obtained a score of 16 out of 25 and was ranked third in the transparency task. There were also efforts from teams such as "The Arctic", which obtained a score of 13, and "IIAI-CV&Med", which obtained a score of 10. These scores show their effort to provide a transparent solution to the polyp and instrument segmentation tasks. We provide the final ranking and task-wise scores in Fig. 9. Notably, team ***"agaldran"*** outperformed others in all three tasks and overall challenge and emerged as the winner of the MedAI 2021 challenge. Overall, ***"mTec"*** secured the second position. Following closely behind, ***"CamAI"*** showcased the third-best solution. The overall rank was computed by combining the mIoU scores of polyp and instrument segmentation tasks and the Transparency score.

Fig. 10(a) illustrates the plot of mIoU reported by each team in their submissions in the two challenges with three different tasks. It can be observed that the *polyp segmentation task* from 2020 to 2021 gained improvement with a larger number of submissions achieving a mIoU of more than 0.80 and the best-performing team with a mIoU of around 0.85. Similar progress can be observed in Fig. 10(b) where an overall mIoU increased by 4.93% when an average score is computed over all participating teams' individual best mIoU in the 2021 polyp

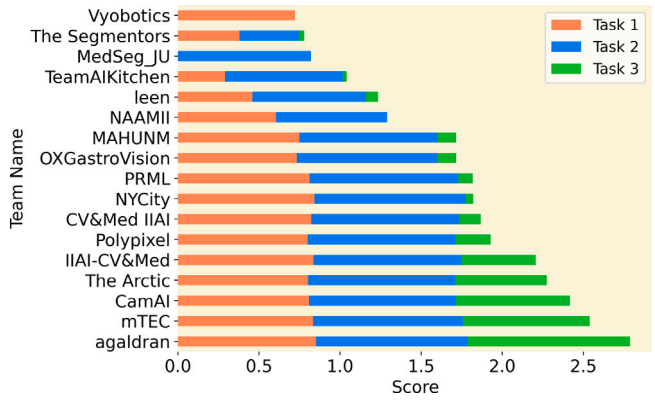


Fig. 9. Task-wise scores achieved by participating teams of MedAI 2021 challenge. Team rankings are decided on the basis of overall scores in all three tasks. Here, we plot the mIoU of Task1 and Task 2, and we have normalized the transparency score to calculate the overall score.

segmentation challenge. We further compared all segmentation metrics, including DSC, recall, precision, mIoU score, accuracy, and F2 score, as shown in Fig. 10(c). Notably, the different evaluation metrics scores are consistent with instrument segmentation tasks in the MedAI challenge. However, there is a high variation in the mIoU between the different teams in the polyp segmentation tasks of Medico 2020 and MedAI 2021 challenges.

These values pertain to the best score corresponding to a particular metric the individual team obtained in different executions. It is to be noted that each team was given the opportunity to submit five different submissions, and the best results for the best submission are reported in the Tables here. From here, it can be observed that most teams in the MedAI 2021 challenge reported overall high scores in terms of various segmentation metrics when compared to Medico 2020 outcomes, thus highlighting the improved performance trends in automated systems over time. Furthermore, it can also be visualized that unlike the high variations shown by teams' scores in the polyp segmentation task, better performance and smaller deviations in scores are reported in the instrument segmentation task. The high variations in the polyp segmentation results also show that polyp segmentation is more challenging because of the presence of variations in the size, structure and appearance of the polyps, and the presence of the artifacts and lighting conditions deteriorate the algorithm's performance.

7. Discussions

The rapid advancement in the AI-based techniques that support CADe and CADx systems has resulted in the introduction of numerous algorithms in the domain of medical image analysis, including colonoscopy. To assess the performance of these algorithms, it is important to benchmark on the particular set of datasets. It enables the comparison and analysis of different techniques and assists in identifying challenging cases that need to be targeted using improved methodologies. This also includes cases that are misled by the presence of artifacts and occlusion by surgical instruments (Ali et al., 2020b). Besides developing and analyzing AI-based algorithms, it is crucial to include explainability and interpretability to infuse trust and reliance when adopting automated systems in clinical settings. Therefore, the challenges discussed in this paper not only focus on lesion and instrument segmentation but also emphasize the importance of transparency in medical image analysis. This section covers the findings, limitations, analysis of failing cases, trust, safety, and interpretability of the methods, and future steps and strategies for Medico 2020 and MedAI 2021.

7.1. Medico 2020 challenge methods

Most of the methods reported in the Medico 2020 challenge focus on encoder–decoder architecture (for example, U-Net, ResUNet++, PraNet, Efficient UNet, etc.). Other networks used include conditional GAN and Faster R-CNN. The overview of the methods is provided in Table 3. For more detailed architectural information, we have also included the backbone and algorithm used by each team. Further, we also report the nature of the algorithm and the choice basis of evaluation, such as mIoU, DSC or FPS. Additionally, we provide information about the augmentation and hyperparameters, such as loss function and optimizers. It is noteworthy that all the top three teams "PRML2020GU", "HBKU_UNITN_SIMULA" and "AI-TCE" used the encoder–decoder architecture. Out of 17 participating teams, only three teams adopted some other architectures. Comparative analysis shows that the highest-scoring encoder–decoder network outperforms the GAN-based approach by a significant margin of 0.3517 in mIoU and 0.2989 in DSC score. Similarly, compared to the R-CNN-inspired networks (team "IRIS-NSYSU"), the best approach (team "PRML2020GU") achieves an improvement of 0.2863 in mIoU and 0.2191 DSC. Medico 2020 challenges provide valuable insight and trends for the polyp segmentation and biomedical image analysis challenges. Most deep learning frameworks submitted for the challenge used the Adam optimizer to optimize their network. However, a handful of teams used other optimizers, such as SGD or RMSProp. Additionally, most of the teams used data augmentation to boost the number of training samples prior to training their frameworks to improve the performance of their architecture. There have been different preferences in loss function where most of the team used "BCE + DSC loss", "binary cross-entropy", IoU loss, etc. However, from the results of the top three teams, it can be concluded that "BCE + DSC loss" is best for this dataset. Similarly, in terms of the backbone for the model architecture, the EfficientNet variant (selected by PRML2020GU) or EfficientNetB4 (selected by AI-TCE) were most favorable.

7.2. MedAI 2021 challenge methods

The summary of the different approaches adopted by the participating teams of the MedAI2021 Challenge is presented in Table 5. To provide a brief overview of the general techniques adopted by the different teams, they can be categorized based on the nature of the approach followed, such as ensemble models, encoder–decoder based architectures, CNN, and hybrid CNN models. Almost all the teams presented the same model for both the tasks proposed in the challenge. Most teams explored ensemble modeling, encoder–decoder networks, or a combination of both in the polyp segmentation task. Another criterion of categorization could be CNN or transformed-based approaches. It is observed that the top-ranked team "agaldran" utilized two encoder–decoder networks and reported a mIoU score of 0.8522. Similarly, "CV&Med IIAI" was ranked second, and Team "NYCity" was ranked third in the polyp segmentation task with a competitive mIoU value of 0.8484 and 0.8418, respectively. Similar to the Medico 2020 polyp segmentation challenge, where GAN-based methods were adopted by teams (for example, Team "leen") failed to perform well in this challenge for polyp and instrument segmentation tasks. It is to be noted that the winning team, "agaldran" used a double encoder–decoder structure with two U-Net, where they incorporated FPN and Resnext101 as the pretrained decoder. They also use SAM and Adam optimizer to optimize the model further. The other competitive team "CV&Med IIAI" used the SINetv2 algorithm with PVTv2 as the backbone, and NYCity used the combination of HarDNet-85 ResNet101.

In the MedAI2021 instrument challenge, participants mainly focused on either ensemble models or encoder–decoder networks similar to the polyp segmentation task. As the majority of the teams utilized the same model that they proposed for the polyp segmentation problem in this task, the categorization of overall methods remains the same

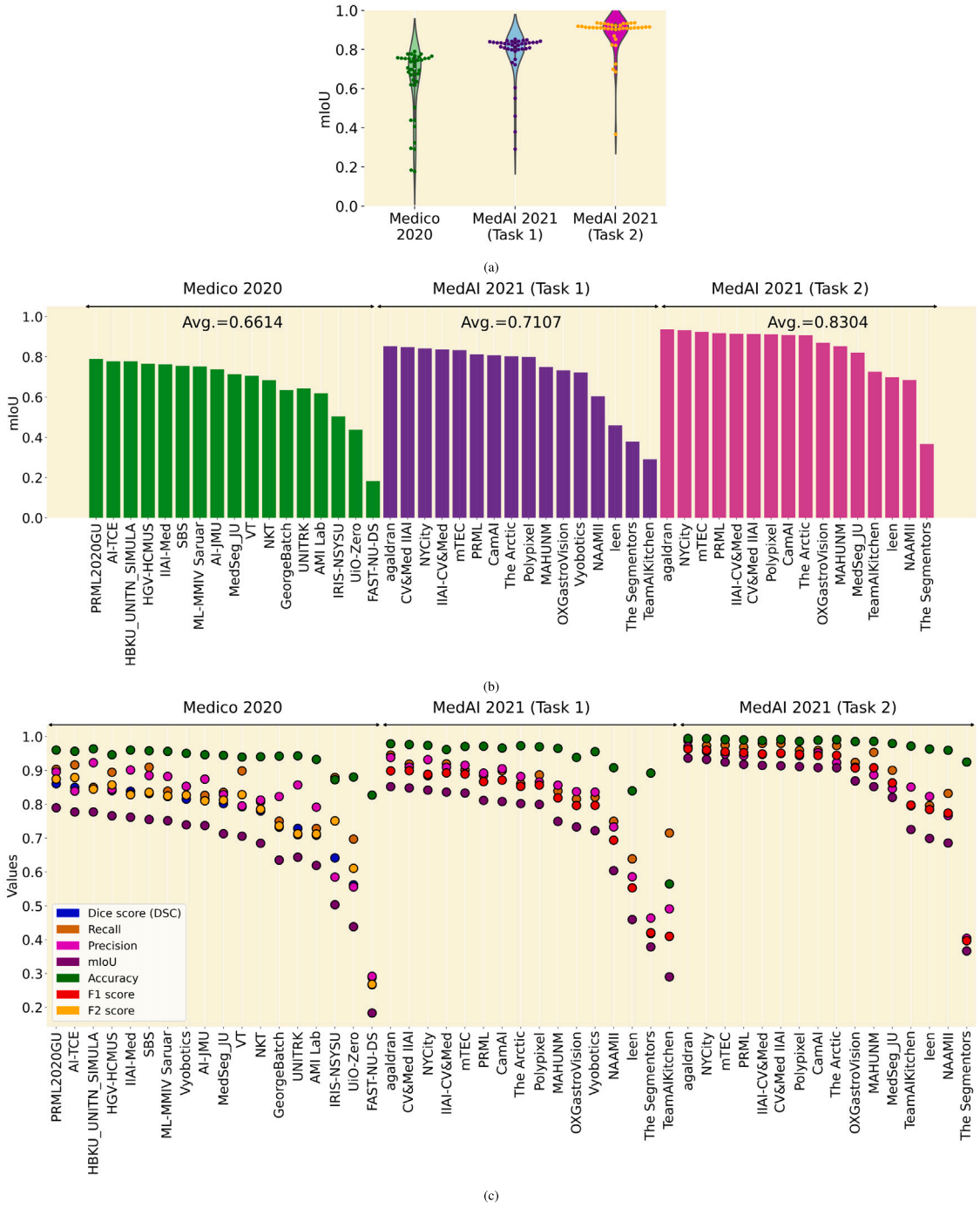


Fig. 10. (a) Violin plots with overlaid swarm plots depicting statistics of submissions received for different tasks for the two challenges, (b) mIoU score comparison of different teams in three tasks of Medico 2020 (polyp segmentation) and MedAI 2021 (Task 1: polyp segmentation and Task 2: instrument segmentation), and (c) Strip plots for all segmentation metrics (Dice score, recall, precision, mIoU score, accuracy, F1 score, and F2 score) reported by different teams in both challenges for all test data samples.

as that of the first task described above. The top rank is secured by Team "agaldran", with encoder–decoder architecture, pyramid network as the decoder, and Resnext101 as the pre-trained decoder. The second-ranked model by Team "NYCity" is the CNN and transformer based ensemble model, which achieved only a slight difference in the scores from the leading model. mTec was ranked third in the challenge, which used dual parallel reverse attention edge network (DPRA-EdgeNet) (Bhattacharya et al., 2021a). The architecture used HardNet (Chao et al., 2019) as the backbone. The challenge shows that most of the teams were reluctant to share their method (refer to Table 10). From the table, it can be seen that only five teams were qualified for the doctor evaluation. Additionally, the quality of the code submitted by most of the team was not satisfactory. Most of the participants did not put much effort into the readme file. Additionally, most teams neglected the failure analysis, ablation study and explainability in their submission. Moreover, based on the doctor's evaluation, only the solution provided by a few teams (for example, "agaldran", "mTEC" "CamAI", "The Arctic", and "IIAI-CV&Med") was considered useful and understandable.

7.3. Analysis of the failed cases

We have analyzed the regular and failing cases in polyp and surgical tool segmentation to highlight the limitations of the current methods

so that these cases can be considered during further algorithm development. Fig. 6 and Fig. 7 show examples of instances where the models fail for most cases. From the results on the test dataset, it was observed that most of the algorithms failed on diminutive and flat polyps located in the left colon. These are the challenging classes in the colon and require effective detection and diagnosis systems. Similarly, although most of the methods performed well on the diagnostic and therapeutic surgical tool, there were issues with the images having caps and forceps. Additionally, the algorithms showed difficulties with challenging cases for polyps on rare cases such as sessile polyps, even though it performed well on overall quantitative metrics (see Fig. 6, 7, 8 and Figures in the supplementary material). Therefore, investigating the cause for misclassification for such polyp and instrument samples in the dataset, along with the failure analysis, will be critical to focus on for future research. This can include evaluating generalization performance on unseen test data from different hospitals. Such investigations can reduce the chances of underperformance on rare cases.

7.4. Trust, safety, and interpretability of methods

Integrating CADe or CADx in clinical settings necessitates addressing factors such as trust, safety, and interpretability to ensure its adoption. The high variations and potential bias in the curated datasets used to train such models and the actual scenarios in which they are adopted create a high chance of biases, impacting the generalizability of the method. Such bias ultimately makes it challenging to infuse trust while adopting CADe or CADx tools and questions the safety of patients. To tackle this issue, we introduced a transparency task in the MedAI2021 challenge that underscores the need for interpretability, reproducibility, and explainability in medical AI research, including polyp and instrument segmentation.

Our initiative aimed to light the potential risk that can arise from wrong decisions based on model and algorithmic bias. Our dataset contained polyp cases with varied appearances in terms of shapes, sizes, the presence of artifacts, lightning conditions, textures, and the different numbers of polyps per image that are encountered in real-world clinical settings. Additionally, we have included frames containing surgical instruments to support the cases of occluded endoluminal elements or polyps that could arise in general. Some of the methods adopted by the participating teams include the submission of intermediate heatmaps using approaches like layer-wise relevance propagation that showed visual explanation and highlighted the model decision-making process. Team ‘‘agaldran’’ provided detailed ablation studies in support of the predictions obtained. By promoting transparency through subjective analysis and addressing potential biases, the MedAI challenge aimed to foster trust in the presented solution and ensure safety in adopting such methods in the clinic.

7.5. Limitation of the Medico 2020 and MedAI 2021

In our study, we aimed to standardize the challenge of polyp and instrument segmentation by providing the same test sets and evaluation metrics to all participants. To achieve this, we introduced variable polyp cases, including polyps with different sizes, noisy frames with artifacts, blurry images, and occlusion. We also added regular frames to the test set to ensure that participants drew the ground truth manually and did not cheat. However, our study has some limitations. Although we used datasets collected from four medical centers in Norway, these images are from a single country, limiting the ethnicity variance though there is very limited differences if any in the mucosal appearance between ethnicities. Nevertheless, there is a need for a more diverse dataset that includes multiple ethnicities and countries also because the prevalence of various diseases varies between regions. Moreover, the current models should be tested on multi-center datasets to assess their generalization ability.

There was no online leaderboard in our challenge due to the Mediaeval policy. Therefore, we manually calculated the predictions submitted by each team. Each team had limitations of 5 submissions for each task, which restricted further optimization opportunities. Although we have also introduced normal findings from the GI tract to trick the participants and models, our challenge only used still frames and did not incorporate video sequence datasets. Even when the best performing algorithms are tested on a temporal video sequence dataset, it is possible that the performance can drop. Most of the images are only from white light imaging. Although our dataset was annotated by one annotator and checked by two gastroenterologists, there is still a possibility of bias in the labels. In the accessory instrument challenge, we had more images from the stomach class than accessory instruments such as biopsy forceps or snares due to the lack of availability of datasets. Finally, despite including diverse cases in the polyp and instrument segmentation challenge, we still had limited flat and sessile polyps, frequently missed during routine colonoscopy examinations. Incorporating multi-center data and video sequences data and addressing label biases will lead to more comprehensive and reliable evaluations of AI-based colonoscopy systems.

7.6. Future steps and strategies

In our study, we aimed to promote transparency and interpretability in machine learning models for the GI tract setting. However, more work is needed to understand how decisions are made and identify potential biases or errors in a quantitative manner to build trust in such systems in a clinical setting. To achieve this, we plan to test the best-performing algorithms on large-scale datasets to observe their scalability. We will consider using more quantitative metrics, such as statistical mixed models, bootstrapping analysis and estimate confidence intervals. Additionally, we will also include metrics such as Hausdorff distance and normalized surface distance.

We will emphasize more transparent decision-making methods and visualize interpretability results while focusing on clinical relevance rated by expert clinicians instead of just one objective metric. To achieve this, we have already started collecting large-scale datasets and plan to build a tool if the algorithms are robust enough and verified by our gastroenterologists. Next, we will propose a challenge to polyp video sequences analysis. We will explore the integration of state space models, such as Video Vision Mamba-based framework (Yang et al., 2024), to capture the temporal information in video sequences that affect the efficiency and accuracy of segmentation tasks. It is worth noting that there has been innovation within hardware (colonoscope) for safer medical colonoscopy devices, such as developing fully flexible automated colonoscopes to offer expanded fields of view rather than 120-170° visualization, which can capture dead spots, improving the lesions’ miss-rate. These scopes are currently in the final stage of development. This hardware would require high processing speed to locate potential lesions in real time for a smooth workflow. We believe these solutions from our challenge could help address the complexities with the improved hardware and improved image quality.

8. Conclusion

Our study aimed to provide a comprehensive analysis of the methods used by participants in the Medico 2020 and MedAI 2021 competitions for different medical image analysis tasks. We designed the tasks and datasets to demonstrate that the best-performing approaches were relatively robust and efficient for automatic polyp and instrument segmentation. We evaluated the challenge based on several standard metrics. In MedAI 2021, we also used a quantitative approach, where a multi-disciplinary team, including gastroenterologists, accessed each submission and evaluated the usefulness and understandability of their results. Through the qualitative results, we found that even the best-performing method underperforms in rare cases. This highlights the

need for further investigation to understand the cause of misclassification. During the "performance task" and "algorithm efficiency" tasks, we observed a trade-off between mIoU and inference time when tested across unseen still frames. For the instrument segmentation challenge, we observed that almost all teams performed relatively well, as segmenting instruments is easier than polyp segmentation. From the transparency task, we observed that more effort is required from the community to enhance the transparency of the proposed model. Overall, we also observed that several teams demonstrated the use of data augmentation and optimization techniques to improve performance on specific tasks. Our study highlights the need for multi-center dataset collection from larger and more diverse populations, including experts from various clinics worldwide. More competitions should be held on polyp video sequences to observe the efficiency difference in still frames and video sequences. Further research should investigate multiple polyp classes that typically fail in clinical settings, multi-center clinical trials, and the emphasis on real-time systems. Additionally, research on transparency and interpretability should be emphasized as it could help build clinically relevant and trustworthy systems.

CRediT authorship contribution statement

Debesh Jha: Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Vanshali Sharma:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation. **Debapriya Banik:** Writing – original draft. **Debayan Bhattacharya:** Writing – original draft. **Kaushiki Roy:** Writing – original draft. **Steven A. Hicks:** Formal analysis. **Nikhil Kumar Tomar:** Writing – review & editing, Writing – original draft, Methodology. **Vajira Thambawita:** Methodology. **Adrian Krenzer:** Methodology. **Ge-Peng Ji:** Methodology. **Sahadev Poudel:** Methodology. **George Batchkala:** Methodology. **Saruar Alam:** Methodology. **Awadelrahman M.A. Ahmed:** Methodology. **Quoc-Huy Trinh:** Methodology. **Zeshan Khan:** Methodology. **Tien-Phat Nguyen:** Methodology. **Shruti Shrestha:** Methodology. **Sabari Nathan:** Supervision. **Jeonghwan Gwak:** Methodology. **Ritika K. Jha:** Writing – original draft, Validation, Methodology. **Zheyuan Zhang:** Writing – original draft. **Alexander Schlaefer:** Methodology. **Debotosh Bhattacharjee:** Methodology. **M.K. Bhuyan:** Supervision. **Pradip K. Das:** Supervision. **Deng-Ping Fan:** Supervision. **Sravanthi Parasa:** Data curation. **Sharib Ali:** Writing – original draft. **Michael A. Riegler:** Supervision, Software, Resources, Project administration, Methodology, Investigation, Data curation, Conceptualization. **Pål Halvorsen:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Resources, Project administration, Methodology, Conceptualization. **Thomas de Lange:** Writing – review & editing, Writing – original draft, Data curation. **Ulas Bagci:** Writing – review & editing, Writing – original draft, Supervision.

Declaration of competing interest

1. Financial Interests:

Author have no financial interests, direct or indirect, in the research or its outcomes presented in the manuscript.

2. Non-Financial Interests:

Author have no non-financial interests that could be perceived as having influenced the research or its presentation in the manuscript.

3. Conflicts of Interest:

Author confirm that there are no known conflicts of interest that could potentially bias the results, analysis, or conclusions presented in the manuscript.

Data availability

Data will be made available on request.

Acknowledgments

D. Jha is supported by NIH (United States) funding: R01-CA246704, R01-CA240639, U01-DK127384-02S1, and U01-CA268808. V. Sharma is supported by the INSPIRE fellowship (IF190362), DST, Govt. of India. D. Bhattacharya is partially funded by the i^3 initiative of the Hamburg University of Technology and by the Free and Hanseatic City of Hamburg (Interdisciplinary Graduate School "Innovative Technologies in Cancer Diagnostics and Therapy"). K. Roy is thankful to DST Inspire Ph.D fellowship, India (IF170366).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2024.103307.

References

Ahmed, A., Ali, M., 2020. Generative adversarial networks for automatic polyp segmentation. arXiv preprint arXiv:2012.06771.

Ahmed, A., Ali, L.A., 2021. Explainable medical image segmentation via generative adversarial networks and layer-wise relevance propagation. arXiv preprint arXiv:2111.01665.

Alam, S., Tomar, N.K., Thakur, A., Jha, D., Rauniyar, A., 2020. Automatic polyp segmentation using u-net-resnet50. arXiv preprint arXiv:2012.15247.

Ali, S., Dmitrieva, M., Ghatwary, N., Bano, S., Polat, G., Temizel, A., Krenzer, A., Hekalo, A., Guo, Y.B., Matuszewski, B., et al., 2021. Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. Med. Imag. Anal. 70, 102002.

Ali, S., Ghatwary, N., Jha, D., Isik-Polat, E., Polat, G., Yang, C., Li, W., Galdran, A., Ballester, M.Á.G., Thambawita, V., et al., 2022a. Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge. arXiv preprint arXiv:2202.12031.

Ali, S., Jha, D., Ghatwary, N., Realdon, S., Cannizzaro, R., Salem, O.E., Lamarque, D., Daul, C., Anonsen, K.V., Riegler, M.A., et al., 2022b. A multi-centre polyp detection and segmentation dataset for generalisability assessment. Sci. Data.

Ali, S.M.F., Khan, M.T., Haider, S.U., Ahmed, T., Khan, Z., Tahir, M.A., 2020a. Depth-wise separable atrous convolution for polyps segmentation in gastro-intestinal tract. In: Proceedings of the MediaEval. pp. 1–3.

Ali, S., Tomar, N.K., 2021. Iterative deep learning for improved segmentation of endoscopic images. Nord. Mach. Intell. 1 (1), 38–40.

Ali, S., et al., 2020b. An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. Sci. Rep. 1–21.

Angermann, Q., Bernal, J., Sánchez-Montes, C., Hammami, M., Fernández-Esparrach, G., Dray, X., Romain, O., Sánchez, F.J., Histace, A., 2017. Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis. In: Proceedings of the Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures: 4th International Workshop, CARE 2017, and 6th International Workshop, CLIP 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, 2017, Proceedings 4. pp. 29–41.

Asplund, J., Kauppila, J.H., Mattsson, F., Lagergren, J., 2018. Survival trends in gastric adenocarcinoma: a population-based study in Sweden. Ann. Surgi. Oncol. 25 (9), 2693–2702.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One 10 (7), e0130140.

Ballesteros, C., Trujillo, M., Mazo, C., Chaves, D., Hoyos, J., 2017. Automatic classification of non-informative frames in colonoscopy videos using texture analysis. In: Proceedings of the Lecture Notes in Computer Science, Vol. 10125. pp. 401–408.

Banik, D., Bhattacharjee, D., 2020. Deep conditional adversarial learning for polyp segmentation. In: Proceedings of the MediaEval. pp. 1–3.

Banik, D., Roy, K., Bhattacharjee, D., 2021. EM-Net: An efficient M-Net for segmentation of surgical instruments in colonoscopy frames. Nord. Mach. Intell. 1 (1), 14–16.

Batchkala, G., Ali, S., 2020. Real-time polyp segmentation using U-net with IoU loss. In: Proceedings of the MediaEval. pp. 1–3.

Bernal, J., Aymeric, H., 2017. Gastrointestinal Image ANAlysis (GIANA) Angiodysplasia D&L challenge. https://endovissub2017-giana.grand-challenge.org/home/. (Accessed 20 November 2017).

Bernal, J., Sánchez, J., Vilarino, F., 2012. Towards automatic polyp detection with a polyp appearance model. Pattern Recognit. 45 (9), 3166–3182.

Bernal, J., Tajkbaksh, N., Sanchez, F.J., Matuszewski, B.J., Chen, H., Yu, L., Angermann, Q., Romain, O., Rustad, B., Balasingham, I., et al., 2017. Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 endoscopic vision challenge. IEEE Trans. Med. Imaging 36 (6), 1231–1249.

Bernal, J., et al., 2018. Polyp detection benchmark in colonoscopy videos using gtcreator: A novel fully configurable tool for easy and fast annotation of image databases. In: Proceedings of the Comput. Assist. Radiol. Surg. CARS.

Bhattacharya, D., Betz, C., Eggert, D., Schlaefer, A., 2021a. Dual parallel reverse attention edge network : DPRA-EdgeNet. Nord. Mach. Intell. 1 (1), 8–10.

Bhattacharya, D., Betz, C., Eggert, D., Schlaefer, A., 2021b. Self-supervised U-net for segmenting flat and sessile polyps. arXiv preprint arXiv:2110.08776.

Cai, Z., Vasconcelos, N., 2019. Cascade R-CNN: High quality object detection and instance segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 43 (5), 1483–1498.

Chao, P., Kao, C.Y., Ruan, Y.S., Huang, C.H., Lin, Y.L., 2019. Hardnet: A low memory traffic network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3552–3561.

Chen, Y.H., Kuo, P.H., Fang, Y.Z., Wang, W.L., 2021. More birds in the hand -medical image segmentation using a multi-model ensemble framework. Nord. Mach. Intell. 1 (1), 23–25.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 801–818.

Chou, Y., 2021. Automatic polyp and instrument segmentation in MedAI-2021. Nord. Mach. Intell. 1 (1), 17–19.

Dong, B., Wang, W., Fan, D.P., Li, J., Fu, H., Shao, L., 2021a. Polyp-pvt: Polyp segmentation with pyramid vision transformers. arXiv preprint arXiv:2108.06932.

Dong, B., Wang, W., Li, J., 2021b. Transformer based multi-model fusion for medical image segmentation. Nord. Mach. Intell. 1 (1), 50–52.

Fan, D.P., Ji, G.P., Cheng, M.M., Shao, L., 2021. Concealed object detection. IEEE Trans. Pattern Anal. Mach. Intell. 44 (10), 6024–6042.

Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L., 2020. Pranet: Parallel reverse attention network for polyp segmentation. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23. pp. 263–273.

Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B., 2020. Sharpness-aware minimization for efficiently improving generalization. arXiv preprint arXiv:2010.01412.

Galdran, A., 2021. Polyp and surgical instrument segmentation with double encoder-decoder networks. Nord. Mach. Intell. 1 (1), 5–7.

Haithami, M., Ahmed, A., Liao, I.Y., Jalab, H., 2021. Employing GRU to combine feature maps in DeeplabV3 for a better segmentation model. Nord. Mach. Intell. 1 (1), 29–31.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conf. Comput. Vis. Pattern Recognit. CVPR, pp. 770–778.

Hicks, S., Jha, D., Thambawita, V., Halvorsen, P., Singstad, B.J., Gaur, S., Pettersen, K., Goodwin, M., Parasa, S., de Lange, T., Riegler, M., 2021. MedAI: Transparency in medical image segmentation. Nord. Mach. Intell. 1, 1–4.

Huang, C.H., Wu, H.Y., Lin, Y.L., 2021. Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps. arXiv preprint arXiv:2101.07172.

Hwang, S., Oh, J., Tavanapong, W., Wong, J., De Groen, P.C., 2007a. Polyp detection in colonoscopy video using elliptical shape feature. In: Proceedings of the IEEE International Conference on Image Processing, Vol. 2. pp. II–465.

Hwang, S., Oh, J., Tavanapong, W., Wong, J., Groen, P., 2007b. Polyp detection in colonoscopy video using elliptical shape feature. In: Proceedings of the International Conference on Image Processing, Vol. 2. ICIP, pp. II – 465.

Jha, D., Ali, S., Emanuelsen, K., Hicks, S.A., Thambawita, V., Garcia-Ceja, E., Riegler, M.A., de Lange, T., Schmidt, P.T., Johansen, H.D., et al., 2021. Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. In: MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II 27. pp. 218–229.

Jha, D., Hicks, S.A., Emanuelsen, K., Johansen, H., Johansen, D., de Lange, T., Riegler, M.A., Halvorsen, P., 2020a. Medico multimedia task at MediaEval 2020: Automatic polyp segmentation. In: Proceedings of the CEUR Worksh. Multim. Bench. Worksh. MediaEval, pp. 1–2.

Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P., Johansen, H.D., 2019. Resunet++: An advanced architecture for medical image segmentation. In: In Proceedings of the IEEE International Symposium on Multimedia. ISM, pp. 225–2255.

Jha, D., et al., 2020b. Kvasir-seg: A segmented polyp dataset. In: Proceedings of the Int. Conf. Multim. Model. MMM, pp. 451–462.

Ji, G.P., Fan, D.P., Zhou, T., Chen, G., Fu, H., Shao, L., 2020. Automatic polyp segmentation via parallel reverse attention network. In: Proceedings of the MediaEval. pp. 1–3.

Kang, J., Gwak, J., 2020. KD-ResUNet++: Automatic polyp segmentation via self-knowledge distillation.. In: Proceedings of the MediaEval. pp. 1–3.

Keprate, A., Pandey, S., 2021. Kvasir-instruments and polyp segmentation using UNet. Nord. Mach. Intell. 1 (1), 26–28.

Khadka, R., 2020. Transfer of knowledge: Fine-tuning for polyp segmentation with attention.. In: Proceedings of the MediaEval. pp. 1–3.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N., 2020. Big transfer (bit): General visual representation learning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. pp. 491–507.

Krenzer, A., Puppe, F., 2020. Bigger networks are not always better: Deep convolutional neural networks for automated polyp segmentation. In: Proceedings of the MediaEval. pp. 1–3.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep Learning, vol. 521, pp. 436–444.

Levin, B., et al., 2008. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. CA Cancer J. Clin. 58 (3), 130–160.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2117–2125.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.

Mahmud, T., Paul, B., Fattah, S.A., 2021. PolypSegNet: A modified encoder-decoder architecture for automated polyp segmentation from colonoscopy images. Comput. Biol. Med. 128, 104119.

Maxwell Hwang, C.W., Hwang, K.S., Xu, Y.S., Wu, C.H., 2020. A temporal-spatial attention model for medical image detection. In: Proceedings of the MediaEval. pp. 1–3.

Mehta, R., Sivaswamy, J., 2017. M-net: A convolutional neural network for deep brain structure segmentation. In: 2017 IEEE 14th International Symposium on Biomedical Imaging. ISBI 2017, pp. 437–440.

Mirza, A., Rajak, R.K., 2021. Segmentation of polyp instruments using UNet based deep learning model. Nord. Mach. Intell. 1 (1), 44–46.

Moriyama, T., Uraoka, T., Esaki, M., Matsumoto, T., 2015. Advanced technology for the improvement of adenoma and polyp detection during colonoscopy. Dig. Endosc. 27, 40–44.

Nathan, S., Ramamoorthy, S., 2020. Efficient supervision net: Polyp segmentation using EfficientNet and attention unit. In: Proceedings of the MediaEval. pp. 1–3.

Nguyen, T.P., Nguyen, T.C., Diep, G.H., Le, M.Q., Nguyen-Dinh, H.P., Nguyen, H.D., Tran, M.T., 2020. HCMUS at medico automatic polyp segmentation task 2020: PraNet and ResUnet++ for polyps segmentation. In: Proceedings of the MediaEval. pp. 1–3.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

Poudel, S., Lee, S.W., 2020. Automatic polyp segmentation using channel-spatial attention with deep supervision. In: Proceedings of the MediaEval. pp. 1–3.

Poudel, S., Lee, S.W., 2021. Explainable U-Net model forMedical image segmentation. Nord. Mach. Intell. 1 (1), 41–43.

Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jagersand, M., 2020. U2-Net: Going deeper with nested U-structure for salient object detection. Pattern Recognit. 106, 107404.

Rauniyar, S., Jha, V.K., Jha, R.K., Jha, D., Rauniyar, A., 2021. Improving polyp segmentation in colonoscopy using deep learning. Nord. Mach. Intell. 1 (1), 35–37.

Riegler, M., et al., 2016. Eir—efficient computer aided diagnosis framework for gastrointestinal endoscopies. In: Proceedings of the Inter. Worksh. Content-Based Multime. Index. CBMI, pp. 1–6.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18. pp. 234–241.

Shrestha, S., Khanal, B., Ali, S., 2020. Ensemble U-net model for efficient polyp segmentation. In: Proceedings of the MediaEval. pp. 1–3.

Siegel, R.L., Miller, K.D., Wagle, N.S., Jemal, A., 2023. Cancer statistics, 2023. CA Cancer J. Clin. 73 (1), 17–48.

Somani, A., Singh, D., Prasad, D., Horsch, A., 2021. T-MIS: Transparency adaptation in medical image segmentation. Nord. Mach. Intell. 1 (1), 11–13.

Tan, M., Le, Q., 2019a. Efficientnet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the International Conference on Machine Learning. pp. 6105–6114.

Tan, M., Le, Q.V., 2019b. EfficientNet: Rethinking model scaling for convolutional neural networks. CoRR, arXiv:1905.11946.

Thambawita, V., Hicks, S., Halvorsen, P., Riegler, M.A., 2020. Pyramid-focus-augmentation: Medical image segmentation with step-wise focus. arXiv preprint arXiv:2012.07430.

Tomar, N.K., 2021. Automatic polyp segmentation using fully convolutional neural network. arXiv preprint arXiv:2101.04001.

Tomar, N.K., Jha, D., Ali, S., Johansen, H.D., Johansen, D., Riegler, M.A., Halvorsen, P., 2021. Ddanet: Dual decoder attention network for automatic polyp segmentation. In: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part VIII. pp. 307–314.

Tomar, N.K., Jha, D., Riegler, M.A., Johansen, H.D., Johansen, D., Rittscher, J., Halvorsen, P., Ali, S., 2022. Fanet: A feedback attention network for improved biomedical image segmentation. IEEE Trans. Neural Netw. Learn. Syst..

Trinh, Q.H., Nguyen, M.V., Huynh, T.G., Tran, M.T., 2020. HCMUS-juniors 2020 at medico task in MediaEval 2020: Refined deep neural network and U-net for polyps segmentation. In: Proceedings of the MediaEval. pp. 1–3.

Tzavara, N.P., Singstad, B.J., 2021. Transfer learning in polyp and endoscopic tool segmentation from colonoscopy images. Nord. Mach. Intell. 1 (1), 32–34.

Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L., 2022. Pvt v2: Improved baselines with pyramid vision transformer. Comput. Vis. Media 8 (3), 415–424.

Yang, Y., Xing, Z., Zhu, L., 2024. Vivim: a video vision mamba for medical video object segmentation. arXiv preprint arXiv:2401.14168.

Yeung, M., 2021. Attention U-Net ensemble for interpretable polyp and instrument segmentation. Nord. Mach. Intell. 1 (1), 47–49.

Zhang, Y., Liu, H., Hu, Q., 2021. Transfuse: Fusing transformers and cnns for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 14–24.

Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., et al., 2022. Resnest: Split-attention networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2736–2746.