An objective validation of polyp and instrument segmentation methods in colonoscopy through Medico 2020 polyp segmentation and MedAI 2021 transparency 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[i], Ge-Peng Ji[o], Sahadev Poudel[m], George Batchkala[q], Saruar Alam[x], Awadelrahman M. A. Ahmed[h], Quoc-Huy Trinh[s], Zeshan Khan[r], Tien-Phat Nguyen[j], Shruti Shrestha[u], Sabari Nathan[v], Jeonghwan Gwak[w], Ritika K. Jha[a], Zheyuan Zhang[a], Alexander Schlaefer[f], Debotosh Bhattacharjee[d], M.K. Bhuyan[e], Pradip K. Das[e], Deng-Ping Fan[aa], Sravanthi Parasa[s], Sharib Ali[n], Michael A. Riegler[b,c], Pål Halvorsen[b,c], Thomas de Lange[y,z], Ulas Bagci[a]

[a]*Machine & Hybrid Intelligence Lab, Department of Radiology, Northwestern University, Chicago, USA*
[b]*Department of Holistic System, 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]*Jaeyong Kang, Information Systems Technology and Design, Singapore University of Technology and Design, Singapore*
[h]*University of Oslo, Norway*
[i]*Julius-Maximilian University of Würzburg, Germany*
[j]*Faculty of Information Technology, University of Science, VNU-HCM, Vietnam*
[k]*Vietnam National University, Ho Chi Minh City, Vietnam*
[l]*Department of Colorectal Surgery, the Second Affiliated Hospital of Zhejiang University School of Medicine, Zhejiang, China*
[m]*Department of IT Convergence Engineering, Gachon University, Seongnam 13120, South Korea*
[n]*School of Computing, University of Leeds, LS2 9JT, Leeds, United Kingdom*
[o]*College of Engineering, Australian National University, Canberra, Australia*
[p]*School of Computer Science, Wuhan University, Hubei, China*
[q]*Department of Engineering Science, University of Oxford, Oxford, UK*
[r]*National University of Computer and Emerging Sciences, Karachi Campus, Pakistan*
[s]*Swedish Medical Center, Seattle, USA*
[t]*Vietnam National University, Ho Chi Minh City, Vietnam*
[u]*NepAL Applied Mathematics and Informatics Institute for Research (NAAMII), Kathmandu, Nepal*
[v]*Couger Inc, Tokyo, Japan*
[w]*Department of Software, Korea National University of Transportation, Chungju-si, South Korea*
[x]*University of Bergen, Bergen, Norway*
[y]*Department of Medicine and Emergencies - Mölndal Sahlgrenska University Hospital, Region Västra Götaland, Sweden*
[z]*Department of Molecular and Clinical Medicin, Sahlgrenska Academy, University of Gothenburg, Sweden*
[aa]*Computer Vision Lab (CVL), ETH Zurich, Zurich, Switzerland*

ARTICLE INFO

Article history:
Received x xxxx xxxx
Received in final form x xxxx xxxx
Accepted x xxxx xxxx
Available online x xxxx xxxx

Communicated by xxxx xxxx

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*

*Corresponding author
e-mail: debesh.jha@northwestern.edu (Debesh Jha)

2021)" competitions. The Medico 2020 challenge pulled submissions from 17 different teams, while the MedAI 2021 challenge gathered submissions from 17 teams. 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 28. Through the comprehensive analysis of the challenge, we not only highlight the advancements in polyp and surgical instrument segmentation but also encourage qualitative 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 with the ultimate goal of reducing the cancer burden and improving patient care.

1. Introduction

Gastrointestinal 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. That's why there has been a push to develop 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, it is challenging to interpret the decision or understand the features

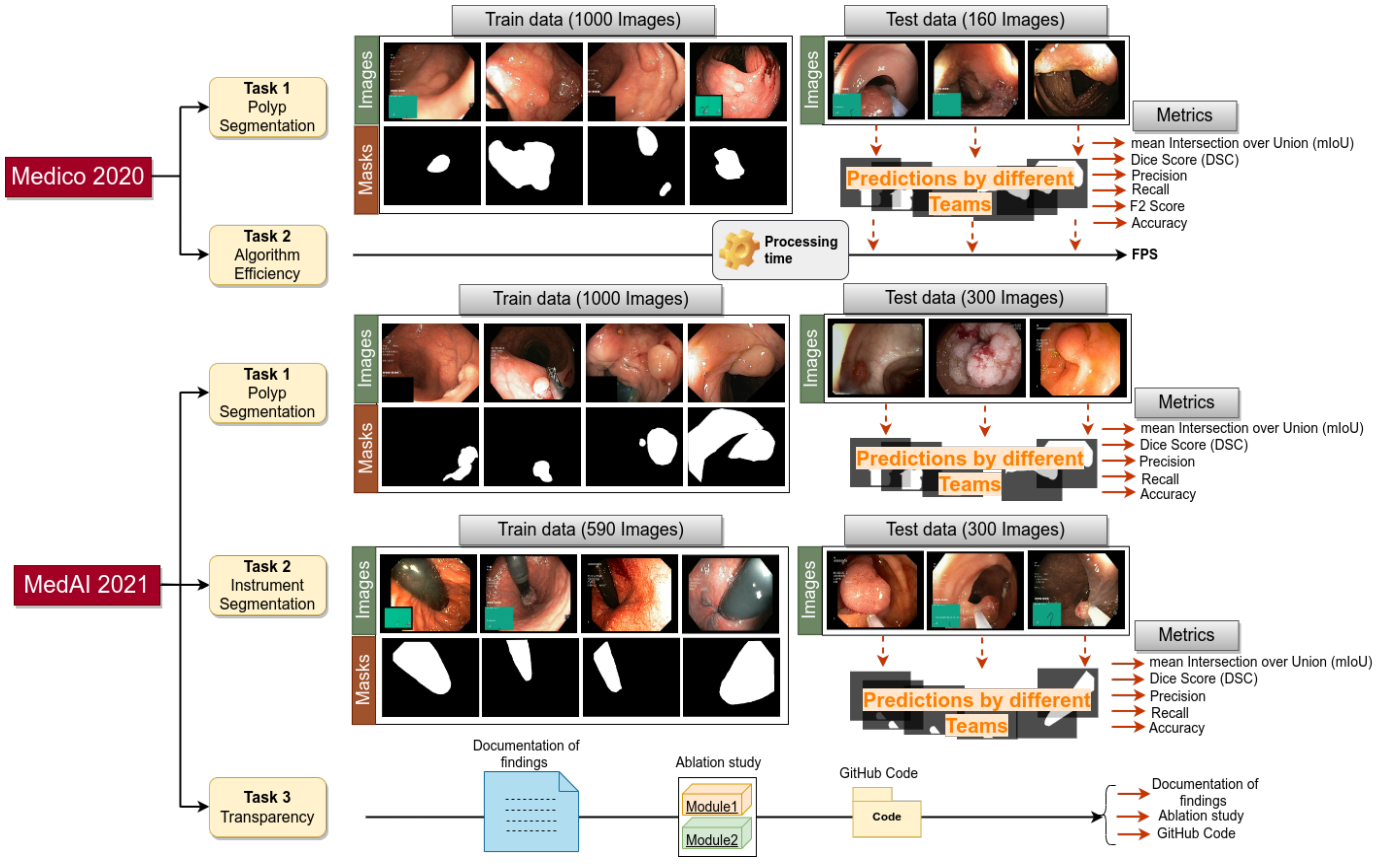


Fig. 1: The overview of the "***Medico 2020 Polyp***" and "***MedAI 2021 Transparency*** " challenges. We describe each task along with the number of training and testing datasets and the evaluation metrics used in the tasks.

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 were designed to foster the development of CADx solutions on the same datasets, with a focus on transparency, explainability, robustness, speed, and generalization, aiming to evaluate the relevance of such algorithms in clinical workflows. The challenges provided posed four distinct tasks:

- Accurate polyp segmentation task to develop state-of-the-art algorithms for early detection and treatment of colon cancer (Medico 2020, MedAI 2021).
- Algorithm efficiency task to develop methods with the least 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 the proposed system from a transparency point of view (for example, explanations of

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

[2] https://www.nora.ai/competition/image-segmentation.html

the training procedure, amount of data used and model's predictions interpretation) (MedAI 2021).

These tasks were focused on the development of state-of-the-art (SOTA) algorithms for polyp, instrument and medical image 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. Figure 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 generalizability issues, multi-center and out-of-distribution testing in context to current limitations of computer-aided diagnosis systems.

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. This challenge aimed to benchmark automated polyp segmentation algorithms using the same dataset and to develop methods that can 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.

2.1.1. Task Description

The participants were invited to submit their solutions for the following tasks.

***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).

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 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.

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

2.2.1. *Task Description*

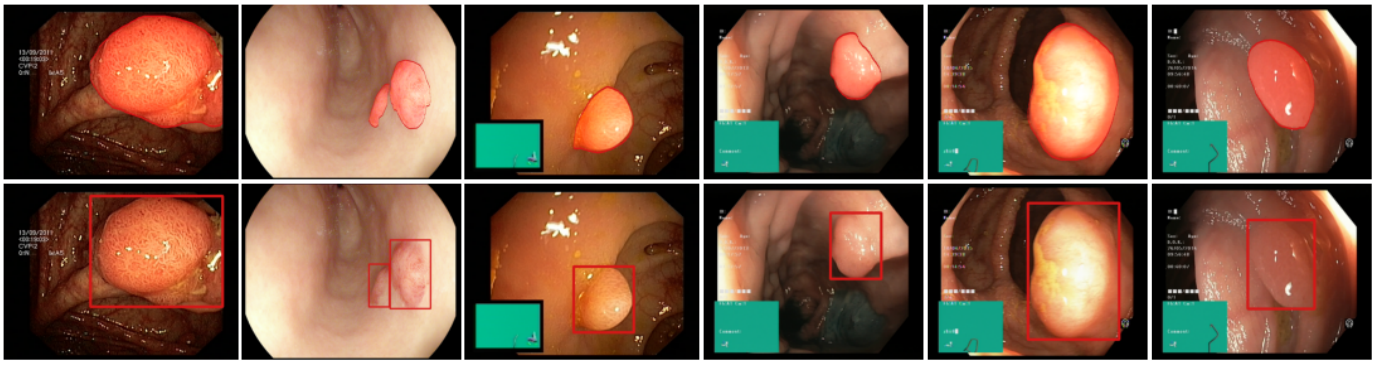
We present three tasks: the polyp segmentation task, the instrument segmentation task, and the transparency task. Each task targets a different requirement within automatic findings segmentation in Gastrointestinal (GI) image analysis.

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.

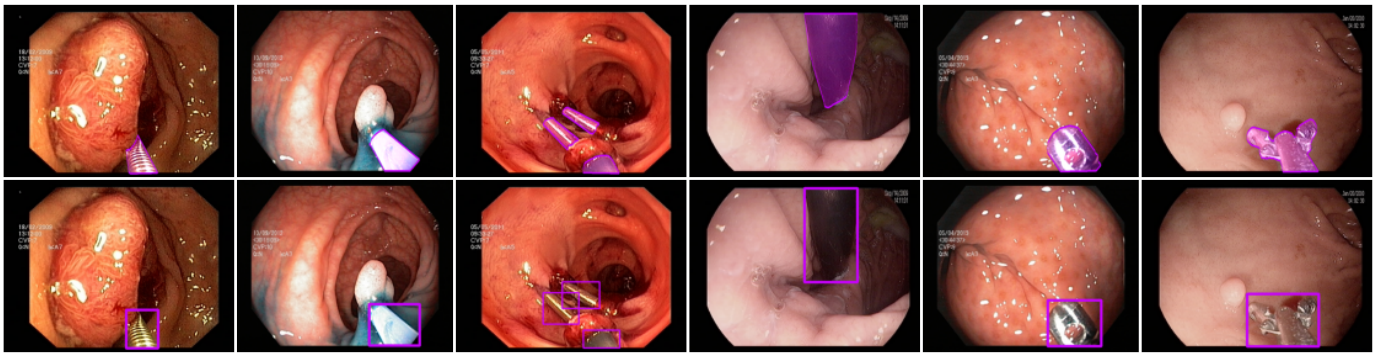
b) Automatic Instrument Segmentation Task: The instrument segmentation task required the development of algorithms that could generate segmentation masks for GI accessory instruments such as biopsy forceps or polyp snares used during live endoscopy procedures. This task aimed to create segmentation models 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 procedures 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 of their proposed algorithms, and share their code in a GitHub repository with clear steps for reproducing the results. In addition, participants were required to submit a one-page document summarizing their findings from the transparency task. 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, tasks, and evaluation metrics can also be found in (Hicks et al., 2021a).

[3] `https://nordicaimeet.com`



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



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

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

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; Shin and Balasingham, 2018). 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 Convolutional Neural Networks (CNNs), polyp segmentation task has seen 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. These challenges have limitations such as small datasets (Ali *et al.*, 2020) or datasets that are not used consistently across different challenges (for example, EndoVis2015 challenge on Early Barrett's cancer detection[4]). Additionally, the algorithms proposed in these challenges are often not publicly available, making it difficult to reproduce and build upon them. Hence, there is a need for benchmarking datasets and for making the algorithms proposed in these challenges reproducible to facilitate progress in this field.

[4] https://endovissub-barrett.grand-challenge.org

Table 1: Overview of **polyp segmentation challenges** in the past 8 years. Here, WL = White Light Endoscopy, NBI = Narrow Band Imaging, WCE = Wireless Capsule Endoscopy, FL = Fluorescence 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 2021 (Hicks et al., 2021b)	Colon	WL	Polyps	300 images (test)	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., 2021a) (Ours)	Colon, bladder	WL	Polyps, Instrument, Normal frames	600 images (test) & (1000 +590) images (train)	Open-academic

Table 1 provides an overview of GI challenges held in the past eight years, including the imaging modalities used. Several challenges have been organized in the past eight years to compare and improve computer vision classification methods and benchmark GI endoscopy image datasets. In 2015, Bernal et al. (Bernal et al., 2017) organized the "Automatic Polyp Detection in colonoscopy videos" challenge. Likewise, they organized the GIANA challenge in 2017 and 2018[5] focused on colonoscopy data and included tasks such as detection of lesions in Video capsule Endoscopy (VCE), polyp detection, and polyp segmentation. The EndoCV2020 challenge[6] included a sub-challenge on "Endoscopy disease detection (EDD2020)" with multi-organ and multi-modal endoscopy data, but only 386 annotated frames and 5 class categories were included. Recent challenges attempted to address generalisability in polyp detection and segmentation (Ali et al., 2022a) with both 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 the detection and classification of GI abnormalities.

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.

Medico 2020 (Jha et al., 2020a) focused on promoting new algorithmic innovations in polyp segmentation and assessing algorithmic efficiency, while MedAI 2021 (Hicks et al., 2021a) emphasizes innovations in both polyp segmentation and instrument segmentation. In addition, the MedAI 2021 challenge also introduces a transparency task that encourages and validates reproducible research and a focus on the explainability of model predictions. Through these two challenges, we aimed to address some of the key challenges in GI endoscopy, including benchmarking of datasets, reproducibility of algorithms, and explainability of model predictions. In this paper, we comprehensively

[5] https://giana.grand-challenge.org/
[6] https://endocv.grand-challenge.org

analyze the outcomes of both challenges.

4. Challenge datasets and evaluation metrics

4.1. Medico 2020 dataset

The dataset contains 1,000 polyp images and their corresponding ground truth mask taken from Kvasir-SEG (Jha et al., 2020b). The datasets were collected from real routine clinical examinations at Vestre Viken Health Trust (VV) in Norway by 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 of the images contain green thumbnails in the lower-left corner of the images showing the position marking from the ScopeGuide (Olympus). The training dataset can be downloaded from [7]. The test dataset contains unique polyp dataset from the same distribution (collected from the same center). It can be downloaded from [8]. Some samples are shown in Figure 2a.

4.2. MedAI Transparency challenge 2021 dataset

For the MedAI transparency challenge as well, we use Kvasir-SEG (Jha et al., 2020b) as the training dataset. The development dataset for the polyp segmentation task can be downloaded from [9] whereas the development dataset for the instrument segmentation task can be downloaded from [10]. Some sample images are shown in Figure 2a and Figure 2b. Figure 3 shows the data distribution of the train and test datasets used in Medico 2020 and MedAI 2021. We have categorized the datasets into "small", "medium" and "large" according to the size of regions of interest and plotted the height versus 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 are crucial for assessing the performance, robustness, and generalizability of the proposed algorithms.

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

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

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

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

4.3. Metrics for polyp and instrument segmentation tasks

We used mIoU as an evaluation metric for the polyp and instrument segmentation tasks. If the teams had the same mIoU values, they are further evaluated based on the higher value of the Dice coefficient (DSC). We also recommend calculating other standard evaluation metrics such as Precision (Pre), pixel accuracy (Acc.), Recall, (Rec), F2-Score (F2), and FPS for a comprehensive evaluation.

4.4. Metrics for efficiency tasks

Efficiency is essential during GI endoscopy as it directly impacts the models' feasibility and practicality in real-world scenarios. For example, in a clinical setting, endoscopists may need to analyze a large number of frames in real-time during routine endoscopy (upper GI endoscopy, enteroscopy or colonoscopy), and lag in the analysis could lead to suboptimal results. Therefore, we strongly recommend calculating processing speed in terms of FPS as an evaluation metric for the polyp segmentation tasks.

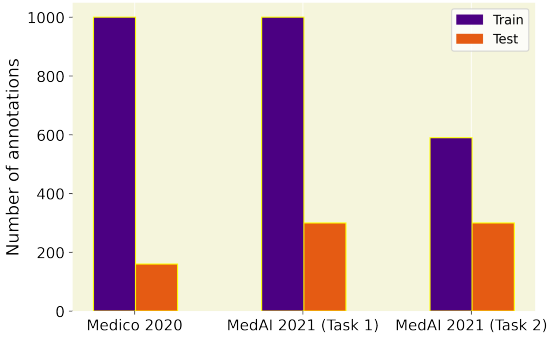
4.5. Metrics for transparency tasks

The transparency task aims to assess the transparency and understandability of algorithms for medical AI by utilizing a qualitative approach in the evaluation metrics. The evaluation team, comprising multiple experts from diverse fields evaluated the submissions based on various attributes such as the availability of code, the depth of evaluation, reproducibility, and the implementation of explainable AI techniques. In addition, participants are expected to provide detailed information about their solution, including rigorous failure analysis, thorough ablation studies, and a comprehensive GitHub repository with clear reproducibility steps. The transparency evaluation is divided into three categories: open source, model evaluation, and clinical evaluation, with corresponding scores outlined in Table 10. Ultimately, this task aims to promote the development of more transparent and interpretable medical AI systems.

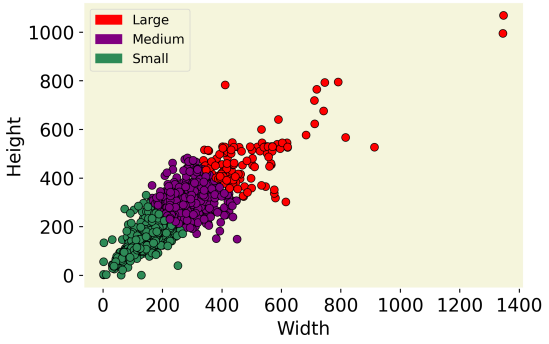
5. Participating Research Teams

5.1. Methods used in Medico 2020

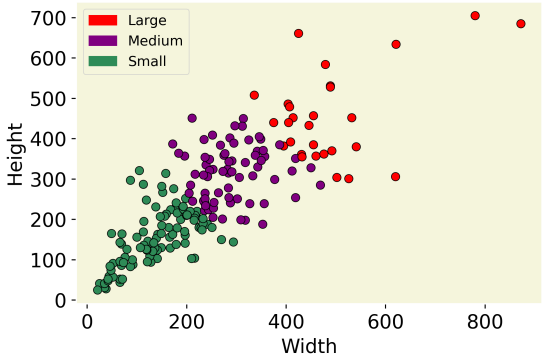
In Table 2 we have provided summary of all the teams who participated in "Medico 2020" challenge. It can be seen from



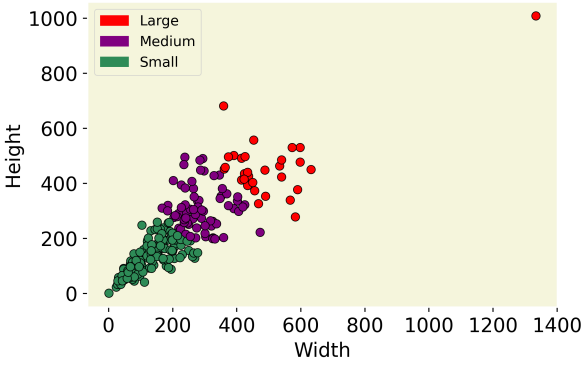
(a) Train and test sample proportion for Medico 2020 and MedAI 2021



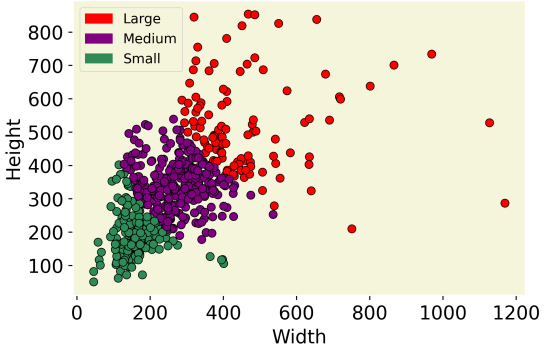
(b) Train data (Medico 2020 and MedAI 2021 Task 1)



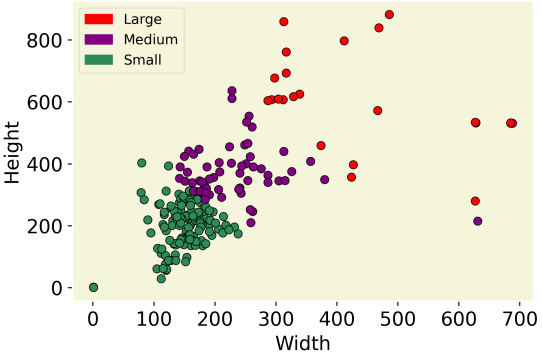
(c) Test data (Medico 2020)



(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 present the size distribution information of regions of interest in the data samples.

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	√	√

Table 2 that most of the teams participated in only one subtask (Task 1) whereas 9 teams participated in both Task 1 and 2 of the challenge.

FAST-NU-DS: Team FAST-NU-DS (Ali et al., 2020b) 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 (ASPP) was also implemented on the ResUNet++ architecture. The purpose of this architectural design was to provide good performance on the image segmentation evaluation metrics as well as in terms of inference time. The implementation of depth-wise separable convolution resulted in less number of parameters and giga-floating point operations (GFLOPs). To get the lightweight model architecture the convolution layer in the atrous bridge was replaced with depth-wise separable convolution and the atrous bridge was also replaced with a deep atrous bridge. The comparison of modification in model architecture was made against UNet (Ronneberger et al., 2015) and ResUNet++. All models were trained on custom mean intersection over union loss.

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 (CSCA) 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 framework for solving the automatic segmentation problem. Perceiving the problem as an image-to-image translation task, conditional generative adversarial networks 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.

HBKU_UNITN_SIMULA: HBKU_UNITN_SIMULA (Trinh et al., 2020) team proposed methods combining the Residual module, Inception module, Adaptive CNN with U-Net model, and PraNet 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 Resnet50 backbone was used to build ResUNet, yielding better obtained results. Last, PraNet was employed for polyp segmentation in colonoscopy images.

AI-JMU: Team AI-JMU (Krenzer and Puppe, 2020) explored various image segmentation models, specifically the Cascade Mask R-CNN and Mask R-CNN with ResNet as well as the ResNeSt 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.

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 (Ronneberger et al., 2015). 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 increases and error rate decreases.

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), 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.

HGV-HCMUS: Team HGV-HCMUS (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.

GS-CDT: Team GS-CDT (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 result in enhanced segmentation performance, with a nearly 3% improvement on the DSC metric while maintaining real-time performance (more than 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 Figure 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: 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 effi-

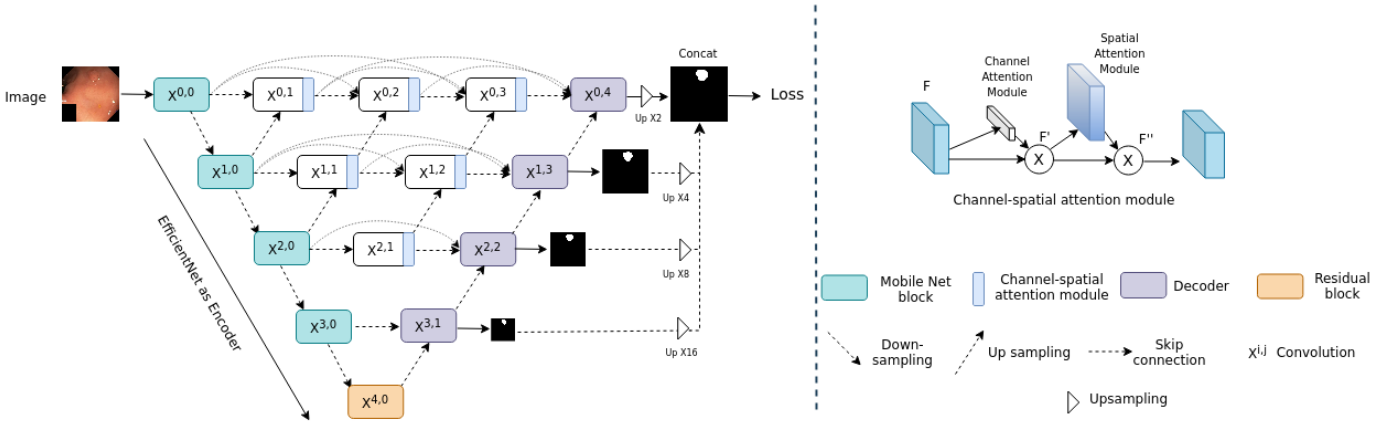


Fig. 4: Overview of the winning solution for the Polyp segmentation task (Task 1) from Team ***PRML2020GU***. The architecture utilizes pre-trained weights from EfficientNet in the encoder. Additionally, it uses dense skip connections, deep supervision and channel-spatial attention for fast convergence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

cient 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 consists 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, loss function, data augmentation, and optimizer used by the different 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. The challenge has three subtasks: polyp segmentation task, accessory instrument (eg. biopsy forceps or polyp snare) segmentation task, and transparency task. 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 four teams which participated in either one or two of the sub-tasks. Most participating teams have used the same architecture in their submission for both 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 accessory 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.

Table 3: Summary of the participating teams algorithm for Medico 2020.

Team Name	Algorithm	Backbone	Nature	Choice basis	Aug.	Loss	Optimizer
FAST-NU-DS (Ali et al., 2020b)	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 + DSC 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
HCMUS-Juniors (Trinh et al., 2020)	Residual module, Inception module, Adaptive CNN with U-Net and PraNet	U-Net and Resnet50	Cascade of residual blocks and inception module	mIoU and DSC	Yes	—	—
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	NA	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 (Nguyen et al., 2020)	PraNet and ResUNet++ with triple path	ResUNet++	Encoder-decoder	mIoU	Yes	Categorical crossentropy	Adam
GS-CDT (Batchkala and Ali, 2020)	U-Net	None	Encoder-decoder	Acc and Speed	Yes	Non-Binarized mIoU	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	—	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

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.

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	√	√	–

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 point-wise 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 Atrous Spatial Pyramid Pooling (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.

leen: Team leen (Ahmed and Ali, 2021) utilized the generative adversarial networks (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.

NYCity: Team NYCity (Chen et al., 2021) presented a novel multi-model ensemble framework for medical image segmentation. The team first collected a set of state-of-the-art models in this field and further improved them through a series of architecture refinement moves and a set of specific training skills. By integrating those fine-tuned models into a more powerful ensemble framework, they were able to achieve improved performance. The proposed multi-model ensemble framework was tested on polyp and instrument datasets and experiment results have shown that it performed satisfactorily.

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.

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.

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 embedded into the SINetV2 (Fan et al., 2021) framework. The SINetV2 framework with camouflaged object detection (COD) 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 Figure 5. The first encoder-decoder network processes the original image and produces output that is 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. To train their models, the team employed a 4-fold cross-validation approach, training with four separate models and used temperature sharpening on the ensemble 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. The evaluation using the unseen testing dataset resulted in an IOU of 0.29 and a DSC of 0.41 for the polyp segmentation task.

CamAI: Team CamAI (Yeung, 2021) presented a deep learning pipeline that is specifically developed to accurately segment

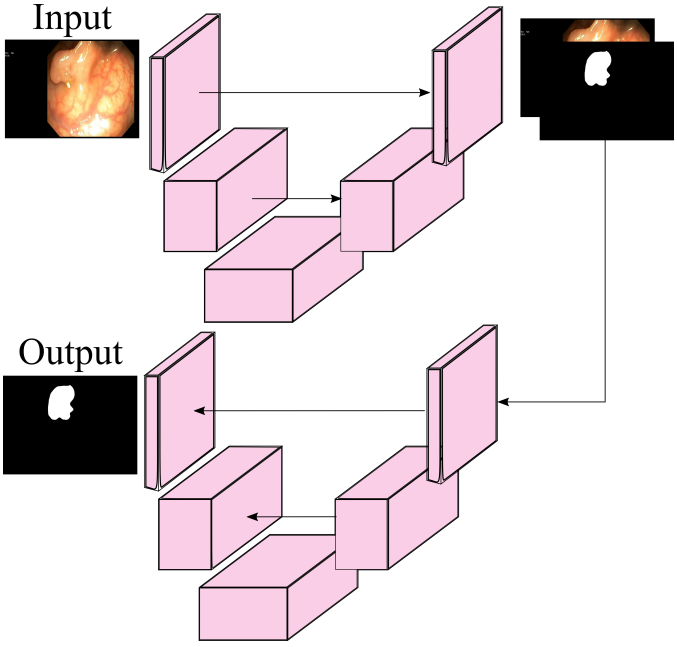


Fig. 5: Overview of winning solution of MedAI 2021 proposed by Team ***agaldran*** . A double encoder-decoder network was used to segment polyps and surgical instruments.

colorectal polyps and various instruments used during endoscopic procedures. To 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.

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.

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 U2-Net 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 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 gave more emphasis was given to mIoU, DSC and FPS. We have highlighted the best and the second-best scores

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	—	—
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	—	—
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	PPA loss	—
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

in boldface and red color, respectively, for all the tasks in the two challenges.

6.1. Medico 2020 Polyp segmentation results

In Table 6 and Table 7, we provide the results for the *polyp segmentation* task and *algorithm efficiency* task for the challenge "Medico 2020". A total number of 17 teams participated in the first task and 9 teams participated in the second task. The teams were ranked based on the mIoU metric for the first task, and for the second task, the teams were ranked based on the FPS metric. It can be observed from Table 6 that Team "PRML2020GU" outperforms other participating teams in the polyp segmentation task. It achieves a mIoU of 0.7897, DSC of 0.8607, recall of 0.9031, precision of 0.8673, accuracy of 0.9546 and F2 of 0.8748. Team "AI-TCE" was the second best performing team with mIoU of 0.7770 and "IIAI-Med" was the third best performing team. The best performing team "PRML2020GU" used an encoder-decoder structure with EfficientNet as the backbone and a U-Net decoder with channel-spatial attention with deep supervision. This architecture had an improvement of 1.24% and 1.31% over the mIoU and DSC achieved by the Team "AI-TCE" which used PraNet and Res2Net backbone.

For the second task, 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 can be observed from Table 7 and Figure 10a. However, it is worth noting that the team obtained a low mIoU of 0.6351 for the polyp segmentation task as evident from Table 6. Despite the two teams, "PRML2020GU"' and "GeorgeBatch", achieving the highest evaluation metric values, there is a trade-off between these performance metrics. 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 obtained by all the research teams who participated in Medico 2020 challenge in Figure 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 polyp and accessory instrument segmentation challenge results

In Tables 8 and 9, we tabulated the evaluation results of all the participating teams in MedAI 2021 for the two tasks, namely polyp segmentation and instrument segmentation. A total of 15 teams participated in both tasks of the MedAI2021 challenge. However, two teams, namely "Vyobotics (Rauniyar et al., 2021)" and "MedSeg_JU (Banik et al., 2021)" participated in only one task. Almost all the teams have used the same architecture for both tasks. The participating teams for the two tasks were ranked based on the mIoU metric. From Table 8, it can be observed that team "agaldran" outperforms other teams in the polyp segmentation task with mIoU of 0.8522, DSC of 0.8965, accuracy of 0.9791, recall of 0.9009 and precision of 0.9242. Team "IIAI-CV&Med" also showed good performance and was ranked 2^{nd} in the polyp segmentation task with a DSC of 0.8927, a very small difference from the best-performing team. In Figure 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. Overall, the qualitative masks produced by teams "agaldran" and "IIAI-CV&Med" were better as compared to the other teams.

We present the results of the accessory instrument segmentation task in Table 9. From the table, 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 2^{nd} 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 predicted segmentation mask compared to team "agaldran". Another interesting observation is the team "agaldran"

Table 6: Polyp segmentation task (Medico 2020)

Team Name	mIoU ↑	DSC ↑	Recall ↑	Precision ↑	Accuracy ↑	F2 ↑
PRML2020GU	**0.7897**	**0.8607**	0.9031	0.8673	0.9546	0.8748
HGV-HCMUS	0.4058	0.5148	0.5072	0.7574	0.9011	0.5007
AI-TCE	0.7770	0.8503	**0.9164**	0.8389	0.9566	**0.8790**
HBKU_UNITN_SIMULA	0.7537	0.8309	0.8399	0.8764	0.9581	0.8303
IIAI-Med	0.7619	0.8385	0.8304	**0.9012**	**0.9602**	0.8283
SBS	0.7550	0.8316	0.8316	0.8851	0.9582	0.8249
ML-MMIV Saruar	0.7516	0.8228	0.8390	0.8822	0.9564	0.8249
AI-JMU	0.7374	0.8143	0.8266	0.8743	0.9463	0.8103
MedSeg_JU	0.7133	0.8019	0.8354	0.8286	0.9446	0.8124
VT	0.7057	0.7926	0.8830	0.7878	0.9331	0.8236
NKT	0.6847	0.7801	0.8077	0.8126	0.9404	0.7854
UNITRK	0.6437	0.7287	0.7098	0.8572	0.9432	0.7131
GeorgeBatch	0.6351	0.7327	0.7500	0.8229	0.9422	0.7361
AMI Lab	0.6195	0.7088	0.7286	0.7914	0.9325	0.7122
IRIS-NSYSU	0.5035	0.6417	0.8791	0.5849	0.8726	0.7508
UiO-Zero	0.4381	0.5618	0.6972	0.5558	0.8806	0.6110
FAST-NU-DS	0.1834	0.2669	0.2744	0.2918	0.8272	0.2676

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).

Team Name	mIoU ↑	DSC ↑	Recall ↑	Precision ↑	Accuracy ↑	F2 ↑	FPS ↑
HCMUS	**0.7364**	0.8074	0.8164	**0.8646**	**0.9572**	0.8067	33.27
SBS	0.7341	**0.8148**	0.8764	0.8145	0.9452	**0.8354**	26.66
NKT	0.6847	0.7801	0.8077	0.8126	0.9404	0.7854	80.60
FAST-NU-DS	0.6582	0.7556	**0.8982**	0.7171	0.9255	0.8109	67.51
UNITRK	0.6437	0.7287	0.7098	0.8572	0.9432	0.7131	116.79
GeorgeBatch	0.6351	0.7327	0.7500	0.8229	0.9422	0.7361	**196.79**
AMI Lab	0.6195	0.7088	0.7286	0.7914	0.9325	0.7122	107.87
AI-JMU	0.7213	0.8017	0.8359	0.8495	0.9345	0.8056	3.36
PRML2020GU	0.5083	0.6265	0.6003	0.7870	0.9149	0.6029	2.25

Table 8: Performance of teams in polyp segmentation task of MedAI:Transperancy in Medical Image Segmentation (MedAI 2021)

Team Name	Accuracy ↑	mIoU ↑	DSC ↑	Recall ↑	Precision ↑
agaldran	**0.9791**	**0.8522**	**0.8965**	0.9009	0.9242
IIAI-CV&Med	0.9593	0.8361	0.8927	**0.9195**	0.8963
NYCity	0.9735	0.8418	0.8885	0.8794	**0.9319**
mTEC	0.9679	0.8334	0.8892	0.9010	0.9096
CV&Med IIAI	0.9766	0.8213	0.8612	0.8602	0.8814
PRML	0.9715	0.8116	0.8669	0.8852	0.8922
CamAI	0.9663	0.8083	0.8701	0.8702	0.9052
The Arctic	0.9730	0.8022	0.8533	0.8604	0.8821
Polypixel	0.9701	0.7997	0.8567	0.8868	0.8659
MAHUNM	0.9654	0.7495	0.8189	0.8397	0.8568
OXGastroVision	0.9385	0.7334	0.7966	0.8158	0.8374
Vyobotics	0.9557	0.7220	0.7967	0.8214	0.8359
NAAMII	0.9082	0.6041	0.6940	0.7499	0.7334
leen	0.8399	0.4595	0.5531	0.6389	0.5860
The Segmentors	0.8922	0.3789	0.4205	0.4178	0.4640
TeamAIKitchen	0.5646	0.2904	0.4100	0.7152	0.4910

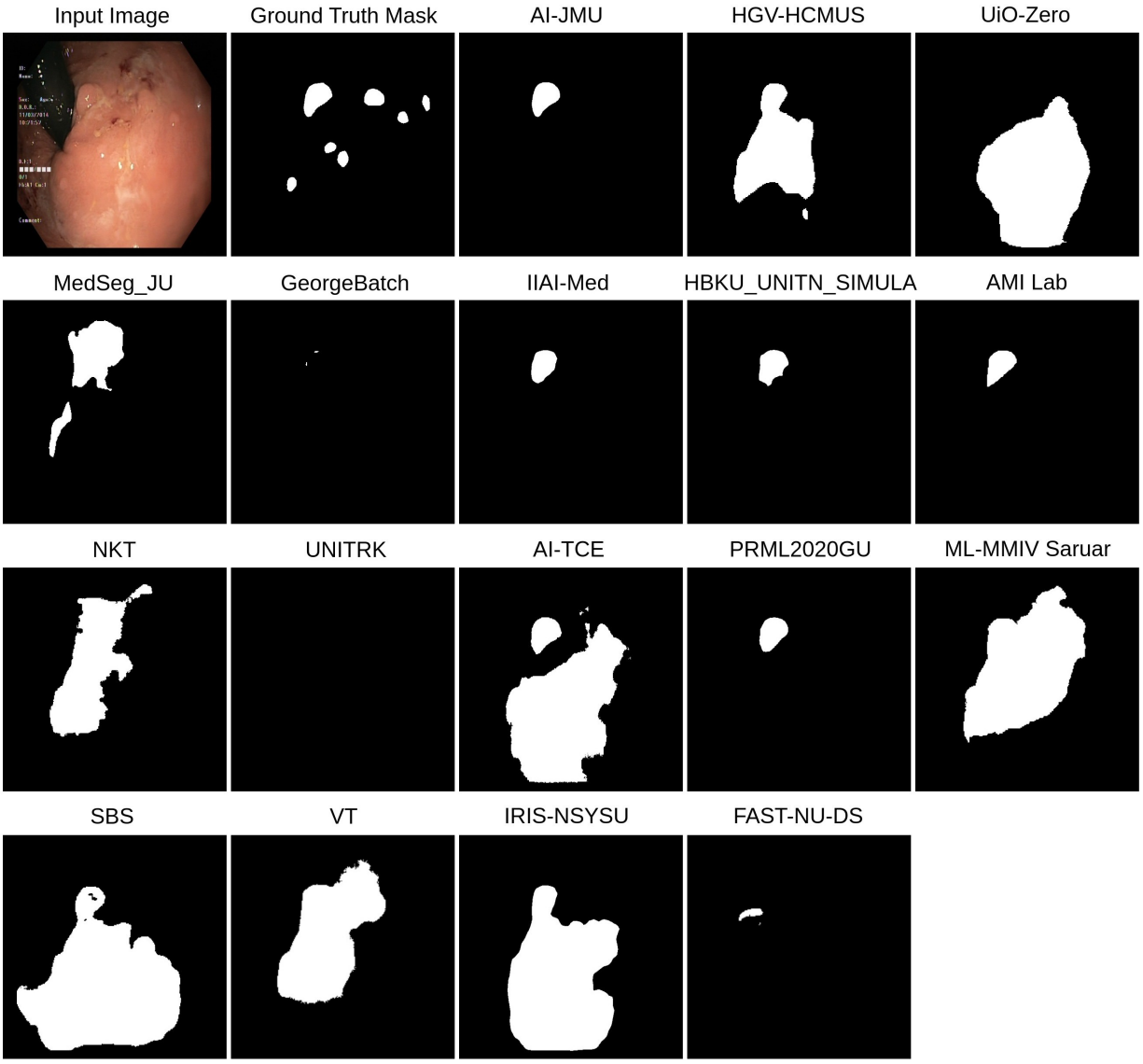


Fig. 6: The figure shows the qualitative results of participating teams for the polyp segmentation task in the Medico 2020 Challenge on challenging scenarios. When each team's predicted mask is compared with its corresponding ground truth, we can observe that none of the teams obtained results that fit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

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 Figure 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.

A detailed score distribution under different criteria is shown in Table 10, which was part of our ***Task 3 (transparency task)***. Here, submissions that did not participate in task 3 are left blank. We have evaluated transparency tasks using a more quantitative approach compared to 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 evalua-

Table 9: Performance of teams in instrument segmentation task of MedAI:Transperancy in Medical Image Segmentation (MedAI 2021)

TeamName	Accuracy ↑	mIoU ↑	DSC ↑	Recall ↑	Precision ↑
agaldran	**0.9941**	**0.9364**	**0.9635**	0.9692	**0.9632**
NYCity	0.9937	0.9326	0.9586	**0.9712**	0.9516
mTEC	0.9913	0.9245	0.9553	0.9687	0.9490
PRML	0.9901	0.9178	0.9528	0.9687	0.9441
IIAI-CV&Med	0.9885	0.9148	0.9490	0.9612	0.9473
CV&Med IIAI	0.9736	0.9136	0.9512	0.9605	0.9500
Polypixel	0.9860	0.9114	0.9478	0.9591	0.9438
CamAI	0.9871	0.9079	0.9442	0.9527	0.9468
The Arctic	0.9912	0.9078	0.9448	0.9735	0.9231
OXGastroVision	0.9854	0.8692	0.9073	0.9236	0.9096
MAHUNM	0.9859	0.8523	0.9080	0.9535	0.8864
MedSeg_JU	0.9799	0.8205	0.8632	0.9005	0.8464
TeamAIKitchen	0.9721	0.7257	0.7980	0.7955	0.8510
leen	0.9634	0.6991	0.7845	0.7963	0.8232
NAAMII	0.9597	0.6857	0.7741	0.8321	0.7669
The Segmentors	0.9250	0.3668	0.3971	0.3985	0.4040

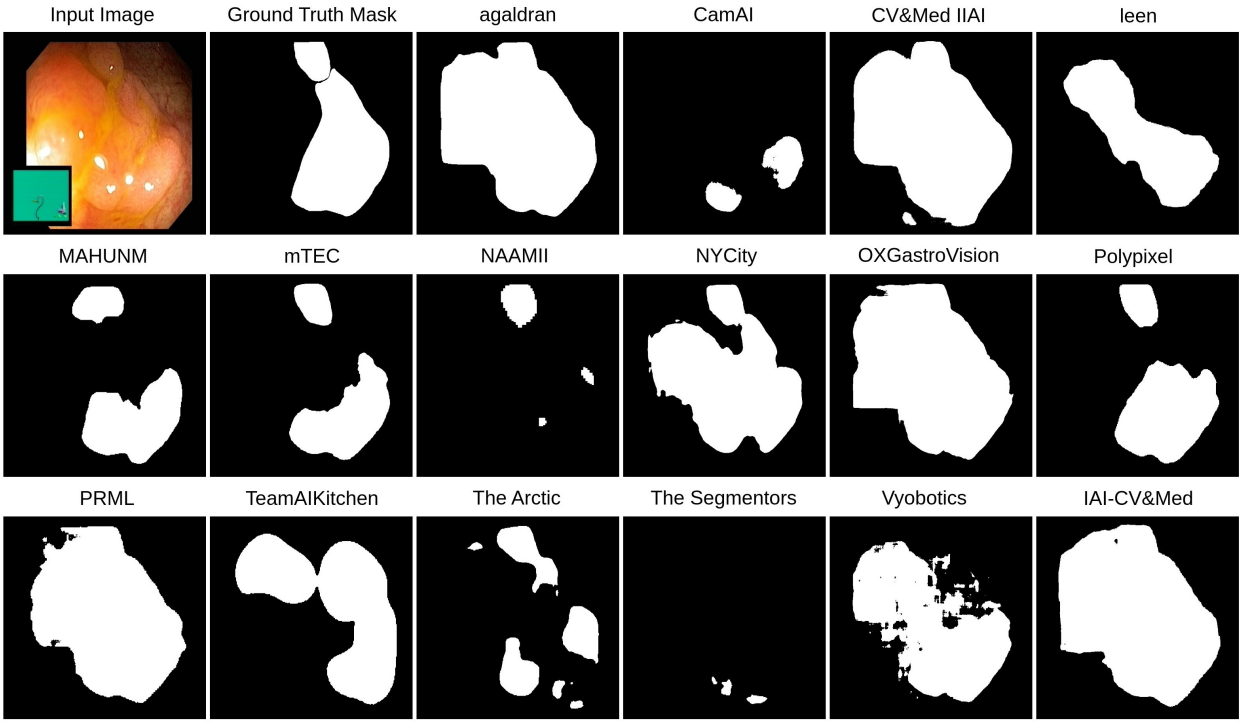


Fig. 7: Qualitative results of all the methods participating in polyps segmentation challenge in MedAI2021.

tion. 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 consider the usefulness of the method and its interpretability. With these three criteria, we aim to measure the transparency of the provided solutions. We present the 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 2^{nd}. Likewise, team "CamAI" obtained a score of 16 out of 25 and was ranked third in the transparency task. There were also efforts from team such as "The Arctic" which obtained a score of 13 out of 25 and IIAI-CV&Med that obtained a score of 10

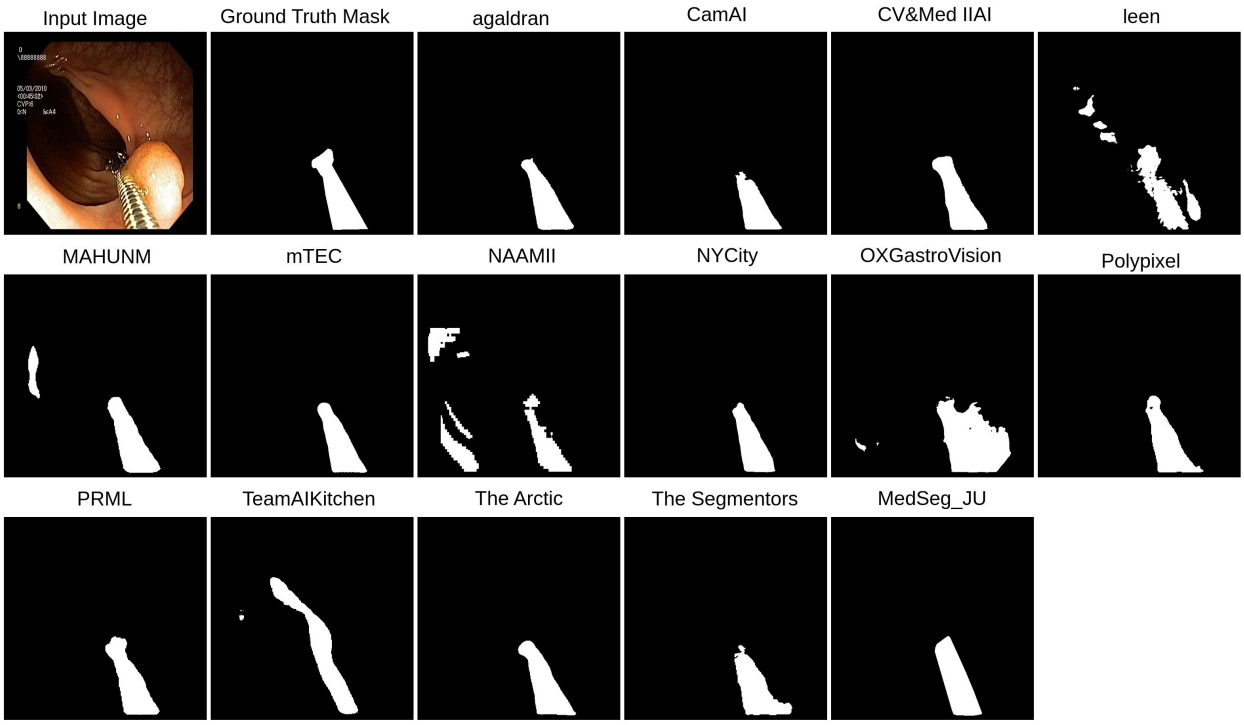


Fig. 8: Qualitative results of all the methods participating in surgical instrument segmentation challenge in MedAI2021.

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).

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

out of 25. 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 Figure 9. Notably, team "agaldran" outperforms others in all three tasks and overall challenge and emerged as the winner of the MedAI challenge. Overall, mTec secured the second position. Following closely behind, CamAI showcased the third-best solution.

Figure 10b illustrates the plot of DSC 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

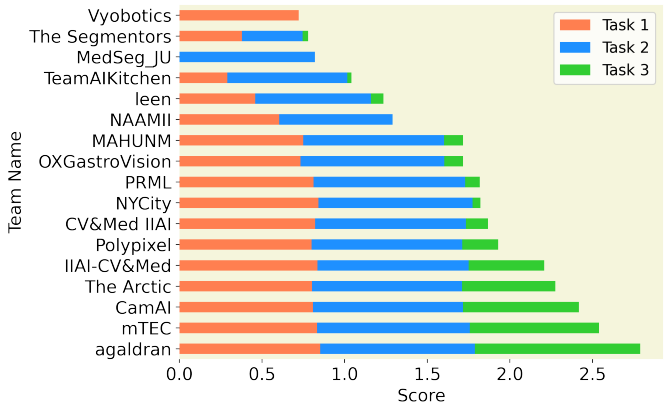


Fig. 9: Task-wise scores achieved by participating teams of MedAI 2021 challenge. Team rankings are decided on the basis of overall scores in all three tasks.

2020 to 2021 gained improvement with a larger number of submissions achieving a DSC of more than 0.80 and the best performing team with a DSC of around 0.90. Similar progress can be observed in Figure 10c where an overall DSC increased by 2.33% when an average score is computed over all participating teams' individual best DSC in the 2021 polyp segmentation challenge. We further compared all segmentation metrics, including DSC, recall, precision, mIoU score, accuracy, and F2 score, as shown in Figure 10d. Notably, all teams' accuracy is consistently high, with a score close to 0.90. Similarly, the different scores of evaluation metrics are consistent for 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 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., 2020a). Besides developing and analyzing AI-based algorithms, it is crucial to include explainability and interpretability to infuse trust and reliance during the adoption of automated systems in clinical settings (Ali, 2022). 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 and limitations of the two challenges, Medico 2020 and 2021.

The findings from Medico 2020 and MedAI 2021 challenges provide valuable insight and trends for the current biomedical image analysis challenge. All the participants in both challenges used deep learning-based frameworks for segmentation tasks. It can be observed that most of the deep learning frameworks submitted for the challenge used Adam optimizer for optimizing their network. However, a handful of teams used other optimizers such as stochastic gradient descent (SGD) and RMSProp. We also observed that most of the deep learning frameworks used by participants in both challenges used the encoder-decoder framework as the backbone network. Additionally, most of the teams used data augmentation to boost the

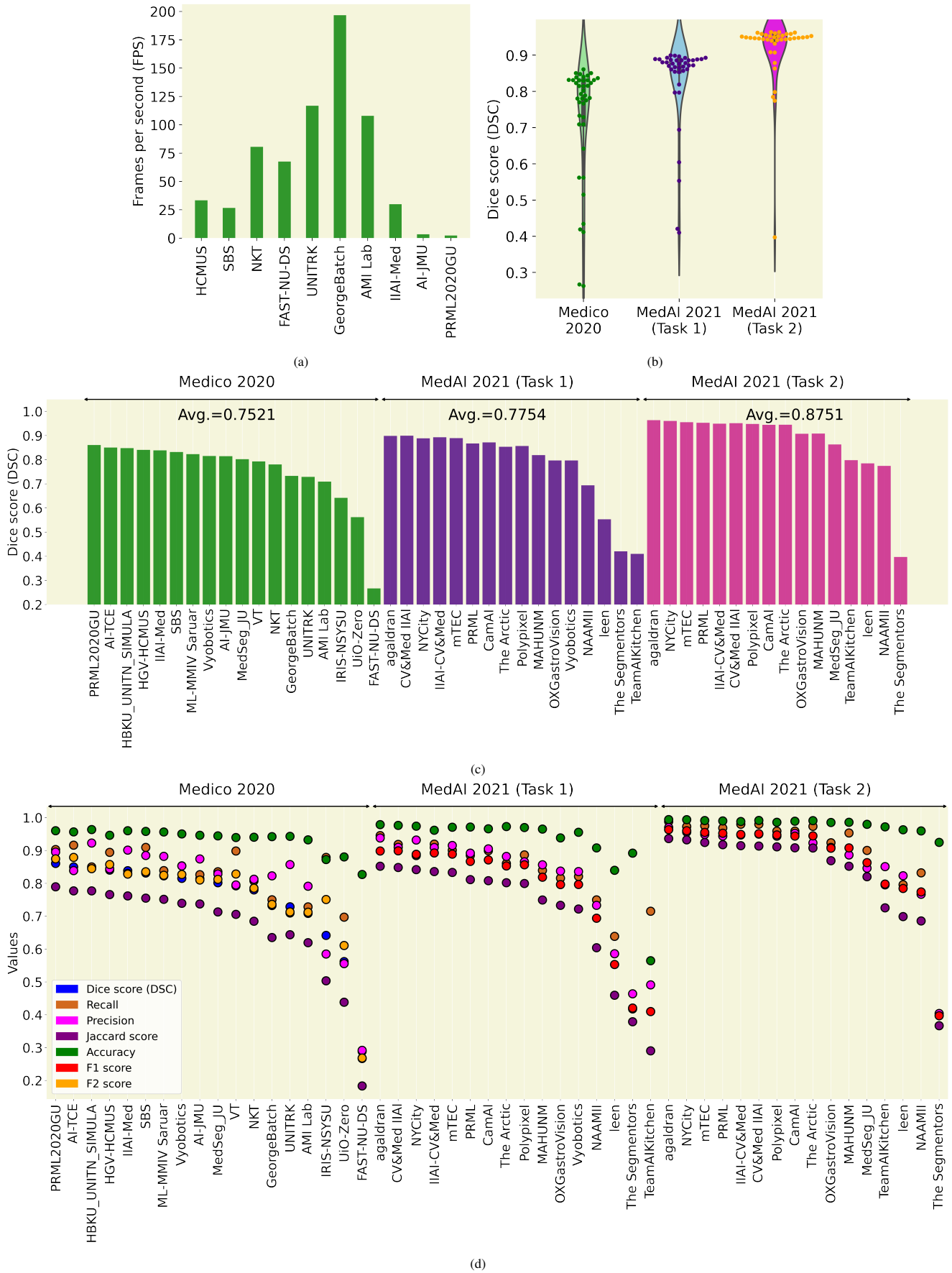


Fig. 10: (a) Processing speed analysis by comparing FPS achieved by participating teams in 2020, (b) violin plots with overlaid swarm plots depicting statistics of submissions received for different tasks for the two challenges, (c) Dice score comparison of different teams in three tasks of Medico 2020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

number of training samples prior to training their frameworks to improve the performance of their architecture. For the MedAI 2021 challenge, excluding two teams, all the others participated in both polyp and instrument segmentation tasks and used the same framework.

7.1. Medico 2020 challenge methods

Most of the methods reported in the Medico 2020 challenge focus on encoder-decoder architecture. Other networks used include GAN and 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 methods used the encoder-decoder architecture and out of 17, 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 35.17% in and mIoU and 29.89% in dice score. Similarly, compared to the R-CNN-inspired networks (team "IRIS-NSYSU"), the best approach (team "PRML2020GU") achieves an improvement of 28.63% in mIoU score and 21.91% dice score.

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. In the polyp segmentation task, most teams explored ensemble modeling, encoder-decoder networks, or a combination of both. 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 85.22%. Contrary to the Medico 2020 polyp segmentation challenge, where GAN-based methods failed to perform well, in this challenge, the second-leading team in polyp and instrument segmentation tasks (IIAI-CV&Med) adopted the GAN framework to generate segmented polyp masks. This approach is observed to be a close competitor to the best performer, with a slight difference of 0.90% in accuracy and 0.38% in the mIoU score. Also, this GAN-based approach performed better than the best method in terms of dice score and recall. The next close competitor in the list is a CNN and transformer based ensemble model.

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 in this task that they proposed for the polyp segmentation problem, the categorization of overall methods remains the same as that of the first task described above. The top rank is secured by "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. From the challenge, it can be observed that most of the team were reluctant to share their method (refer Table 10). Additionally, the quality of the code submitted by most of the team was was not satisfactory. Most of the participants did not put much effort into the readme file. Additionally, most of the 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, mTEC, CamAI,agaldran, and IIAI-CV&Med) was useful and understandable.

7.2.1. 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. Figure 6 and Figures 7 shows the example 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 system. 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. Similarly, the performance on the challenging images for polyps and instruments (see Figure 6, Figures 7 and 8) demonstrate that we can not solely rely on image-based metrics. Therefore, investigating the cause for misclassification for each sample in the dataset and failure analysis will be critical to focus for future research.

7.2.2. Trust, safety, and interpretability of methods

Integrating CADe or CADx in healthcare necessitates addressing factors such as trust, safety, and interpretability to ensure its adoption in clinical settings. The high variations 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. We addressed this issue by incorporating a transparency task in the MedAI2021 challenge.

The main aim of this task was to emphasize the need for interoperability, reproducibility and explainability in the AI-based submissions and shed light on the potential bias and wrong decisions that could have resulted from model and algorithmic bias. Our dataset contained polyp cases with varied shapes and sizes. We included samples with artifacts to make them closer to real clinical settings. Further, the inclusion of frames containing surgical instruments supported 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 LRP and a detailed ablation study in support of the predictions obtained. By promoting transparency and addressing potential biases, the challenge aimed to foster trust in the presented solution and ensure safety in adopting such methods in the clinic.

7.3. Transparency

One of the main aim of the MedAI 2021 challenge was to promote transparency in polyp and surgical instrument segmentation. In addition to the metric-based evaluation, we also evaluated submissions based on the transparency of their work. We encouraged participants to share how their models were trained, how datasets were used, and the insight on the interpretation of the model prediction. We allowed the participants to submit, considering the transparency and left them to decide what to deliver for the task. We gave some suggestions, such as providing rigorous analysis of the failing cases and a detailed GitHub repository with clear steps to run the code for reproducibility. We encouraged the participants to list package dependencies, provide the code for the architecture (guiding through building, compiling, and training of the proposed architecture), provide documentation for the code, 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 and statistical analysis and provide a detailed ablation study to show the contributing blocks in the network. Although heatmaps might not be the best choice, other alternatives, such as SHapley Additive exPlanations (SHAP), should be explored.

The submissions of the transparency task were evaluated using a more qualitative approach. A multi-disciplinary team accessed each submission and evaluated the transparency of the proposed solution. Then, we provided a report on the transparency of their submission and highlighted the details about which parts were good and which required more clarity to achieve transparency. Using this approach, we hope that the participant will be more responsible in making their submission in a public repository and including transparency parameters in their future work.

7.3.1. 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.

In our challenge, there was no online leaderboard due to the policy of Mediaeval. Therefore, we calculated the predictions submitted by each team manually. 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. 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, video sequences and addressing label biases will lead to more comprehensive and reliable evaluations of AI-based colonoscopy systems.

7.3.2. 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 organize a competition on multi-center datasets, including images from various modalities such as Flexible spectral Imaging Color Enhancement (FICE) and Blue laser imaging (BLI), and challenging cases. We will also consider metrics that weight speed, accuracy, and robustness for better objective assessments and introduce more distance-based metrics such as Hausdorff distance and normalized surface distance for improved fairness.

Furthermore, we will do a reliability test on the best-performing methods on easy, medium, and challenging real-world cases. 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. As a next step, we have organized Medical Visual Question Answering for GI Task - MEDVQA-GI challenge in 2023 [11].

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 more generalizable and transparent methods are needed to be

[11] https://www.imageclef.org/2023/medical/vqa

integrated into real-world clinical settings. During the "performance task" and "algorithm efficiency" tasks, we observed a trade-off between accuracy and inference time when tested across unseen still frames. For the instrument segmentation challenge, we observed that all most all teams performed relatively well as a segmenting instrument is easy compared to polyp segmentation. From the transparency task, we observed that more effort is required from the community to enhance the transparency of their work. 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 to achieve the goal of generalization in polyp segmentation. 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 could help build more clinically relevant and trustworthy systems.

Acknowledgment

D. Jha is supported by the NIH funding: R01-CA246704 and R01-CA240639. V. Sharma is supported by the INSPIRE fellowship (IF190362), DST, Govt. of India. D. Bhattacharya is funded partially 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 (IF170366).

Authors contribution

D. Jha conceptualized, initiated, and coordinated the work. He also led the data collection, curation, and annotation processes for Medico 2020 and evaluated the Medico 2020 Challenge. S. Hicks and M.A. Riegler initiated the MedAI 2021 Challenge, organized the data collection together with D. Jha and conducted all evaluations for the challenges, organized the reviews and coordinated with all the authors. V. Sharma analyzed the results and prepared most graphs for technical validation along with N. K. Tomar. She also wrote a part of the results and discussion. D. Banik, D. Bhattacharya and K. Roy wrote part of the introduction, related work, and participants' methods and provided subsequent feedback on the method's tables. M.A. Riegler and P. Halvorsen facilitated the data and organization for both competitions. Our gastroenterologists, T. de Lange and S. Parasa, reviewed the annotations and provided the required feedback during dataset preparation and evaluation. Challenge participants provided the method details for Medico 2020. All authors read the manuscript, provided substantial input, and agreed to the submission.

References

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

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

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., 2022. Where do we stand in AI for endoscopic image analysis? deciphering gaps and future directions. npj Digit. Medicine 5. doi:10.1038/s41746-022-00733-3.

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. Medical Image Analysis 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 multicentre polyp detection and segmentation dataset for generalisability assessment. Scientific Data doi:10.1038/s41597-023-01981-y.

Ali, S., Tomar, N.K., 2021. Iterative deep learning for improved segmentation of endoscopic images. Nordic Machine Intelligence 1, 38–40.

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

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

Ali *et al.*, S., 2020. Endoscopy disease detection challenge 2020. arXiv preprint arXiv:2003.03376 .

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: 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, Springer. 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, 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, 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: In Proceedings of the Lecture Notes in Computer Science, pp. 401–408.

Banik, D., Bhattacharjee, D., 2020. Deep conditional adversarial learning for polyp segmentation., in: 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. Nordic Machine Intelligence 1, 14–16.

Batchkala, G., Ali, S., 2020. Real-time polyp segmentation using u-net with iou loss., in: 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: 2017-11-20.

Bernal, J., Sánchez, J., Vilarino, F., 2012. Towards automatic polyp detection with a polyp appearance model. Pattern Recognition 45, 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 transactions on medical imaging 36, 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: In proceedings of the Comput. Assist. Radiol. Surg. (CARS), pp. –.

Bhattacharya, D., Betz, C., Eggert, D., Schlaefer, A., 2021a. Dual parallel reverse attention edge network : DPRA-EdgeNet. Nordic Machine Intelligence 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 .

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.

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. Nordic Machine Intelligence 1, 23–25.

Chou, Y., 2021. Automatic polyp and instrument segmentation in medai-2021. Nordic Machine Intelligence 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. Nordic Machine Intelligence 1, 50–52.

Fan, D.P., Ji, G.P., Cheng, M.M., Shao, L., 2021. Concealed object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 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.

Galdran, A., 2021. Polyp and surgical instrument segmentation with double encoder-decoder networks. Nordic Machine Intelligence 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. Nordic Machine Intelligence 1, 29–31.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: 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., 2021a. Medai: Transparency in medical image segmentation. Nordic Machine Intelligence 1, 1–4.

Hicks, S., Jha, D., Thambawita, V., Hammer, H., de Lange, T., Parasa, S., Riegler, M., Halvorsen, P., 2021b. Medico multimedia task at mediaeval 2021: Transparency in medical image segmentation, in: Proceedings of MediaEval 2021 CEUR Workshop, pp. 1–2.

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

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

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: 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: 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: In Proceedings of the MediaEval, pp. 1–3.

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

Keprate, A., Pandey, S., 2021. Kvasir-instruments and polyp segmentation using UNet. Nordic Machine Intelligence 1, 26–28.

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

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

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–44.

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: a Cancer Journ. Clinici. 58, 130–160.

Mahmud, T., Paul, B., Fattah, S.A., 2021. Polypsegnet: A modified encoder-decoder architecture for automated polyp segmentation from colonoscopy images. Computers in Biology and Medicine 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: 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. doi:`10.1109/ISBI.2017.7950555`.

Mirza, A., Rajak, R.K., 2021. Segmentation of polyp instruments using UNet based deep learning model. Nordic Machine Intelligence 1, 44–46.

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

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

Poudel, S., Lee, S.W., 2021. Explainable U-Net model forMedical image segmentation. Nordic Machine Intelligence 1, 41–43.

Rauniyar, S., Jha, V.K., Jha, R.K., Jha, D., Rauniyar, A., 2021. Improving polyp segmentation in colonoscopy using deep learning. Nordic Machine Intelligence 1, 35–37.

Riegler, M., et al., 2016. Eir—efficient computer aided diagnosis framework for gastrointestinal endoscopies, in: 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, Springer. pp. 234–241.

Shin, Y., Balasingham, I., 2018. Automatic polyp frame screening using patch based combined feature and dictionary learning. Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society 69, 33–42.

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

Siegel, R.L., Miller, K.D., Wagle, N.S., Jemal, A., 2023. Cancer statistics, 2023. CA: a cancer journal for clinicians 73, 17–48.

Somani, A., Singh, D., Prasad, D., Horsch, A., 2021. T-MIS: Transparency adaptation in medical image segmentation. Nordic Machine Intelligence 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 abs/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, Springer. 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 Transactions on Neural Networks and Learning Systems .

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: 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. Nordic Machine Intelligence 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. Computational Visual Media 8, 415–424.

Yeung, M., 2021. Attention U-Net ensemble for interpretable polyp and instrument segmentation. Nordic Machine Intelligence 1, 47–49.