

A comprehensive analysis of classification methods in gastrointestinal endoscopy imaging

Debesh Jha, Sharib Ali, Steven Hicks, Vajira Thambawita, Hanna Borgli, Pia H. Smedsrud, Thomas de Lange, Konstantin Pogorelov, Xiaowei Wang, Philipp Harzig, Minh-Triet Tran, Wenhua Meng, Trung-Hieu Hoang, Danielle Dias, Tobey H. Ko, Taruna Agrawal, Olga Ostroukhova, Zeshan Khan, Muhammad Atif Tahir, Yang Liu, Yuan Chang, Mathias Kirkerød, Dag Johansen, Mathias Lux, Håvard D. Johansen, Michael A. Riegler, Pål Halvorsen

Angaben zur Veröffentlichung / Publication details:

Jha, Debesh, Sharib Ali, Steven Hicks, Vajira Thambawita, Hanna Borgli, Pia H. Smedsrud, Thomas de Lange, et al. 2021. "A comprehensive analysis of classification methods in gastrointestinal endoscopy imaging." *Medical Image Analysis* 70: 102007. <https://doi.org/10.1016/j.media.2021.102007>.



ELSEVIER

Contents lists available at ScienceDirect

Medical Image Analysis

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

Challenge Report

A comprehensive analysis of classification methods in gastrointestinal endoscopy imaging

Debesh Jha^{a,b,*}, Sharib Ali^{c,v}, Steven Hicks^{a,d}, Vajira Thambawita^{a,d}, Hanna Borgli^{a,e}, Pia H. Smedsrud^{a,e,f}, Thomas de Lange^{a,f,g,h}, Konstantin Pogorelovⁱ, Xiaowei Wang^j, Philipp Harzig^k, Minh-Triet Tran^l, Wenhua Meng^m, Trung-Hieu Hoang^l, Danielle Diasⁿ, Tobey H. Ko^o, Taruna Agrawal^p, Olga Ostroukhova^q, Zeshan Khan^r, Muhammad Atif Tahir^r, Yang Liu^s, Yuan Chang^t, Mathias Kirkerødⁱ, Dag Johansen^b, Mathias Lux^u, Håvard D. Johansen^b, Michael A. Riegler^a, Pål Halvorsen^{a,d}

^a SimulaMet, Oslo, Norway^b UiT The Arctic University of Norway, Tromsø, Norway^c Department of Engineering Science, University of Oxford, Oxford, UK^d Oslo Metropolitan University, Oslo, Norway^e University of Oslo, Oslo, Norway^f Augere Medical AS, Oslo, Norway^g Sahlgrenska University Hospital, Molndal, Sweden^h Bærum Hospital, Vestre Viken, Oslo, Norwayⁱ Simula Research Laboratory, Oslo, Norway^j DeepBlue Technology, Shanghai, China^k University of Augsburg, Augsburg, Germany^l University of Science, VNU-HCM, Vietnam^m ZhengZhou University, ZhengZhou, Chinaⁿ University of Campinas, Brazil^o The University of Hong Kong, Hong Kong^p University of Southern California, Los Angeles, USA^q Research Institute of Multiprocessor Computation Systems, Russia^r School of Computer Science, National University of Computer and Emerging Sciences, Karachi Campus, Pakistan^s Hong Kong Baptist University, Hong Kong^t Beijing University of Posts and Telecom., China^u Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria^v Oxford NIHR Biomedical Research Centre, Oxford, UK

ARTICLE INFO

Article history:

Received 18 July 2020

Revised 20 January 2021

Accepted 16 February 2021

Available online 19 February 2021

Keywords:

Gastrointestinal endoscopy challenges

Artificial intelligence

Computer-aided detection and diagnosis

Medical imaging

Medico Task 2017

Medico Task 2018

BioMedia 2019 grand challenge

ABSTRACT

Gastrointestinal (GI) endoscopy has been an active field of research motivated by the large number of highly lethal GI cancers. Early GI cancer precursors are often missed during the endoscopic surveillance. The high missed rate of such abnormalities during endoscopy is thus a critical bottleneck. Lack of attentiveness due to tiring procedures, and requirement of training are few contributing factors. An automatic GI disease classification system can help reduce such risks by flagging suspicious frames and lesions. GI endoscopy consists of several multi-organ surveillance, therefore, there is need to develop methods that can generalize to various endoscopic findings. In this realm, we present a comprehensive analysis of the Medico GI challenges: Medical Multimedia Task at MediaEval 2017, Medico Multimedia Task at MediaEval 2018, and BioMedia ACM MM Grand Challenge 2019. These challenges are initiative to set-up a benchmark for different computer vision methods applied to the multi-class endoscopic images and promote to build new approaches that could reliably be used in clinics. We report the performance of 21 participating teams over a period of three consecutive years and provide a detailed analysis of the methods used by the participants, highlighting the challenges and shortcomings of the current approaches and dissect their credibility for the use in clinical settings. Our analysis revealed that the participants achieved an

* Corresponding author at: SimulaMet, Oslo, Norway.

E-mail address: debesh@simula.no (D. Jha).

improvement on maximum Mathew correlation coefficient (MCC) from 82.68% in 2017 to 93.98% in 2018 and 95.20% in 2019 challenges, and a significant increase in computational speed over consecutive years.

© 2021 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)

1. Introduction

Gastrointestinal (GI) cancers contribute to a large part of cancer-related deaths worldwide. Colorectal Cancer (CRC) ranks third in terms of cancer incidences and second in terms of mortality (Bray et al., 2018). The 5-year survival rates for colon cancer is 68% and that of stomach cancer is only up to 44% (Asplund et al., 2018). Detection and removal of pre-cancerous lesions provides the opportunity to prevent cancer and improve the survival rate to almost 100% (Levin et al., 2008). Early diagnosis and treatment can be facilitated by regular screening of patients at average risks before the disease becomes symptomatic. Screening of high-prevalence areas of infection, such as stomach and the large bowel (CRC), is particularly important to prevent cancer through early detection. The endoscopic procedures are the gold-standard for the diagnosis of GI abnormalities and cancers (Pogorelov et al., 2018b). The design of an automated Computer Aided Detection (CADE) and Computer Aided Diagnosis (CADx) system that can be integrated into the clinical workflow is essential (Suzuki, 2012), however, it requires careful evaluation of the built methods on a benchmark dataset. Additionally, these methods need to be assessed for their clinical applicability such as generalization in context to patient variability, and real-time processing capability.

This paper presents a comprehensive analysis of the results of Multimedia for Medicine Task (Medico) Task at MediaEval 2017 (Riegler et al., 2017) (Medico 2017), Medico Task at MediaEval 2018 (Pogorelov et al., 2018b) (Medico 2018), and the BioMedia Grand Challenge 2019 (Hicks et al., 2019a) at ACM Multimedia (BioMedia 2019). These challenges pose four clinically relevant rigorous tasks on GI endoscopic images and videos that include:

1. Algorithm performance evaluation through a frame level “classification task” (CADx) for multi-class GI tract findings
2. An “efficiency task” to evaluate the methods designed to achieve a trade-off between speed and accuracy
3. An “automated reporting task” on patient endoscopy video to analyse the efficacy of the built methods on videos
4. A “hardware task” to benchmark algorithms on the same system

1.1. Relevance of GI challenges

The Medico 2017 was the first challenge that utilizes a multi-class dataset (eight classes) for GI endoscopic image classification. The challenge was based on a multi-center, multi-modal, and multi-organ dataset that includes 8,000 endoscopic images collected, annotated, and verified by experienced endoscopists from four hospitals in Norway. With the success of the first challenge, we further collected and annotated 14,033 endoscopic images that were used at the Medico Task 2018 and the BioMedia Challenge 2019. The goal of organizing these challenges is to benchmark endoscopic image classification Machine Learning (ML) approaches with the specific focus on speed and robustness of the methods, which are essential for any clinical translation. These challenges have encouraged us to annotate and further release the dataset such as Kvasir-Capsule (Smedsrud et al., 2020), Kvasir-SEG (Jha et al., 2020) and Hyper-Kvasir dataset (Borgli, 2020).

1.2. Motivation of the study

The introduction of new imaging technology and progress in Artificial Intelligence (AI) system for detailed observation and interpretation to improve the diagnostic capability of medical images has motivated a wide range of multimedia researchers. GI endoscopy requires the integration of experienced endoscopists' knowledge to overcome the missed classification of diseases that subsequently ensure effective early disease detection. This could significantly reduce the miss-detection rate during an endoscopy examination. Therefore, there is a need for efficient CADx systems that can support endoscopists in real-time to locate clinically relevant markers and regions that are overlooked during the endoscopic procedure. A CADx system could reduce the workload of expert endoscopists during the examinations. Moreover, it could also aid inexperienced endoscopists for decision-making, which would significantly help to solve the problem of inter- and intra-observer variability in clinical endoscopies worldwide. Furthermore, the automatic reporting generated by AI methods can help reduce an endoscopist's workload, thereby improving their productivity and focus for critical cases.

Most designed computer vision methods and datasets focus on a limited set of lesions and very often limited to a specific organ. In practice, in particular to GI organs, routine surveillance can include multiple organs. For example, an upper GI surveillance can include oesophagus, stomach and first part of duodenum while lower GI can include small intestine to large intestine. Similarly, disease types can vary from organ to organ which will make it hard to detect all lesion occurrence at multiple GI locations in any surveillance. At times, both gastroscopy (upper GI endoscopy) and colonoscopy (lower GI endoscopy) are recommended for some patients. In these scenarios, the methods built with one specific organ or disease type is likely to have minimal clinical applicability and would not provide thorough clinical evaluation. We aimed to curate multi-organ gastroscopy datasets and challenge researchers to design methods for a comprehensive and challenging real-world dataset.

1.3. Task descriptions

Each challenge included four tasks. The teams were required to participate in the main “classification” task. However, the remaining three tasks were optional. Below, we briefly describe each task.

1.3.1. Classification task (required)

The goal of this task is to evaluate the classification methods for classifying anatomical landmarks (e.g., z-line, pylorus, cecum), pathological findings (esophagitis, polyps, ulcerative colitis), polyp removal cases (dyed and lifted polyps, dyed resection margins), and normal and regular cases (e.g., normal colon mucosa, stool, instrument etc.) inside the GI tract. This is to address the requirement for high classification accuracy needed for the development of computer-aided tools in the GI endoscopy. The teams are ranked based on their classification algorithm accuracy on 16 classes of GI dataset (refer Fig. 1).

The participants were instructed to design, train, and implement a classifier on the available training dataset. Subsequently,

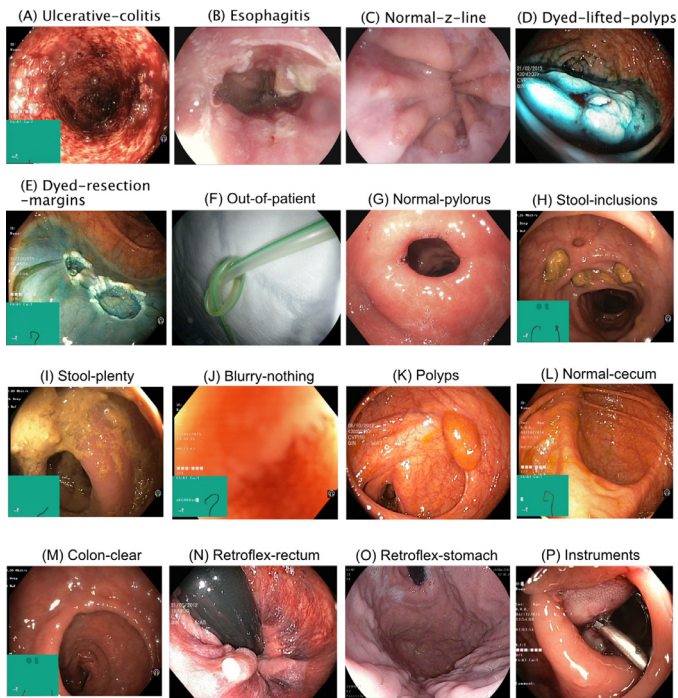


Fig. 1. Examples images from the 16 classes of Medico 2018 and BioMedia 2019 dataset.

the test dataset was released where the participants could test their model and predicted labels were sent to the organizers for evaluation. For the task submission, the participants were asked to create a “.csv” file. The “.csv” file should contain information about the image label prediction in a single line starting with the “name of the predicted file”, “predicted label” and “model’s confidence of the prediction”. Different standard metrics were used to evaluate these methods that are detailed in Section 4.

1.3.2. Efficiency task (optional)

Real-time performance of algorithms is required for clinical applicability of the methods. Analysis of the GI procedure in real-time can provide an opportunity for the experts to acquire feedback in real-time. However, fast inference models often compromise in accuracy. Thus, the goal for the efficiency task was to design the model that provides the best trade-off between speed and accuracy.

In most high-resolution GI endoscopes, the standard frame rates is over 45 Frames per second (FPS). Therefore, this task is aimed at building an efficient lightweight model that has the least latency in the inference time. For this task, the participants were required to capture processing time in millisecond for the inference of each test image on their system and report this time along with the GPU/CPU architecture to the organizers. The task submission procedure is quite similar to the classification task with only one difference, i.e., in efficiency task, the “processing time (in millisecond) for each image” must be included in the “.csv” file after the model’s confidence in the prediction line. The metrics for calculating “classification performance” in both classification and efficiency tasks are the same, however, with an additional FPS metric for the efficiency task. FPS was estimated from the average time reported by each team. A final ranking was computed by using a weighted score based classification accuracy metric and FPS (refer Section 4.2). It is to be noted that the participating team can submit the same or different models for classification and efficiency tasks for all 3 challenges.

1.3.3. Automatic report generation task (optional)

Among several responsibilities one of the crucial task of gastroenterologists is to generate endoscopic procedure reports after each endoscopy session. The World Endoscopy Organization (WEO) recommends using Minimal Standard for Reporting (MSR) and Minimal Standard Terminology (MST) for describing the endoscopic findings. This is often time-consuming and requires huge amount of administrative work (Woolhandler and Himmelstein, 2014). In addition, due to the inter operator variability, there is a large variation in such reporting which leads to inconsistent interpretation of findings and reporting mechanism (Aabakken et al., 2014). Intending to generate the standardized endoscopy reports automatically, we have offered this task in MediaEval and Biomedica challenges (Hicks et al., 2019b). A systematic and structured report preparation that describes the endoscopic findings can play a vital role in the development of an fast, automated and accurate reporting system. This will enable to accelerate the clinical procedures and minimize operator variability. The extensive use of GI endoscopy for diagnosis and treatment demands the requirement of standardized and user-friendly automated reporting systems at present.

In the presented task, the participants were required to automatically generate a text report of the endoscopic procedure that describes the detected findings according to the WEO protocol (Hicks et al., 2019b). The organizers provided the description (list of requirements) of what should be generated in the report. The assessment follows the list of requirements, and the reports were manually checked by two of the medical partners. We provided three videos for Medico 2017 and Medico 2018 for an automatic report generation task. For the BioMedia 2019, the number of videos was increased to six. The medical experts checked the practical usefulness of the report in terms of the medical domain (hospital).

1.3.4. Hardware task (optional)

In BioMedia 2019, we introduced the hardware task. In this challenge, the participants were asked to submit a docker image that included checkpoint of the trained model and test script for their submission. The requirement for this submission included the model trained in the classification task (Task 1). Each docker submission was then run on the test images by the organizers on NVIDIA GTX 1080 Ti GPU. This provided an opportunity to benchmark the built methods on the same hardware by an independent organizing team. Both the accuracy and speed were taken into account for the ranking of the methods for this task. The detailed information on the submission procedure can be found here.¹

2. Related work

While automatic classification, detection and segmentation of various GI lesions and anatomical landmarks have been recently studied, most of these focus on colonoscopy data that include polyp detection and segmentation (Poon et al., 2020; Lee et al., 2020; Song et al., 2020; Yamada et al., 2019; Akbari et al., 2018; Jha et al., 2021), intestinal cancer detection (Wan et al., 2019), stomach lesion detection (Krebs et al., 2020) and ulcerative colitis detection (Khorasani et al., 2020). However, the very nature of GI endoscopic procedures can range from esophageal to stomach to small and large intestine. Some recent works have taken this into account and have designed models for multi GI organ classification and detection (Thambawita et al., 2020; Iakovidis et al., 2018; Ali et al., 2020a; Chhedha et al., 2020; Poudel et al., 2020).

In addition to the research from the individual research group, recently, a few challenges have been initiated in the field of GI

¹ <https://github.com/stevenah/biomedica-2019-submission-evaluation>.

Table 1

Overview of GI endoscopy challenges. 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 task are summed and presented in '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
Medico 2017 (Riegler et al., 2017)	Entire GI	WL	Polyps, esophagitis, ulcerative colitis, z-line, pylorus, cecum, dyed polyp, dyed resection margins, stool	8,000 images	Open academic
GIANA 2017 (Bernal and Aymeric, 2017)	Colon	WL	Polyps & angiodysplasia	3462 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
Medico 2018 (Pogorelov et al., 2018b)	Entire GI	WL	Blurry-nothing, colon-clear, dyed-lifted-polyp, dyed-resection-margin, esophagitis, instrument, normal-cecum, normal-pylorus, normal z-line, out-of-patient, polyp, retroflex-rectum, retroflex-stomach, stool-inclusion, stool-plenty, ulcerative-colitis	14,033 images	Open academic
EAD 2019 (Ali et al., 2019)	Entire GI & bladder	NBI, WL, FL, WCE	Blur, bubbles, contrast, imaging artefact, saturation, specularity, instrument	2,192 images	Open academic
BioMedia 2019 (Hicks et al., 2019a)	Entire GI	WL	Blurry-nothing, colon-clear, dyed-lifted-polyp, dyed-resection-margin, esophagitis, instrument, normal-cecum, normal-pylorus, normal Z-line, out-of-patient, polyp, retroflex-rectum, retroflex-stomach, stool-inclusion, stool-plenty, ulcerative-colitis	14,033 images	Open academic
EAD 2020 (Ali et al., 2019)	Entire GI & bladder	NBI, WL, FL, WCE	Blur, bubbles, blood, contrast, imaging artefact, saturation, specularity, instrument	2,916 images	Open academic
EDD 2020 (Ali et al., 2020a)	Entire GI	NBI, WL	Barrett's esophagus, high-grade dysplasia, suspicious (low-grade), polyp, cancer	386 images	Open academic

endoscopy that uses either still images or both still images and videos. Several ML based methods have been proposed on these endoscopy challenge datasets. However, most of the endoscopy challenges focused only on colorectal polyp and cancer localization, detection and segmentation (Bernal et al., 2017). Additionally, the used datasets are either scarce (only 386 image frames were released for 5 disease classes in (Ali et al., 2020a)) or have not been benchmarked on the same dataset for different challenges over time (for example, EndoVis2015 challenge on Early Barrett's cancer detection²). As a result, the conclusions drawn from these challenges are not comparable from one challenge to the other. In addition, many such datasets are not publicly available, making it difficult for further analysis and comparison (Wang et al., 2018; Bernal et al., 2017; Bernal and Aymeric, 2017; Angermann et al., 2017; Bernal et al., 2018).

To address the need of benchmarking methods on the same dataset, different international challenges have been organized. Polyp detection challenge on colonoscopy videos was organized by (Bernal et al., 2017) at IEEE International Symposium on Biomedical Imaging (ISBI), and Medical Image and Computing and Computer Assisted Intervention (MICCAI) conference in 2015³. The organizers released 808 still images and 38 videos. A comprehensive study of the results on this dataset from 8 different participating teams concluded that there was still a potential for improvement (Bernal et al., 2017) in the polyp detection task.

Our team organized the first MediaEval Medico challenge in 2017 (Riegler et al., 2017) that aimed to compare baseline for computer vision classification methods. With over 8,000 annotated video frames consisting of multiple endoscopic findings for the entire GI tract, including pre- and post-treatment patients and eight different categories, we established a first comprehensive dataset

that mimics various endoscopic procedures as a whole. Bernal et al. launched GIANA challenge (2017 and 2018)⁴ where they broaden the scope of their past challenge by including additional tasks such as detection of lesions in Wireless Capsule Endoscopy (WCE), polyp detection, and polyp segmentation task. However, their task assignment was still focused on colonoscopy data only. To further quantify and improve baseline methods and promote algorithm development, we organized a consecutive Medico task 2018 challenge (Pogorelov et al., 2018b). This challenge had an extended dataset of 14,033 GI endoscopy frames and aimed at classifying 16 class categories for multiple GI endoscopy organs. For better longitudinal analysis and method benchmarking, we used the same dataset to organize a recent BioMedia challenge 2019 (Hicks et al., 2019a). Another challenge in 2019 dedicated for artefact detection and segmentation in endoscopy (EAD2019, (Ali et al., 2020b)) released more than 2,192 still endoscopy frames that included multi-organ and multi-center data and aimed at classifying 6 different artefact classes⁵. A comprehensive analysis of the methods evaluated on EAD2019 challenge revealed the need for more quantifiable metrics and the requirement of clinical applicability tests with current Deep Learning (DL) approaches. The same team launched EndoCV2020 challenge⁶ this year with an additional sub-challenge on "Endoscopy disease detection (EDD2020)". Even though this sub-challenge incorporated multi-organ and multi-modal endoscopy data, the released dataset has only 386 annotated frames and was included only 5 class categories (Ali et al., 2020a). Table 1 presents the overview of GI challenges held and imaging modalities used over past 5 years.

In summary, there is still a need for comprehensive algorithm benchmarking datasets in GI endoscopy, especially due to the var-

² <https://endovissub-barrett.grand-challenge.org>.

³ <https://polyp.grand-challenge.org>.

⁴ <https://giana.grand-challenge.org/>.

⁵ <https://ead2019.grand-challenge.org/>.

⁶ <https://endocv.grand-challenge.org/>.

ied nature of endoscopic findings and abnormalities. Mainly, as most current datasets are limited by sample size, single modality and single organ data, methods built on them cannot be applied to wider endoscopy settings and GI organs. Additionally, most of these datasets are not easily accessible as they require special permissions and email correspondences prior to their use. Such a practice could discourage computational scientists to build and validate their method on these benchmarks.

Motivated by the success of DL techniques in other medical imaging domains, we initiated collaborations with four hospitals in Norway to collect, curate, annotate, and publish open-access datasets. Medico 2017, Medico 2018, and Biomedica 2019 are few attempts to fulfill the challenges related to method comparison for the multi-class GI endoscopy and to address the lack of availability of publicly available datasets. In this paper, we detail on our three challenge datasets from 2017 to 2019 under “MediaEval Medico GI Endoscopy Challenge Dataset” and provide a comprehensive analysis of their outcomes.

3. Medico GI-endoscopy challenge datasets

3.1. Medico 2017

The dataset for Medico 2017 consists of both images and videos. The “Kvasir” dataset (Pogorelov et al., 2017b) is a multi-class dataset consisting of 1,000 images per class with a total of 8,000 images altogether for eight different classes. These classes consist of pathological findings (esophagitis, polyps, ulcerative colitis), anatomical landmarks (z-line, pylorus, cecum), and normal and regular findings (normal colon mucosa, stool), and polyp removal (post-treatment) cases (dyed and lifted polyps, dyed resection margins).

In the Medico 2017, the entire dataset was divided into training and test dataset. The training and test set consists of 4,000 images each. The participants were provided with pre-split train-test categories for all 8 classes with 500 images per class in each split. However, the labels for test set were not provided. The image size varied from 720×576 up to 1920×1072 pixels taken from a high-resolution Olympus endoscope. Some of the images in the dataset contained a green box in the left-bottom corner of the image showing the position of the scope inside the bowel (Pogorelov et al., 2017b) (see Fig. 1). In addition, we provided a separate folder with the extracted visual global features (GFs) for each of the images that included global features such as Joint Composite Descriptor (JCD), Tamura, ColorLayout (CL), edge histogram (EH), AutoColorCorrelogram, and Pyramid Histogram of Oriented Gradients (PHOG) (Lux and Chatzichristofis, 2008).

Three videos containing polyps, bleeding, and Z-line were provided for automatic report generation task. The videos contain the diseases or findings included in the Kvasir dataset. The aim was to use the video cases to generate automated text reports that described the findings in all three videos.

3.2. Medico 2018

The Medico 2018 dataset is the combination of the Kvasir dataset (Pogorelov et al., 2017b) and Nerthus dataset (Pogorelov et al., 2017c). The Medico 2018 dataset consists of 16 classes. Fig. 1 shows the sample images used in Medico 2018 and BioMedia 2019. Initially, the training dataset that consisted of 5,293 images was released. The participants were asked to develop the algorithms based on this dataset. Later on, 8,740 test images were released. The Medico challenge 2018 dataset contains the images from the previous challenge and 6,033 additional images and eight new classes. The additional classes used in the task are colon-clear, stool-inclusions, stool-plenty, blurry-nothing,

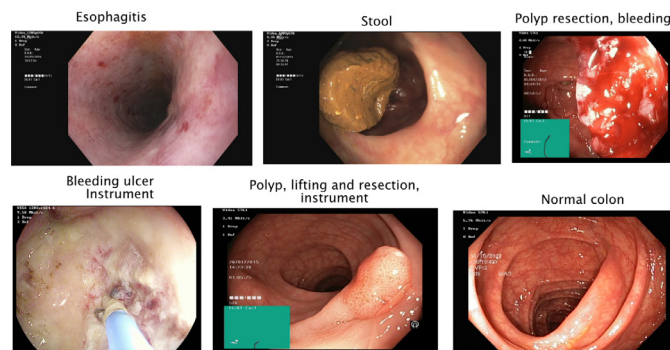


Fig. 2. Example of extracted frame from each of the 6 videos provided to the participants for for automatic report generation task.

out-of-patient, and the pre-, while and therapeutic findings such as dyed-lifted-polyps, dyed-resection-margins, and the instrument class (Pogorelov et al., 2018b). Both the training and test datasets were imbalanced (refer Fig. 3) due to increased class numbers and very few samples for some classes, for example, only four images for out-of-patient class while 613 samples were present for the polyp class. In addition to this, similar to the 2017 challenge, we provided the same three videos for the text-report generation task.

3.3. BioMedia 2019

The BioMedia 2019 consisted of the same two types of datasets as proposed in the 2018 challenge. However, in addition to the classification task, we increased the total number of videos to six for the report generation tasks, we also included a hardware task for fair comparison of submissions. The details on the image dataset is the same as for 2018 presented above and in summary Fig. 3. The video dataset consisted of six videos ranging from 720×576 to 1920×1072 pixels. The length of the video varies from 51 s up to 5 min and 11 s. A sample of an extracted video frame from each video dataset for the automatic report generation task is shown in Fig. 2. The tasks on the videos were similar to those of the image frames. The details about the video dataset is presented in Table 2. More details about the dataset can be found in our task overview paper (Hicks et al., 2019a).

The participants had a total of three months for submission in all of the challenges. The test datasets were provided one month after the release of the training dataset. The challenge datasets can be found here (Pogorelov et al., 2017b; 2017c).

4. Evaluation metrics

Standard evaluation metrics used to quantify image classification methods such as recall, precision, F1-score and accuracy (Eq. (1)–(4)) were used for all three challenges. To determine the final score and rank of the participating teams, we used Matthews correlation coefficient (MCC) (Matthews, 1975), which provides a

Table 2

An overview of video dataset with expected findings, length, and resolution provided for automatic report generation (Hicks et al., 2019a).

Expected Findings	Length	Resolution
Esophagitis	00:51	1920×1072
Stool	00:02	1920×1072
Polyp resection, bleeding	02:00	720×576
Bleeding ulcer, instrument	01:08	1280×1024
Polyp, lifting and resection, instrument	05:11	720×576
Normal colon	00:57	720×576

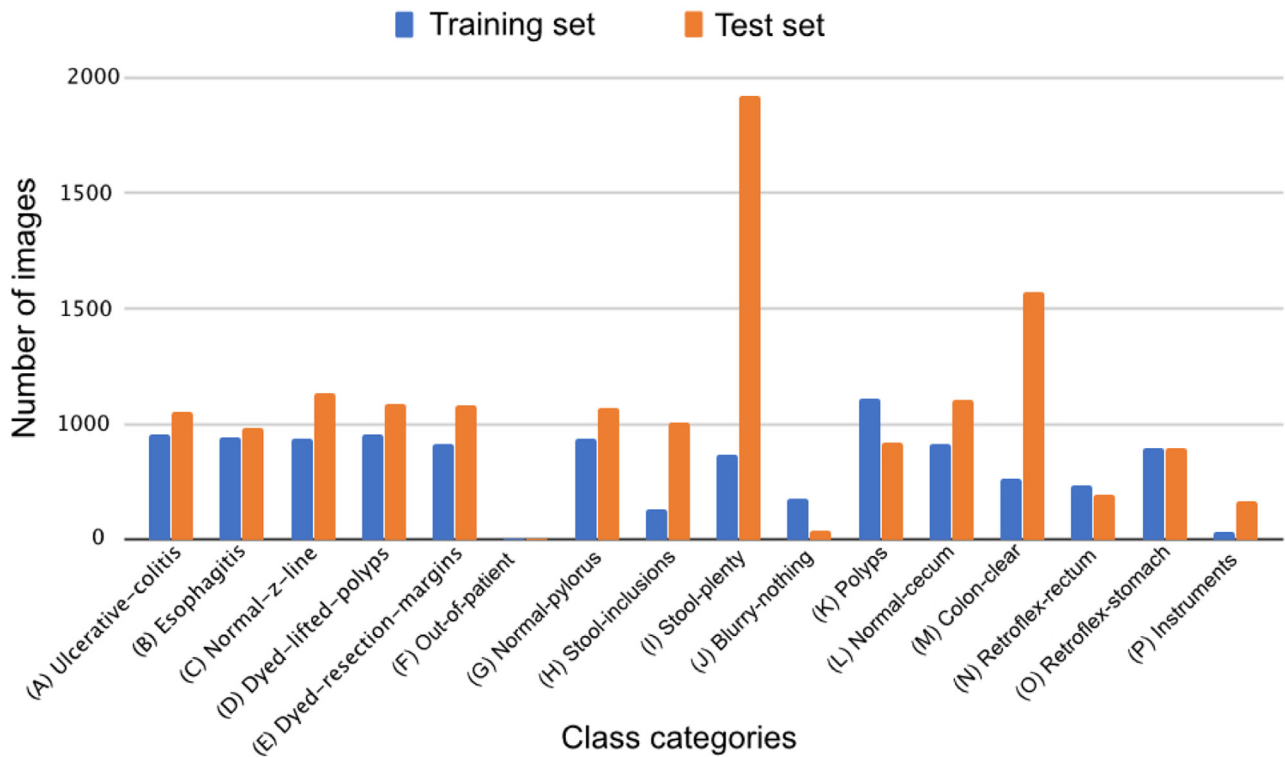


Fig. 3. Summary of the Medico 2018 and BioMedia 2019 dataset.

reliable statistical measure and can handle class imbalance problems in datasets. MCC can be computed from the confusion matrix of true and false positives and negatives (see Eq. (5)).

$$\text{Recall (REC)} = \frac{TP}{TP + FN} \quad (1)$$

$$\text{Specificity (SPEC)} = \frac{TN}{TN + FP} \quad (2)$$

$$\text{Precision (PREC)} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{Accuracy (ACC)} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$\text{MCC} = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{n}}, \quad (5)$$

where $n = (TP + FN)(TN + FP)(TP + FP)(TN + FN)$

$$F_1\text{-score (F1)} = 2 \times \frac{(p \times r)}{p + r} \quad (6)$$

$$\text{Frame Per Second (FPS)} = \frac{1}{\text{sec/frame}} \quad (7)$$

In the above equations, p is precision, r is recall, and TP , FP , TN , FN represent true positives, false positives, true negatives, and false negatives, respectively, for the classification outputs. If the MCC values are equal for more than one team, the efficiency task criteria was considered where we considered processing speed of the algorithms, and the amount of the training data used to obtain the best result (Pogorelov et al., 2018b). The participants were allowed to submit the results up to five runs in total. The more detailed de-

scriptions of the challenge can be found on their respective challenge webpages.^{7,8,9}

4.1. Metrics for classification task

The classification task aimed at achieving higher accuracy for the multi-class classification task of the GI endoscopy findings and diseases. To perform a complete and thorough evaluation of this task, we provided all standard classification metrics, including sensitivity, specificity, precision, accuracy, and F1-score. However, due to the class imbalance in some classes, MCC was used for ranking the participants.

4.2. Metrics for efficiency task

The goal of the efficient classification task is to score the participants based on the test time recorded for their algorithm. The main motivation behind this task is to identify the clinical usability of these methods as speed is one of the required criteria. For this task, we used the FPS estimation of each method on the provided image dataset.

The same evaluation metrics ‘‘MCC,’’ was used. The ‘‘speed’’ was calculated based on the average time the algorithm takes to classify the single image in milliseconds. The submissions were ranked on the basis of the combination of ‘‘classification performance’’ and ‘‘speed’’. For balancing the two requirements, a threshold of 85% was set on specificity and sensitivity (Pogorelov et al., 2018a) that is a standard threshold for an automatic detection system for colonoscopies in industry. Only those submissions that reached or surpassed this threshold was considered as a valid submission. If more than one teams have the same time, higher sensitivity and

⁷ <http://www.multimediaeval.org/mediaeval2017/medico/>.

⁸ <http://www.multimediaeval.org/mediaeval2018/medico/>.

⁹ <https://github.com/kelkalot/biomedica-2019>.

Table 3

Summary information of participating teams in Medico 2017, Medico 2018, and the BioMedia 2019, 'X' = Team participated, '-' = No participation.

Chal.	Team Name	Task 1	Task 2	Task 3	Task 4
2017	HKBU	X	X	-	-
	ITEC-AAU	X	X	-	-
	SLC-UMD	X	X	-	-
	FAST-NU-DS	X	X	-	-
	SIMULA	X	X	-	-
2018	LesCats	X	X	-	-
	RUNE	X	-	-	-
	UMM-SIM	X	-	-	-
	ParaNoMundo	X	X	-	-
	AAUITEC	X	-	-	-
	SIMULA	X	X	-	-
	FAST-NU-DS	X	X	-	-
	NOAT	X	-	-	-
	HKBU	X	X	-	-
	S@M	X	-	-	-
2019	HCMUS	X	X	-	-
	uniaugsburg	X	X	X	X
	CIISR	X	-	X	X
	DeepBlueAI	X	X	-	-
	Mcdull	X	-	-	-
	HCMUS	X	X	-	-

specificity were taken as the better performing one (Hicks et al., 2019a).

4.3. Automatic report generation task

Teams participating in this task were asked to provide the generated text report describing the detection results on the provided video dataset. Two medical experts ranked these automatically generated reports. To aid the senior gastroenterologists in their assessment, they were provided with five team ranking protocols. These included:

1. Does the provided report has clarity and pass the confidence from a clinical point of view?
2. Limitations of the generated report (if any)
3. How useful would the report be in the clinic?
4. Did the teams incorporated any useful suggestions for improvement or additions?
5. Did the teams provide any useful findings as other comments in their report?

5. Participating methods

Table 3 summarizes the participation of each team with 'X' denoting the information about the participants who participated in the particular task for 2017, 2018, and 2019 challenges and the tasks posed in the consecutive years. A wide range of methods were developed in each challenge for which a summary is provided in Table 4.

5.1. Methods used in Medico 2017

In this challenge, there were 5 participating teams that included the organizers. However, the organizers submissions were not considered in the ranking of the challenge. Below we briefly describe method of each team.

HKBU: Team HKBU (Liu et al., 2017) designed a two-stage learning strategy for the classification of GI endoscopy images. In the first stage, they used a manifold learning method called Bidirectional Marginal Fisher Analysis (BMFA) to project the original dataset to a low dimensional space with the key discriminant information being well preserved. In the second stage, a multi-class Support Vector Machine (SVM) was used for the classification.

ITEC-AAU: The method proposed by team ITEC-AAU (Petscharnig et al., 2017) used an Inception-like Convolutional Neural Network (CNN) architecture with a GoogleNet (Szegedy et al., 2015) backbone. Data augmentation with fixed-cropping was also used on both training and test datasets. This step provided an advantage for obtaining low inference time.

SCL-UMD: Transfer learning-based feature extraction technique was used by team SCL-UMD (Agrawal et al., 2017). The team used pre-trained CNN models that included VGGNet (Simonyan and Zisserman, 2014) and Inception-v3 trained on ImageNet (Deng et al., 2009) dataset and fine-tune them on the provided training data. The obtained features were combined with the features provided by the organizers. Their best model was the combination of three features, namely, baseline features provided by organizers, Inception-V3 features, and VGGNet features. A multi-class SVM classifier was trained on these extracted features. The hyperparameter of SVM was tuned using 5-fold cross-validation in the training dataset. The optimal kernel choice for SVM was a linear kernel in their case.

FAST-NU-DS: Team FAST-NU-DS (Naqvi et al., 2017) used an ensemble of texture features for classification of GI endoscopic images. The main motivation of their approach was to combine information from various local features that included Haralick texture features and local binary patterns for successful classification. These features were selected at the training stage using a 10-fold cross-validation strategy. A Logistic Regression (LR) classifier was used to train the model. The outputs of the model were combined using a majority voting strategy.

SIMULA: Team SIMULA (Pogorelov et al., 2017a) approached the task by utilizing both GFs and CNNs. For GFs based approach, 6 GF were experimented with a random tree, Random Forest (RF), and Logistic Model Tree (LMT) classifiers from the WEKA software (Hall et al., 2009). The best classification results were obtained for LMT. Similarly, for the CNN based approach, the team experimented with the Inception-v3 and ResNet-50 (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009). Their best performing approach was using extracted features from fine-tuned ResNet-50 architecture pre-trained on ImageNet and LMT classifier.

5.2. Methods used in Medico 2018

10 teams participated in the Medico 2018. Additionally, there was 1 submission from the organizers team, however, this was not considered in the ranking. Below we briefly present methods for each participating team.

FAST-NU-DS: Team FAST-NU-DS (Khan and Tahir, 2018) investigated various combinations of Haralick texture features, LIRE features, and Deep features. Deep features were extracted using VGG19 pre-trained on the ImageNet dataset. Various models were then trained using an ensemble of classifiers, including LR, RF, and extremely random trees. Each model was trained using 10-fold cross-validation of the training data and with various combinations of features. On test data, the best results were obtained from the combination of Haralick and LIRE features.

HCMUS: Team HCMUS (Hoang et al., 2018) used a combination of residual neural network and Faster R-CNN model (both pre-trained on ImageNet) for classification of the GI endoscopic images. Their approach included data preparation, augmentation, and classification. As a data preparation step, regions containing symptoms of diseases were annotated to train the abnormality localization module. Additionally, some labels of the development dataset were cleaned, and dataset augmentation strategies were applied to balance the number of images between different classes. Their best result was obtained by ResNet-101 and Faster R-CNN trained on the re-labeled training dataset combined with their augmented instrument dataset. This is because the instrument class has rela-

Table 4
Summary of the participating teams algorithm for Medico 2017, and Medico 2018, and the BioMedia 2019. Here, ED = Eigen decomposition, GD = Gradient Descent, SMO = Sequential minimal optimization, BMFA= Bidirectional Marginal Fisher Analysis, SGD = Stochastic gradient descent.

Challenge	Team Name	Algorithm	Backbone	Nature	Choice Basis	Data Aug.	Loss function	Optimizer	GPU/CPU
	HKBU (Liu et al., 2017)	BMFA + ν SVM	N/A	Cascade	Context-speed	No	Hindge loss	ED SMO	Intel Quad-Core i7
Medico 2017	ITEC-AAU (Petscharnig et al., 2017)	CNN (Pre-trained Network)	GoogleNet	General	Speed	Yes	-	-	-
	SLC-UMD (Agrawal et al., 2017)	CNN (Pre-trained Network)	Inception-v3, VGGNet	Ensemble	Accuracy	No	Hindge loss	SGD	N/A
	FAST-NU-DS (Naqvi et al., 2017)	Texture feature + LIRE features + LRC	N/A	Ensemble	Accuracy	Yes	Cross-entropy	N/A	Intel Core i5-10600
	SIMULA (Pogorelov et al., 2017a)	ResNet + LMT	Inception-v3, ResNet-50	Combined feature	Accuracy	No	Cross-entropy	-	GTX 1080Ti
	HCMUS (Hoang et al., 2018)	ResNet + Faster R-CNN	ResNet-101	Feature pyramid	Accuracy	Yes	Cross-entropy	Adam	Tesla K80
Medico 2018	ParaNoMundo (Dias and Dias, 2018a)	DenseNet	DenseNet-201	General	Accuracy	No	Cross-entropy	SGD	N/A
	UMM-SIM (Kirkerød et al., 2018)	GAN + InceptionResNet-v2	InceptionResNet-v2	cascade	Accuracy	No	Cross-entropy	Adam	GTX 1080Ti
	S@M (Thambawita et al., 2018)	ResNet + DenseNet + MLP	ResNet-152, DenseNet-161	Ensemble	Accuracy	Yes	Cross-entropy	SGD	GTX 1080Ti
	AAUITEC (Taschwer et al., 2018)	GF + GoogleNet+ L-SVM	GoogleNet	Combined feature	Accuracy	No	-	-	-
	LesCats (Hicks et al., 2018)	DenseNet	DenseNet-169	Ensemble	Accuracy	Yes	Cross-entropy	Nadam	GTX 1080Ti
	FAST-NU-DS (Khan and Tahir, 2018)	GF + Majority voting(LR, RF, ETC)	N/A	Ensemble	Accuracy-Speed	No	-	GD	Tesla K80
	NOAT (Steiner et al., 2018)	Global feature + CNN	N/A	Combined feature	Speed	No	Cosine distance	-	-
	RUNE (Borgli et al., 2018)	DenseNet	DenseNet-169	General	Accuracy	Yes	Cross-entropy	SGD	GTX 1080Ti
	SIMULA (Ostroukhova et al., 2018)	InceptionNet	Inception-v3	General	Accuracy-Speed	Yes	Cross-entropy	RMSprop	GTX 1080Ti
	HKBU (Ko et al., 2018)	WDE + CS-NN	N/A	Cascade	Context	No	-	ED	Intel Quad-Core i7
Biomedica 2019	CIISR (Meng et al., 2019)	ResNet + Softmax	ResNet-50	General	Accuracy-speed	Yes	Cross-entropy	Adam	Tesla P4
	Mcdull (Chang et al., 2019)	ResNet + SE-ReNeXt + Attention-Inceptionv3	ResNet-34	Feature pyramid	Accuracy-speed	Yes	Focal loss, Cross-entropy	Adam	Tesla P100
	uniaugsburg (Harzig et al., 2019)	MobileNet	MobileNet-V2, DenseNet-121	General	Accuracy-Speed	Yes	Cross-entropy	Adam	TITAN XP
	HCMUS (Hoang et al., 2019)	ResNet+ Faster R-CNN	ResNet-101	Feature pyramid	Accuracy	Yes	Cross-entropy	Adam	GTX 1080Ti
	DeepBlue (Luo et al., 2019)	10 pre-trained CNN from ImageNet	SE_ResNeXt50, SE_ResNeXt101, SENet154, DenseNet201, DenseNet161, ResNet152, ResNet101, ResNet34, InceptionV4 and Inception-ResNetV2	Ensemble	Accuracy	Yes	Cross-entropy	SGD	RTX 2080 Ti

tively fewer samples compared to other classes. Their team won the Medico 2018 challenge for the classification task.

ParaNoMundo: Team ParaNoMundo (Dias and Dias, 2018b) evaluated 10 CNN architectures all of which were pre-trained on ImageNet. Their best model included DenseNet-201 (Huang et al., 2017) and ResNet. On the test dataset, DenseNet-201 outperformed ResNet by a small margin on F1-score and MCC metrics. However, the ResNet model was two times faster than DenseNet-201.

UMM-SIM: Team UMM-SIM (Kirkerød et al., 2018) used an unsupervised context-aware Conditional Generative Adversarial Network (CGAN) (Denton et al., 2016; Goodfellow et al., 2014) as data pre-processing step to remove the green corners of the image marked by “ScopeGuide” with the probe marking (see some image samples from Figs. 1 and 2). They used CGAN to regenerate the areas covered by the green area to help model perform better on the clean dataset. For the image classification task, they used an Inception-ResNet-v2 (Szegedy et al., 2017) with softmax classifier.

AAUITEC: For classifying GI disease and findings, team AAUITEC (Taschwer et al., 2018) used early fusion and late fusion strategies. In the early fusion strategy, they combined GFs and CNN-based features, and for the late fusion strategy, they applied soft voting for combining the output of multiple classifiers. Their approach that resulted in their top score out of five runs was the combination of GFs extracted using LIRE (Lux and Chatzichristofis, 2008) and GoogleNet features. With the combined features, linear SVM performed best compared to KSVM, RF, RF-KSVM-LR, and the LR classifiers.

NOAT: Team NOAT (Steiner et al., 2018) classified the GI images in three steps. First, pre-trained DL models were used for the extraction of features. Then, LIRE was used for indexing these generated features. In the final step, the team searched for the index of the most similar images using a cosine distance function. Out of the four submitted runs, they achieved the best results with the integer features using bit sampling and a hashing technique.

S@M: Team S@M (Thambawita et al., 2018) made a comprehensive evaluation by using a ML-based approach to DL based solution for the multi-class classification of GI tract findings. For the ML-based solution, the extracted GFs were passed through a simple logistic classifier and a LMT classifier. They performed an extensive study by using different pre-trained models and combinations of the pre-trained models. Their best model was the combination of ResNet-152 and DenseNet-161 along with the additional multi-layer perceptron for the classification of the provided 16 classes. Their team held the second position in the classification task.

LesCats: Team LesCats (Hicks et al., 2018) hypothesized that pre-training the models with a medical dataset could outperform models pre-trained on ImageNet (Deng et al., 2009) for the provided dataset. Out of the submitted models, they found that a DenseNet-169 pre-trained on ImageNet performed best. They found that the large and diverse datasets were better to pre-train on rather than smaller datasets, even if they were similar to the target domain.

RUNE: Team RUNE (Borgli et al., 2018) approached the task with a specific focus on automatic hyperparameter optimization and data pre-processing. They used Bayesian optimization for optimizing their pre-trained CNN model. As a pre-processing step, they added extra images to the “out-of-patient” class and also performed a split on the “esophagitis” class into lower and upper. The classes, “esophagitis” and “z-line”, would often be confused, so this split was meant to improve their classification performance by making the image distribution space smaller for the esophagitis class. They achieved the best results with DenseNet-169, standard gradient descent optimizer, and a delimiting layer of 0.

SIMULA: Team SIMULA (Ostroukhova et al., 2018) presented a method proposed by the organizer team. Their main motivation to approach the task was to provide a baseline for method compari-

son. They used the Inception-v3 model pre-trained on ImageNet. To address the imbalanced dataset, they added randomly duplicated images to the classes with fewer image samples. Their best model was the one trained using the balanced training set and a non-prioritized classifier.

HKBU: Team HKBU (Ko et al., 2018) approached the task with a particular focus on dimensionality reduction. They used a two-stage learning strategy, which first performs the weighted discriminant embedding (WDE) to project the original data to a low-dimensional feature subspace and then utilizes the cost-sensitive nearest neighbor (CS-NN) method in the learned subspace for disease prediction.

5.3. Methods used in BioMedia 2019

There were five participating teams in the BioMedia 2019. The methods of each participating team are summarized below.

CIISR: Team CIISR (Meng et al., 2019) participated in the classification task for which they used data enhancement techniques to address the class imbalance problem. Augmentation techniques, such as flipping, rotation, cropping, and color change were used. Their best performing model used ResNet-50 that was pre-trained on ImageNet with a softmax classifier.

Mcdull: The core idea of team Mcdull (Chang et al., 2019) was learning different feature representations for multi-label images using CNN-based models. The team only participated in the classification task. They experimented with a variety of different models, including ResNet-34 (He et al., 2016), SE-ReNeXt (Xie et al., 2017) and attention-Inception-v3 (Szegedy et al., 2016), but found that attention-Inception-v3 achieved the best performance. All models were trained using multi-epoch fusion and adaptive thresholding techniques with an automatic data augmentation scheme.

Uniaugsburg: The main objective of the team Uniaugsburg (Harzig et al., 2019) was to design an improved approach for endoscopic image classification that could potentially run on mobile phones and also generate reports based on the findings of the algorithm. They participated in all four tasks. For the classification task, DenseNet121 (Huang et al., 2017) achieved the best result. For the efficiency task, the team proposed MobileNet-V2 (Sandler et al., 2018) with a width multiplier of 1.0 for an efficient detection model. For the automatic report generation task, they used the same model that was used for the classification task. However, they extended this model with class activation maps (CAM) (Zhou et al., 2016) to detect the spatial location (one of top-left, top-right, bottom-left, bottom-right, or center) for the classification. In combination with a per-frame classification, they were able to generate a report consisting of three clinically relevant sections (main findings, brief summary, and a detailed summary).

HCMUS: Team HCMUS (Hoang et al., 2019) used stacked model of ResNet-101 (He et al., 2016) pre-trained on the ImageNet (Deng et al., 2009), and a Faster R-CNN (Ren et al., 2015). For the classes having a limited number of training samples, such as instruments class, they cropped the area covered by the disease or instruments and their edges. Consequently, these patches were put randomly with affine transformed patches on top of various images from the other classes. Such data augmentation techniques enhanced their performance for both the classification and localization of class categories. In order to reduce the confusion between various types of abnormalities that appeared in the same image, the team used multiple classifiers, introducing a multi-task learning approach. An ablation study revealed the effectiveness of this technique and the data augmentation strategy.

DeepBlue: Team DeepBlue (Luo et al., 2019) used 10-fold cross-validation to train ten different models pre-trained on the ImageNet dataset leading to ten sub-models. They utilized the data augmentation technique to overcome the class imbalance in the

Table 5

Team performances for 2017 Medico Classification task.

Reference	TP	TN	FP	FN	REC	SPEC	PREC	ACC	MCC	F1
HKBU (Liu et al., 2017)	2811	26811	1189	1189	0.7027	0.9575	0.7027	0.9256	0.6626	0.7027
FAST-NU-DS (Naqvi et al., 2017)	3066	27066	934	934	0.7665	0.9666	0.7665	0.9416	0.7331	0.7665
ITEC-AAU (Petscharnig et al., 2017)	3021	27021	979	979	0.7552	0.9650	0.7552	0.9388	0.7202	0.7552
SIMULA (Pogorelov et al., 2017a)	-	-	-	-	0.8260	0.9750	0.8290	0.9570	0.8020	0.8260
SLC-UMD (Agrawal et al., 2017)	3390	27390	610	610	0.8475	0.9782	0.8475	0.9618	0.8257	0.8475

Table 6

Team performance for Medico Efficiency task 2017. Method design is based on the trade-off between the accuracy and speed of each algorithm.

Reference	TP	TN	FP	FN	REC	SPEC	PREC	ACC	MCC	F1	FPS
HKBU (Liu et al., 2017)	2908	26908	1092	1092	0.7270	0.9610	0.7270	0.9317	0.6946	0.7270	2.2
FAST-NU-DS (Naqvi et al., 2017)	2981	26981	1019	1019	0.7452	0.9636	0.7452	0.9363	0.7114	0.7452	2.3
ITEC-AAU (Petscharnig et al., 2017)	3021	27021	979	979	0.7552	0.9650	0.7552	0.9388	0.7202	0.7552	1.4
SIMULA (Pogorelov et al., 2017a)	3248	27248	752	752	0.8120	0.9731	0.9530	0.7851	0.7856	0.7851	46.0
SLC-UMD (Agrawal et al., 2017)	3390	27390	610	610	0.8475	0.9782	0.8475	0.9618	0.8257	0.8475	1.3

challenge dataset. Each of these models was used to obtain the probability of prediction maps, which was then combined and used as data for learning an adaptive ensemble model. They used a linear weight, RF, and LightGBM to learn the relationship between the new data and the labels. Their ensemble model showed that LightGBM produced best MCC.

6. Results

In this section, we present the results of all 21 participating teams over the past three years of our GI endoscopy challenges. Below we condense the outcomes of each team's method. It should be noted that only the best scores from the allowed five runs are provided for each task.

6.1. Medico 2017

All teams participated in classification, and speed task, while there was no submission for the hardware tasks, and report task. The average MCC value of all five teams for the classification task on the provided test dataset was 0.7487, with the score ranging from 0.6626 up to 0.8257. A detailed breakdown of the 2017 challenge can be found in Tables 5 and 6. We observe that team SCL-UMD (Agrawal et al., 2017) obtained the best MCC score of 0.8257, which is over 16% increment over HKBU (Liu et al., 2017) who used Bidirectional Marginal Fisher Analysis (BMFA) features and an SVM classifier. Team SIMULA (Pogorelov et al., 2017a) achieved the second-best MCC score and fastest inference time. Both SCL-UMD and SIMULA used Inception-v3 model with one additional CNN model. The high FPS obtained by team SIMULA was due to the use of residual networks, in particular ResNet-50, unlike SLC-UMD team who used VGGNet, which has nearly six times the parameters when compared to ResNet50. A similar trend for the results can be seen for the algorithm efficiency task in Table 6.

6.2. Medico 2018

The 2018 challenge was similar to the one held in 2017 but had an increase of images and classes (14,033 images and 16 classes). The average MCC score for the 11 participating teams was 0.8175, with the score ranging from a minimum of 0.5357 to a maximum of 0.9398. Tables 7 and 8 presents the detailed results of the 2018 challenge. It can be seen that team HCMUS (Hoang et al., 2018) had increment of 40.3% over team HKBU (Ko et al., 2018) which used a combination of Weighted Discriminant Embedding (WDE) and cost-sensitive nearest neighbor (CS-NN) for GI endoscopy image classification. Team S@M achieved the second-highest MCC of

0.9397, with only a marginal gap of 0.0001 than the winning team. The winning team HCMUS (Hoang et al., 2019) used a combination of Residual Neural Network (RNN) and Faster R-CNN to obtain an MCC score of 0.9398.

Six teams participated in the algorithm efficiency task. Table 8 shows the average FPS and classification metrics for the best performing run for each of the participating teams. In GI endoscopy, any team with above 45 FPS can be considered to have real-time system building capability. Therefore, methods from LesCats (Hicks et al., 2018), FAST-NU-DS (Khan and Tahir, 2018), and HKBU (Ko et al., 2018) are considered efficient to be used in a real-time system. However, among these three teams, LesCats (Hicks et al., 2018) has the best MCC score with a reasonable speed. Therefore, we consider the method proposed by team LesCats as the best method for the algorithm efficiency task. To achieve this, LesCats used AlexNet (Krizhevsky et al., 2012).

6.3. BioMedia 2019

The structure of the BioMedia 2019 is similar to that of Medico 2018. A slight change in hardware task was made by introducing Docker-based submission (please see Section 1.3.4 for details). A detailed breakdown of the 2019 challenge results can be found in Table 9, Table 10, and Table 11. In the 2019 challenge, the average MCC for all submitted runs was 0.9287, with scores ranging from 0.8542 to 0.9520. All teams participated in the classification task, of which team Mcdull (Chang et al., 2019) achieved the best result for the classification task.

Three teams participated in the algorithm efficiency task. An $FPS \geq 45$ can be considered real-time performance. Team DeepBlue (Luo et al., 2019) achieved highest MCC and near real-time FPS of 41.51 by utilizing 10 pre-trained ImageNet models and LightGBM. Only two teams participated in the automatic report generation task, namely team uniaugsburg and team CIISR. The submitted reports were manually evaluated by two senior gastroenterologists, where the usefulness in a real-world clinical environment and the correctness of the reporting were the most important criteria.

A defined protocol stated in Section 4.3 as used to assess the report generation task. The submission that was found most useful and accurate by both clinical experts was by the team uniaugsburg (Harzig et al., 2019). Fig. 4 illustrates the sample of the generated report by this team for one of the videos (out of 6 videos) for the automatic report generation task. The report provides a brief summary of the detected findings (frame-level classification) in the provided video and a more detailed summary that includes timestamps for each. Furthermore, by using class activation maps of the predictions, they also provided an approximate location of where

Table 7
Results of 2018 Medico Classification task (Pogorelov et al., 2018b).

Reference	TP	TN	FP	FN	REC	PREC	SPEC	ACC	MCC	F1
LesCats (Hicks et al., 2018)	513.12	8160.62	33.12	33.12	0.9218	0.9378	0.9959	0.9924	0.9325	0.9236
RUNE (Borgli et al., 2018)	510.37	8150.87	35.37	35.37	0.8572	0.8708	0.9956	0.9918	0.9280	0.8555
UMM-SIM (Kirkerød et al., 2018)	501	8148.5	45.25	45.25	0.8433	0.8514	0.9944	0.9896	0.9082	0.8367
ParaNoMundo (Dias and Dias, 2018a)	496.06	8143.56	50.18	50.18	0.8205	0.8414	0.9938	0.9885	0.8983	0.8114
AAUIITEC (Taschwer et al., 2018)	492.06	8139.56	54.18	54.18	0.8673	0.8826	0.9933	0.9876	0.8897	0.8662
SIMULA (Ostroukhova et al., 2018)	474.18	8121.68	72.06	72.06	0.8236	0.8281	0.9911	0.9835	0.8539	0.8145
FAST-NU-DS (Khan and Tahir, 2018)	358.75	8006.25	187.5	187.5	0.6203	0.7173	0.9767	0.9570	0.6302	0.5868
NOAT (Steiner et al., 2018)	314.43	7961.93	231.81	231.81	0.4219	0.5146	0.9717	0.9469	0.5368	0.3913
HKBU (Ko et al., 2018)	315.31	7962.81	230.93	230.93	0.5005	0.4916	0.9715	0.9471	0.5357	0.4829
S@M (Thambawita et al., 2018)	516.62	8164.12	29.62	29.62	0.9361	0.9319	0.9963	0.9932	0.9397	0.9297
HCMUS (Hoang et al., 2018)	516.75	8164.25	29.5	29.5	0.9281	0.9426	0.9963	0.9932	0.9398	0.9342

Table 8
Results of 2018 Medico Efficiency task (Pogorelov et al., 2018b). Method design is based on the trade-off between the accuracy and speed of each algorithm.

Reference	TP	TN	FP	FN	REC	PREC	SPEC	ACC	MCC	F1	FPS
LesCats (Hicks et al., 2018)	498.68	8146.18	47.56	47.56	0.8986	0.8993	0.9941	0.9891	0.9035	0.8883	624.24
ParaNoMundo (Dias and Dias, 2018a)	495.25	8142.75	51	51	0.8194	0.8379	0.9937	0.9883	0.8965	0.8096	8.61
FAST-NU-DS (Khan and Tahir, 2018)	454.43	8101.93	91.81	91.81	0.7527	0.8160	0.9888	0.9789	0.8132	0.7522	43328.71
HKBU (Ko et al., 2018)	315.31	7962.81	230.93	230.93	0.5005	0.4916	0.9715	0.9471	0.5357	0.4829	3744.38
HCMUS (Hoang et al., 2018)	516.75	8164.25	29.5	29.5	0.9281	0.9426	0.9963	0.9932	0.9398	0.9342	23.14

Table 9
Result of the BioMedia challenge 2019 Classification task (Hicks et al., 2019a).

Reference	TP	TN	FP	FN	PREC	REC	SPEC	ACC	MCC	F1
CIISR (Meng et al., 2019)	7570	129888	1167	1167	0.8664	0.8664	0.9911	0.9833	0.8542	0.8664
DeepBlue (Luo et al., 2019)	8329	130647	408	408	0.9533	0.9533	0.9969	0.9941	0.9480	0.9533
HCMUS (Hoang et al., 2019)	8269	130587	468	468	0.9464	0.9464	0.9964	0.9933	0.9406	0.9464
Mcdull (Chang et al., 2019)	8360	130678	377	377	0.9569	0.9569	0.9971	0.9946	0.9520	0.9569
uniaugsburg (Harzig et al., 2019)	8291	130609	446	446	0.9490	0.9490	0.9966	0.9936	0.9490	0.9105

Table 10
Results of BioMedia challenge 2019 Algorithm Efficiency task (Hicks et al., 2019a). Method design is based on the trade-off between the accuracy and speed of each algorithm.

Reference	TP	TN	FP	FN	PREC	REC	SPEC	ACC	MCC	F1	FPS
DeepBlue (Luo et al., 2019)	8270	130588	467	467	0.9465	0.9465	0.9964	0.9933	0.9406	0.9465	41.51
HCMUS (Hoang et al., 2019)	8269	130587	468	468	0.9464	0.9464	0.9964	0.9933	0.9406	0.9464	3.61
uniaugsburg (Harzig et al., 2019)	8108	130426	629	629	0.9280	0.9280	0.9952	0.9910	0.9201	0.9280	3238.87

Table 11
Results of BioMedia challenge 2019 Hardware task (Hicks et al., 2019a).

Reference	TP	TN	FP	FN	PREC	REC	SPEC	ACC	MCC	F1	FPS
CIISR (Meng et al., 2019)	7570	129888	1167	1167	0.8664	0.8664	0.9911	0.9833	0.8542	0.8664	98.90
uniaugsburg (Harzig et al., 2019)	8108	130426	629	629	0.9280	0.9280	0.9952	0.9910	0.9201	0.9280	1271.97

the detected finding was located in the frame. For the hardware task, we had only 2 teams in which uniaugsburg (Harzig et al., 2019) obtained the best MCC and FPS (see Table 11).

Fig. 5 shows a plot of the MCC scores presented by each participant over the three challenges. When we compare the results from 2017 to the results from 2019, we see an average increase of MCC by 18%, and an increase of the best performing MCC by 12.63%. This improvement highlights the progress achieved toward developing an automated system in the field of GI endoscopy and also creates a benchmark for similar challenges in the future.

7. Discussions

We organized the first GI endoscopy challenge that offered the largest multi-class dataset for classification and algorithm efficiency evaluation. Additionally, the automatic report generation task was also an initiative to reduce the endoscopist burden and minimize operator dependence. Below, we provide detailed discus-

sions on findings and limitations of our 2017, 2018 and 2019 challenges.

7.1. Challenge methods

Table 4 presents the summary of different approaches used in all three challenges. To better understand these methods we categorized each method based on their nature (cascaded networks, general CNN models, ensemble models, combined feature approaches, and feature pyramid models) and basis-of-choice that included speed, accuracy, and context choices. Below we provide insight on methods capability for some of the best methods used in these challenges.

For 2017 challenge (Tables 5 and 6), two teams used classical ML approach while the three (out of five) teams explored CNN based approach. Ensemble method designed by the team SLC-UMD (Inception-V3 (Szegedy et al., 2015) and VGGNet (Simonyan and Zisserman, 2014)) and the combined feature approach used by the team SIMULA secured the best results on the final MCC met-

Main findings: =====	
The video mostly shows polyps (69.51%), followed by normal-cecum (28.21%).	
Brief summary: =====	
The video sequence shows the following events in this chronological order: polyps, normal-cecum, polyps, normal-cecum, polyps, normal-cecum, polyps.	
Detailed summary: =====	
FROM - TO	Description of current time period within the video.
00:00-00:02	Polyps can be seen mostly in the center.
00:02-00:07	A normal cecum can be seen mostly in the top-left.
00:07-00:10	Polyps can be seen mostly in the center.
00:10-00:10	The image is blurry and it is hard to identify what currently can be seen.
00:10-00:21	Polyps can be seen mostly in the center.
00:21-00:22	A normal cecum can be seen mostly in the center.
00:22-00:23	Polyps can be seen mostly in the bottom-left.
00:23-00:24	A normal cecum can be seen mostly in the top-left.
00:24-00:27	Polyps can be seen mostly in the center.
00:27-00:28	A normal cecum can be seen mostly in the top-left.
00:28-00:28	Polyps can be seen mostly in the center.
00:28-00:30	A normal cecum can be seen mostly in the center.
00:30-00:32	Polyps can be seen mostly in the bottom-left.
00:32-00:33	A normal cecum can be seen mostly in the center.
00:33-00:38	Polyps can be seen mostly in the top-left.
00:38-00:40	A normal cecum can be seen mostly in the bottom-right.
00:40-00:41	The image is blurry and it is hard to identify what currently can be seen.
00:41-00:42	Polyps can be seen mostly in the bottom-left.
00:42-00:44	A normal cecum can be seen mostly in the center.
00:44-00:57	Polyps can be seen mostly in the top-left.

Fig. 4. Generated report for polyp resection, bleeding videos from automatic report generation task.

ric for classification (0.8257 and 0.8020, respectively). This improvement was nearly 10% more than the other general CNN-based method (e.g., team ITEC-AAU). Similarly, ensemble of combined feature approaches also made a mark on the score chart in 2018 and 2019 challenges (see Tables 7 and 9). Team HCMUS that used a box regression network together with the feature extraction network won the challenge in 2018, while team Mcdull won the 2019 challenge where they fused several network backbones and implemented an attention mechanism with Inception-V3 architecture. These results demonstrate that while ensemble or fused feature-based methods resulted in improved performances, the choice of each network in these methods affect the algorithm performance, reliability and usability. For example, the choice of Inception-v3 and VGGNet by SLC-UMD limits the depth of feature extraction and risk of vanishing gradient problem. Addition-

ally, VGGNet has extremely high number of trainable parameters (e.g., VGG-16 (Simonyan and Zisserman, 2014) has roughly 138 million) compared to the ResNet-50 counterpart (only 23 million). Clearly, in this context, the approach taken by SIMULA team has more strength where they exploited ResNet-50 that includes feature fusion through skip-connection and less number of trainable parameters compared to VGGNet. Table 6 for algorithm efficiency task also demonstrates this case where the computational speed is largely compromised in the method presented by SLC-UMD (FPS of 1.3 only) compared to near real-time speed for team SIMULA with FPS of 46.0.

Most methods that topped the evaluation chart for 2018 and 2019 used ensemble or feature fusion networks. Similar to 2017, the network choices can be seen to have a direct consequence on the applicability issue and model strength. For e.g., the winning team HCMUS used the detection method for classification task using deeper ResNet-101 (He et al., 2016) model (44.5 million parameters) as backbone and bounding box regressor network which showed serious consequence in compromise in speed (FPS of only 23.14) when tested for efficiency task (refer Table 8). On contrary, LesCats which used DenseNet-169 (Huang et al., 2017) (14.3 million) with fewer parameters than ResNet models and embodied skip-connections without fusion has a clear advantage in terms of trade-off between computation speed and accuracy. It can be observed in Table 8 that the MCC for team LesCats is 0.9035 at FPS of 624.24 compared to MCC of 0.9398 with just over 23 FPS for HCMUS. The choice of method by LesCats has clearly more strength and provided a promise for real-time clinical applicability.

Again, for 2019 (Table 9–11), the model choice as well as the GPU choices (Table 4) directly implicated the strength of each designed method and its clinical applicability such as speed. For example, the best performing team on classification task (MCC = 0.9520) fused several feature extraction backbones including ResNet and SE-ResNeXt (Hu et al., 2018) and several Inception-V3 (Szegedy et al., 2015) models (24 million parameters) for attention mechanism (see Fig. 6). The second best method with MCC = 0.9490 on classification task used MobileNet-V2 (Howard et al., 2017) (6.9 million parameters only) and efficient DenseNet-121 model (reduced parameters compared to ResNet). Due to the choice of the models a best trade-off between speed and accuracy was observed when tested for algorithm efficiency where the team used only MobileNet-V2. For this task, the team obtained a real-time performance with FPS of 3238.87 at a competitive MCC score of 0.9201. While the second best DeepBlue used 10 sub-models

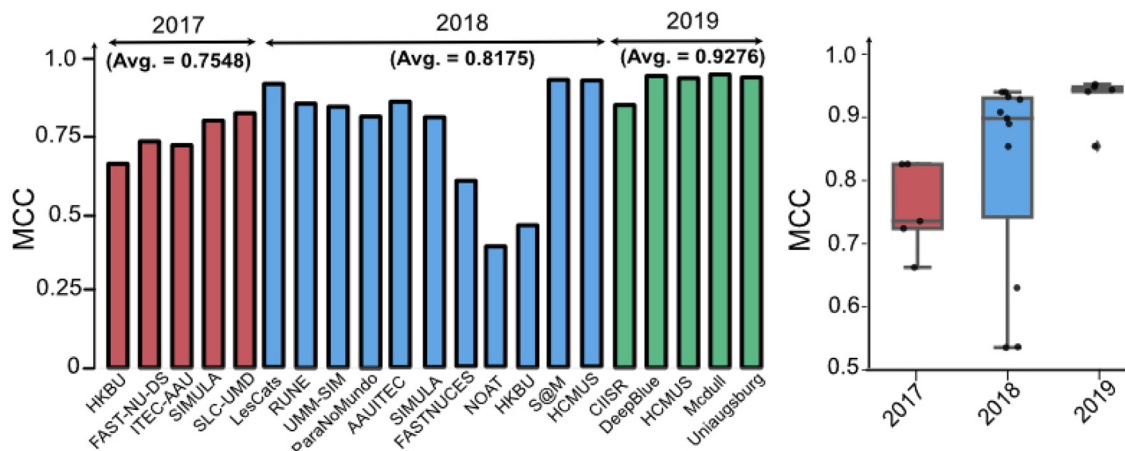


Fig. 5. MCC score comparison of different participating teams in Medico 2017, Medico 2018, and BioMedia 2019 challenges. On left individual team scores (bar plot), and on the right statistics of each year submission (box plot).

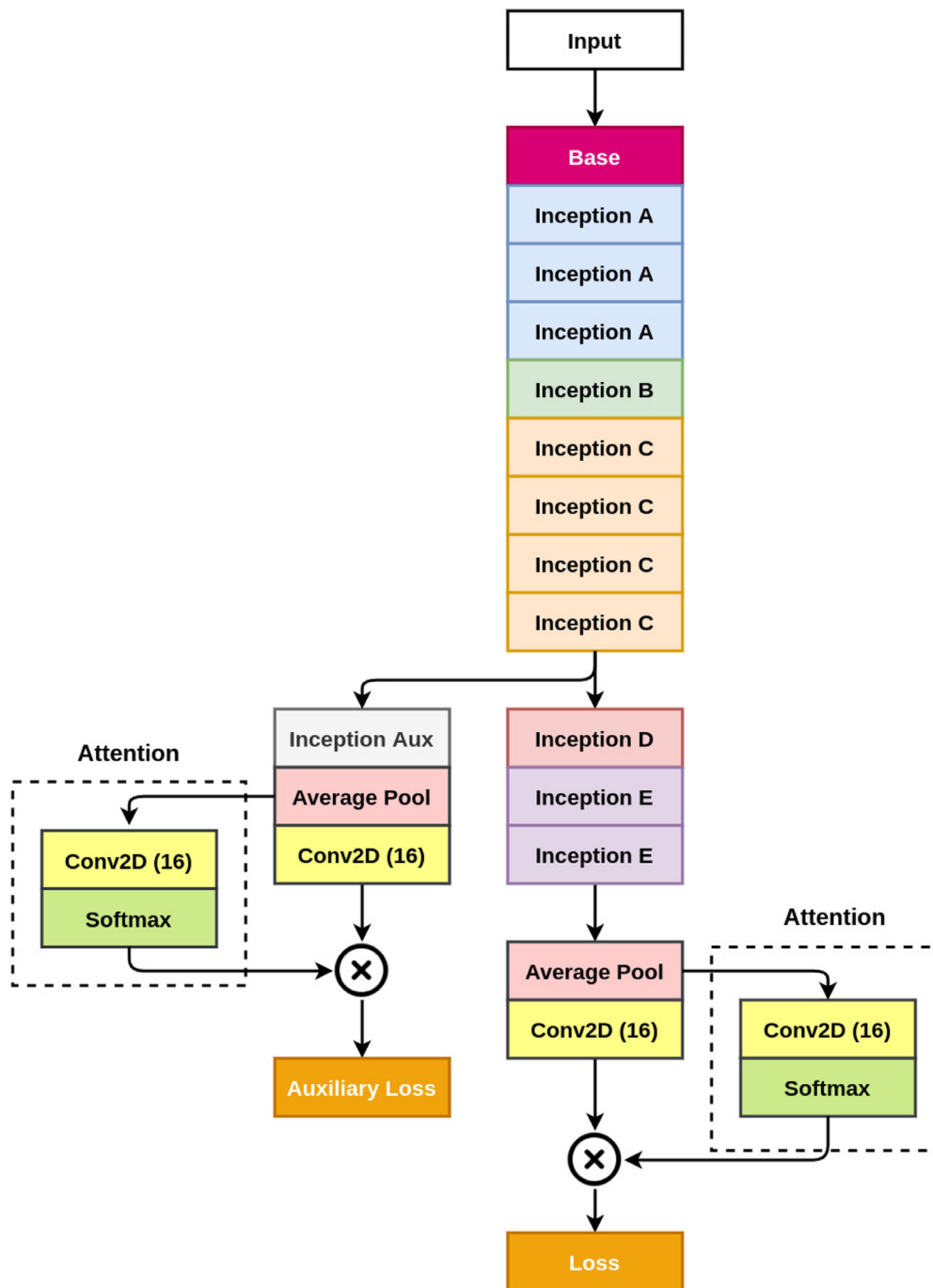


Fig. 6. Best architecture in Medico 2018 “classification task” Team Mcdull (Chang et al., 2019).

(Fig. 7) with 10 cross-fold validation resulting in improved MCC score but with a sacrifice in speed achieving FPS of only 41.51.

For the hardware task (i.e., methods tested on the same NVIDIA GTX 1080 Ti GPU), uniaugsburg team which used light weight model MobileNet-V2 (Howard et al., 2017) (6.9 million parameters only) provided a real time application strength of FPS of 1271.98 with MCC above 0.92. Similarly, for other teams which used only single model, their accuracy also depended on the model choice itself. For example, team RUNE that used DenseNet-169 (Huang et al., 2017) has nearly 7% improvement over team SIM-ULA that used Inception-V3 model. Compared to DL methods, all classical ML methods including the teams that utilized ensemble or fusion networks (e.g., team Fast-Nu-DS in 2017 and 2018, and

NOAT and HKBU in 2018) resulted in a worse performance even though they provided a promise for real-time application (for e.g., teams HKBU and Fast-NU-DS in Table 8). There is no surprise that no team in 2019 competition used classical ML approaches. It is to be noted that other metrics such as precision, recall, specificity and accuracy appear to be proportional to the MCC metric used for evaluating the methods in these challenges and hence have been tabulated but not discussed here.

Even though only two teams participated in our automated report generation task, it provided an evidence of the strength of automated methods and their clinical usability for reporting (e.g., location of disease or anatomy, timestamp in video, % of occurrence of different findings (see Fig. 4)). While, manual post-analysis

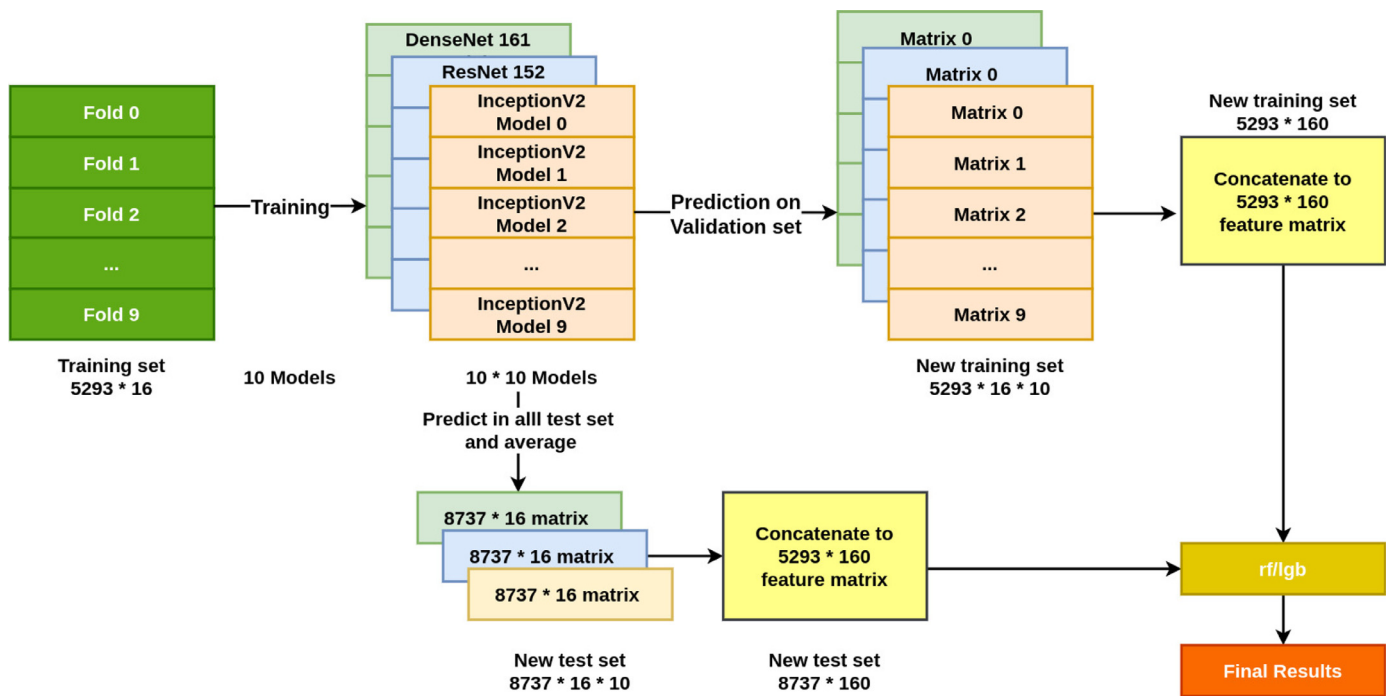


Fig. 7. The architecture of the best performing team in Biomedica 2019 challenge (Team DeepBlue (Luo et al., 2019)).

of the acquired raw videos is close to impossible, and evidently most recorded procedures are almost never re-visited for retrospective case understanding, our automated report generation task demonstrated an utmost feasibility and strength of the deep learning methods that can be utilized to obtain clinically valuable automated reporting and provide a potential for post-analysis of patients. However, the reliability of such approaches need to be rigorously studied in the future.

7.2. Challenge outcomes and clinical applicability

As detailed in the previous section, each method had its strengths and weaknesses based on their choice of the approach. A major outcome of each year’s challenge revealed several interesting findings, such as the evolution of methods in the classification task, their ability to provide reliable accuracy when evaluated on the same machine (robustness test) and their inference speed. In Table 12, we ranked each team based on these important criteria. When a longitudinal comparison was done, methods submitted in 2018 and 2019 surpassed those in 2017. Similarly, most top-ranking methods were from 2019.

In the literature, there are several useful recommendations towards developing clinically acceptable CADx systems for colonoscopy (Mori et al., 2017) or polyp detection (Bernal et al., 2017). For example, models performing over 64 FPS can be in general considered to provide real-time performance, which is very critical in a clinical environment. The clinical applicability of these methods is one important dissection in Table 12 (refer to the last column), which is based on accuracy (MCC Rank), speed, and robustness (RR) ranks. In our ranking, it can be observed that team DeepBlue had the best clinical translation capability with 41.51 FPS in speed and 1st and 2nd ranks in the robustness and accuracy, respectively. Similarly, team uniaugsburg from 2019 challenge and team HCMUS from 2018 achieved 2nd rank in our clinical applicability test. It is to be noted that HCMUS ranked 1st in the robustness rank while uniaugsburg ranked 1st in the speed rank and only 3rd in the robustness rank. However, developed methods by all the teams in 2017 have low clinical translation capability.

	500	0	0	0	0	0	0	0	0	0	0	39	0	3	0	1	1	0	7
(A) Ulcerative colitis	3	432	48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(B) Esophagitis	1	121	513	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
(C) Normal-z-line	1	0	0	522	31	0	0	0	0	0	0	0	0	2	0	0	0	0	34
(D) Dyed-resection-polyps	0	0	0	33	532	0	0	0	0	0	0	0	0	1	0	0	0	0	17
(E) Out-of-patient	0	0	0	0	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0
(F) Normal-pylorus	3	3	2	0	0	0	559	0	0	0	0	0	2	0	0	0	0	0	0
(G) Normal-inclusions	0	0	0	0	0	0	0	501	7	0	0	0	0	0	0	0	0	0	0
(H) Stool-plenty	1	0	0	0	0	0	0	0	1918	0	0	0	0	0	0	0	0	0	1
(I) Blurry-nothing	1	0	0	0	0	0	0	0	1	37	0	0	0	0	0	0	0	0	0
(J) Polyps	10	0	0	1	0	0	1	0	0	0	0	0	358	6	0	1	0	46	
(K) Normal-cecum	18	0	0	0	0	0	0	0	0	0	0	0	6	578	0	0	0	0	2
(L) Colon-clear	1	0	0	0	0	0	0	5	0	0	0	0	0	0	1063	0	1	0	
(M) Retroflex-rectum	3	0	0	0	0	0	0	0	0	0	0	0	2	0	188	1	0		
(N) Retroflex-stomach	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2	395	1		
(O) Instruments	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	165	
	(A) Ulcerative colitis	(B) Esophagitis	(C) Normal-z-line	(D) Dyed-resection-polyps	(E) Out-of-patient	(F) Normal-pylorus	(G) Normal-inclusions	(H) Stool-plenty	(I) Blurry-nothing	(J) Polyps	(K) Normal-cecum	(L) Colon-clear	(M) Retroflex-rectum	(N) Retroflex-stomach	(O) Instruments				

Fig. 8. Confusion matrix plot of Team S@M (Thambawita et al., 2020). A-P represents class labels.

7.3. Limitations of Medico challenges

7.3.1. Analysis of the failed classes

In this section, we analyze the results based on performance of each class of the dataset.

Esophagitis vs normal Z-line. In most of the presented approaches in the three challenges, the significant misclassification was observed between ‘esophagitis’ and ‘normal-z-line’ classes. In Fig. 8, it can be observed that the esophagitis class (B) and the normal-z-line class (C) were the most confused classes. The same problem was observed for all teams (Hicks et al., 2018; Meng et al.,

Table 12

Clinical applicability of the participants methods that considers MCC, efficiency, and speed into account. Here Clas. = MCC classification, AR = MCC Algorithm Robustness, RR = Robustness-rank, SR = Speed-rank, Rank= MCC Rank, CAR = Clinical applicability rank and na = not available. 10 is the imputed rank for speed and robustness ranks.

Year	Team	Clas.	AR	Speed	RR	SR	Rank	CAR
2017	HKBU	0.6626	0.6946	2.2	4	10	18	8
	FAST-NU	0.7331	0.7114	2.3	3	10	16	7
	ITEC-AAU	0.7202	0.7202	1.4	1	10	17	7
	SIMULA	0.8220	0.7856	46	10	2	14	5
	SLC-UMD	0.8257	0.8257	1.3	1	10	15	7
2018	LesCats	0.9325	0.9035	624	3	1	7	3
	RUNE	0.928	na	na	na	na	8	na
	UMM-SIM	0.9082	na	na	na	na	9	na
	ParaNoMundo	0.8983	0.8965	8.61	1	10	10	4
	AAUITEC	0.8897	na	na	na	na	11	na
	FAST-NU-DS	0.6302	0.8132	43329	10	1	19	7
	NOAT	0.5368	na	na	na	na	20	na
	HKBU	0.5357	0.5357	3744.4	1	1	21	6
	S@M	0.9397	na	na	na	na	6	na
	HCMUS	0.9398	0.9342	23	1	3	5	2
	CIISR	0.8542	0.8542	98.9	1	1	12	3
2019	DeepBlue	0.948	0.9406	3226	1	1	2	1
	HCMUS	0.9406	0.9406	3.6	1	10	4	4
	Mcdull	0.9520	na	na	na	na	1	na
	uniaugsburg	0.9490	0.9201	1272	3	1	3	2

2019; Dias and Dias, 2018b; Agrawal et al., 2017). One of the reasons is their location as they both exist very close to each other (see Fig. 9).

Dyed-resection-margins vs dyed-lifted-polyp Other significant challenges were observed in the 'dyed-resection-margins' (class E) and 'dyed-lifted-polyps' (class D) classes. This is again evident from confusion matrix Fig. 8. For the Medico test dataset, there were a total of 64 misclassification for these two classes in the method of Team S@M (Thambawita et al., 2020). Similar problems were also seen in other teams performance. The primary reason for misclassification can be due to similarity between these two classes, for example, in terms of their color properties (see Fig. 1, first row, fourth column, and second-row first column). The other reasons behind the class confusion in both the above cases can be due to the model choice, use of simple data augmentation, and choice of the loss function.

7.3.2. Limitations of the study

The curated dataset consisted of green patches that are present in the real clinical endoscopy data used for location guidance by endoscopists. However, this may have affected some of the methods' performance due to the confusion of these local patches with other classes that consisted of a similar green patch. Additionally, in terms of color and semantic features, some chosen class labels were very similar (e.g., class B - Esophagitis and class C - normal-z-line). There can be presence of label biases due to presence of both instrument class and disease class category as well. As a result, the conducted study is susceptible to algorithmic errors due to

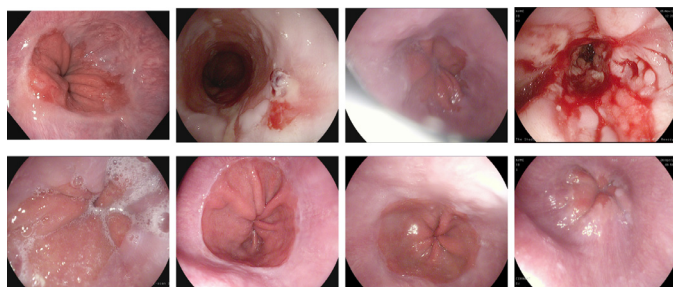


Fig. 9. Example frames from 'Esophagitis' and 'Normal-z-line' class.

dataset complexity. Additionally, very similar images were present even though they were taken from different videos. This can create pseudo data balance due to which algorithms can fail to generalize. Even though we have taken a larger patient cohort, the ability of methods to generalize on different endoscopic data or on a different patient cohort can result in unpredictable outcomes. Other limitation of our challenges was not having an automated leaderboard as a result the prediction maps sent by the teams may be sub-optimal and could have error in some metrics especially in inference time reporting. Similarly, manual scoring of the report generation task can be prone to human errors and biases.

7.4. Trust, safety, and interpretability of methods

With the hardware and software advancements over past years, it is evident from the presented challenge series that a significant improvement on reliability of methods is observed over time (see Fig. 5). However, with the case variability it is also vital to incorporate more challenges in the dataset to be addressed. We almost doubled data in 2017 in 2018 and 2019 challenges.

Other important issue is the assessment of methods on real clinical settings where the negative samples are tremendously higher than in the curated data for research and development. Often patient safety is of direct concern as wrong detection of any lesion can result in wrong procedure. Thus, an assessment of methods on real-world clinical scenarios is needed. With the report generation task in 2019, we attempted to address this issue by providing 6 raw videos to the participating teams. Though this attracted only two participating teams, efficacy and reliability of methods were tested.

Confusion between similar looking samples which are easily distinguishable by the human but methods may fail to interpret due to the lack of enough samples in training is a common problem (e.g., failed cases presented in Section 7.3). It is therefore vital to improve interpretability of the methods by injecting more negative samples to improve context-awareness of methods. In all challenges of this series, while most methods were designed to regress heavily on the presented features using heavy augmentation similar to natural scene domain, a more careful approaches must be built for clinical videos or images, particularly in endoscopy, by preserving both geometric and contextual features dur-

ing any transformation strategy. Additionally, including temporal awareness (e.g., use of LSTMS (Xiao et al., 2018) for sequences) or use of metric learning approaches by understanding the embedding distances (e.g., use of few-shot learning (Tian et al., 2020)) can be next step forward to improve reliability of methods. Both of these were not exploited by any team in these challenges.

7.5. Future steps and strategies

The three consecutive challenges revealed a progress in method development, and competence of teams to achieve improved scores. However, the choice-of-methods still depended on fine-tuning approaches and use of off-the-shelf methods. Almost all teams that used DL approaches used pre-trained methods that were trained on natural images (e.g., ImageNet dataset). Only a few team tried to use medical image datasets. A major challenge in the medical imaging community is the availability and accessibility of large datasets. As a result, the complex medical features cannot be learnt. This becomes more prominent problem when the images are merged for multi-class classification as in our case. To this note, we have been working immensely on increasing the dataset size and at the same time making it accessible for researchers. Our effort has lead to the HyperKvasir (Borgli, 2020), open-access dataset that contains 110,079 images and 373 videos and Kvasir-Capsule, an open access video capsule endoscopy dataset that consists of 18 videos which can be used for extraction of 4,820,739 image frames (Smedsrud et al., 2020). While, clinical image data are vague and is prone to severe distortions when generated using adversarial networks or performing unrealistic data augmentation, it is vital step in overcoming the unreliability of built technologies in this regime.

A second key strategy that we have learnt from our challenges is to provide algorithm performances on different baseline approaches so that the teams do not have to try off-the-shelf methods by themselves, giving them more time to better design the methods to overcome the limitations of ML approaches on provided GI dataset.

Finally, it is important to address the clinical applicability of each developed method and independently rank teams based on their merit of clinical usability. A metric can be a weighted score between speed, accuracy and robustness. Additionally, the built models can be tested on the real clinical environment for its accuracy, speed and reliability tests in real-world clinical settings. For this we intend to use clinical hardware systems and integrate these models during endoscopy procedures and confirm the reliability with the clinical expert on both easy and hard cases.

8. Conclusion

A comprehensive evaluation, comparison, and summarization of different presented methods in the MediaEval Medico 2017, MediaEval Medico 2018, and BioMedia 2019 challenges are presented in this paper. Varied methodologies were used: from traditional Machine Learning methods based on global features to recent state-of-the-art Convolutional Neural Network methods. Several teams also demonstrated the use of specialized data augmentation techniques. Here, we have provided an overview of several baseline methods using standard computer vision metrics on a common publicly available benchmark dataset. We advocate that using such a systematic approach of method evaluation and analysis is necessary and provides the best practice towards method development in GI endoscopy imaging.

Each year we observed significant improvements in both both classification and algorithm robustness tasks. More importantly, the efficiency results of Medico task 2018 and BioMedia 2019 show that it is possible to achieve real-time for GI endoscopy. The automatic reporting task was one of the first effort to communicate the

algorithmic findings with clinical experts. Thus, this study highlighted the significance of collaboration between endoscopists and computer scientists to develop a meaningful medical image analysis tools that can assist endoscopists to reduce their clinical workload.

The study also highlighted the need for the collection of larger endoscopic image dataset that incorporates wider class categories, and different modalities. We showed that both objective and subjective metrics are critical for obtaining insights in the developed methods and their reliability for use in clinical settings. From the different submissions, we observed that there is a trade-off between speed and accuracy. So, we ranked each team based on these scores and provided an average score determining their clinical relevance rank. Our analysis showed that teams that achieved one of the highest classification accuracy ranked lower than team with a modest accuracy.

Further research direction includes investigation on tackling the challenges related to integration of multi-modality, multi-centered and multi-organ data and feedback from endoscopists for developing more robust systems. A consensus should be reached to improve understanding and interpretability of the results of CNN models. A potentially optimized combination of them could be helpful to build clinically useful method.

Author contribution

D. Jha conceptualized the work. The paper was written and revised mostly by D. Jha and S. Ali with the input from all the co-authors. T. D. Lange provided input on clinical aspect of the study. All the analyses presented in the paper was done by challenge organizers (S. Hicks, K. Pogorelov, M. A. Riegler, P. Halvorsen), D. Jha and S. Ali. The details on the methods were provided by the participating teams. All authors participated in the revision of this manuscript and provided substantial input.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research is partially funded by the PRIVATON project (#263248) and the Autocap project (#282315) from the Research Council of Norway (CRN). Our experiments were performed on the Experimental Infrastructure for Exploration of Exascale Computing (eX3) system, which is financially supported by CRN under contract 270053. D. Jha is funded by PRIVATON project and S. Ali is supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

- Aabakken, L., et al., 2014. Standardized endoscopic reporting. *J. Gastroenterol. Hepatol.* 29 (2), 234–240.
- Agrawal, T., Gupta, R., Sahu, S., Espy-Wilson, C.Y., 2017. SCL-UMD at the medico task-mediaeval 2017: transfer learning based classification of medical images. In: *Proc. CEUR Worksh. Multim. Bench. Worksh. (MediaEval)*.
- Akbari, M., Mohrekeesh, M., Nasr-Esfahani, E., Soroushmehr, S.R., Karimi, N., Samavi, S., Najarian, K., 2018. Polyp segmentation in colonoscopy images using fully convolutional network. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 69–72.
- Ali, S., Dmitrieva, M., Ghatwary, N., Bano, S., Polat, G., Temizel, A., ... Rittscher, J., 2021. Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. *Med. Image Anal.*, 102002.
- Ali, S., et al., 2019. Endoscopy artifact detection (EAD 2019) challenge dataset. arXiv:1905.03209.

- Ali, S., et al., 2020a. Endoscopy disease detection challenge 2020. arXiv:2003.03376.
- Ali, S., et al., 2020. An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. *Sci. Rep.* 1–21.
- Angermann, Q., et al., 2017. Towards real-time polyp detection in colonoscopy videos: adapting still frame-based methodologies for video sequences analysis. In: *Comput. Assist. and Robot. Endoscopy and Clin. Image-Based Proced. (CARE CLIP)*, Vol. 10550, 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. Surg. Oncol.* 25 (9), 2693–2702.
- 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., et al., 2017. Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 endoscopic vision challenge. *IEEE Trans. Med. Imaging* 36 (6), 1231–1249.
- Bernal, J., et al., 2018. Polyp detection benchmark in colonoscopy videos using GTCreator: a novel fully configurable tool for easy and fast annotation of image databases. In: *Proc. Comput. Assist. Radiol. Surg. (CARS)*.
- Borgli, H., et al., 2020. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Sci. Data* 7, 1–14 1.
- Borgli, R.J., Halvorsen, P., Riegler, M., Stensland, H.K., 2018. Automatic hyperparameter optimization in keras for the mediaeval 2018 medico multimedia task. In: *Proc. CEUR Worksh. Multim. Bench. Worksh. (MediaEval)*.
- Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A., 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68 (6), 394–424.
- Chang, Y., Huang, Z., Chen, W., Shen, Q., 2019. Gastrointestinal tract diseases detection with deep attention neural network. In: *Proc. ACM Internat. Conf. Multim.*, pp. 2568–2572.
- Chhedha, T., Iyer, R., Koppaka, S., Kalbande, D., 2020. Gastrointestinal tract anomaly detection from endoscopic videos using object detection approach. In: *International Symposium on Visual Computing*, pp. 494–505.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: a large-scale hierarchical image database. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 248–255.
- Denton, E., Gross, S., Fergus, R., 2016. Semi-supervised learning with context-conditional generative adversarial networks. arXiv:1611.06430.
- Dias, D., Dias, U., 2018. Transfer learning with CNN architectures for classifying gastrointestinal diseases and anatomical landmarks. In: *Proc. CEUR Worksh. on Multim. Bench. Worksh. (MediaEval)*.
- Dias, D., Dias, U., 2018. Transfer learning with CNN architectures for classifying gastrointestinal diseases and anatomical landmarks. In: *Proc. CEUR Worksh. on Multim. Bench. Worksh. (MediaEval)*.
- Goodfellow, I., et al., 2014. Generative adversarial nets. In: *Proc. NIPS*, pp. 2672–2680.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. *ACM SIGKDD Explora. Newslett.* 11 (1), 10–18.
- Harzig, P., Einfalt, M., Lienhart, R., 2019. Automatic disease detection and report generation for gastrointestinal tract examination. In: *Proc. ACM Internat. Conf. Multim.*, pp. 2573–2577.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778.
- Hicks, S., et al., 2019. ACM multimedia biomed 2019 grand challenge overview. In: *Proc. ACM Int. Conf. Multim.*, pp. 2563–2567.
- Hicks, S., et al., 2019. Deep learning for automatic generation of endoscopy reports. *Gastrointest. Endosc.* 89 (6), AB77.
- Hicks, S.A., Smedsrud, P.H., Halvorsen, P., Riegler, M., 2018. Deep learning based disease detection using domain specific transfer learning. In: *Proc. CEUR Worksh. Multim. Bench. Worksh. (MediaEval)*.
- Hoang, T.-H., Nguyen, H.-D., Nguyen, T.-A., 2018. An application of residual network and faster - RCNN for medico: multimedia task at mediaeval 2018. In: *Proc. CEUR Worksh. Multim. Bench. Worksh. (MediaEval)*.
- Hoang, T.-H., Nguyen, H.-D., Nguyen, V.-A., Nguyen, T.-A., Nguyen, V.-T., Tran, M.-T., 2019. Enhancing endoscopic image classification with symptom localization and data augmentation. In: *Proc. ACM Internat. Conf. Multim.*, pp. 2578–2582.
- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861.
- Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141.
- Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4700–4708.
- Iakovidis, D.K., Georgakopoulos, S.V., Vasilakakis, M., Koulaouzidis, A., Plagianakos, V.P., 2018. Detecting and locating gastrointestinal anomalies using deep learning and iterative cluster unification. *IEEE Trans. Med. Imaging* 37 (10), 2196–2210.
- Jha, D., Smedsrud, H.P., Johansen, D., De Lange, T., Johansen, H.D., Halvorsen, P., Riegler, M.A., 2021. A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation. *IEEE J. Biomed. Health Inf.*
- Jha, D., et al., 2020. Kvasir-SEG: a segmented polyp dataset. In: *Proc. Int. Conf. Multim. Model. (MMM)*, pp. 451–462.
- Khan, Z., Tahir, M.A., 2018. Majority voting of heterogeneous classifiers for finding abnormalities in the gastro-intestinal tract. In: *Proc. CEUR Worksh. Multim. Bench. Worksh. (MediaEval)*.
- Khorasani, H.M., Usefi, H., Peña-Castillo, L., 2020. Detecting ulcerative colitis from colon samples using efficient feature selection and machine learning. *Sci. Rep.* 10 (1), 1–9.
- Kirkerød, M., Thambawita, V., Riegler, M., Halvorsen, P., 2018. Using preprocessing as a tool in medical image detection. In: *Proc. CEUR Worksh. Multim. Bench. Worksh. (MediaEval)*.
- Ko, H.T., Gu, Z., Tahir, L.Y., 2018. Weighted discriminant embedding: Discriminant subspace learning for imbalanced medical data classification. In: *Proc. CEUR Worksh. Multim. Bench. Worksh. (MediaEval)*.
- Krebs, A., Benezeth, Y., Bazin, T., Marzani, F., Lamarque, D., 2020. Pre-cancerous stomach lesion detections with multispectral-augmented endoscopic prototype. *Appl. Sci.* 10 (3), 795.
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. In: *Proc. NIPS*, pp. 1097–1105.
- Lee, J.Y., Jeong, J., Song, E.M., Ha, C., Lee, H.J., Koo, J.E., Yang, D.-H., Kim, N., Byeon, J.-S., 2020. Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets. *Sci. Rep.* 10 (1), 1–9.
- Levin, B., et al., 2008. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the american cancer society, the us multi-society task force on colorectal cancer, and the american college of radiology. *CA Cancer J. Clin.* 58 (3), 130–160.
- Liu, Y., Gu, Z., Cheung, W.K., 2017. HKBU at mediaeval 2017 medico: medical multimedia task. In: *Proc. CEUR Worksh. Multim. Bench. Worksh. (MediaEval)*.
- Luo, Z., Wang, X., Xu, Z., Li, X., Li, J., 2019. Adaptive ensemble: Solution to the biomedica ACM MM grandchallenge 2019. In: *Proc. ACM Int. Conf. on Multim.*, pp. 2583–2587.
- Lux, M., Chatzichristofis, S.A., 2008. Lire: lucene image retrieval: an extensible java CBIR library. In: *Proc. ACM Int. Conf. Multim.*, pp. 1085–1088.
- Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophys. Acta (BBA)-Prot. Struct.* 405 (2), 442–451.
- Meng, W., Zhang, S., Yao, X., Yang, X., Xu, C., Huang, X., 2019. Biomedica ACM MM grand challenge 2019: using data enhancement to solve sample unbalance. In: *Proc. ACM Int. Conf. Multim.*, pp. 2588–2592.
- Mori, Y., Kudo, S.-e., Berzin, T.M., Misawa, M., Takeda, K., 2017. Computer-aided diagnosis for colonoscopy. *Endoscopy* 49 (08), 813–819.
- Naqvi, S.S.A., Nadeem, S., Zaid, M., Tahir, M.A., 2017. Ensemble of texture features for finding abnormalities in the gastro-intestinal tract. In: *Proc. CEUR Worksh. Multim. Bench. Worksh. (MediaEval)*.
- Ostroukhova, O., Pogorelov, K., Riegler, M., Dang-Nguyen, D.-T., Halvorsen, P., 2018. Transfer learning with prioritized classification and training dataset equalization for medical objects detection. In: *Proc. CEUR Worksh. Multim. Bench. Worksh. (MediaEval)*.
- Petschmann, S., Schöffmann, K., Lux, M., 2017. An inception-like CNN architecture for Gi disease and anatomical landmark classification. In: *Proc. CEUR Worksh. Multim. Bench. Worksh. (MediaEval)*.
- Pogorelov, K., Ostroukhova, O., Jeppsson, M., Espeland, H., Griwodz, C., de Lange, T., Johansen, D., Riegler, M., Halvorsen, P., 2018. Deep learning and hand-crafted feature based approaches for polyp detection in medical videos. In: *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 381–386.
- Pogorelov, K., et al., 2017. A comparison of deep learning with global features for gastrointestinal disease detection. In: *Proc. CEUR Worksh. Multim. Bench. Worksh. (MediaEval)*.
- Pogorelov, K., et al., 2017. KVASIR: a multi-class image dataset for computer aided gastrointestinal disease detection. In: *Proc. ACM Multim. Sys. Conf. (MMSys)*, pp. 164–169.
- Pogorelov, K., et al., 2017. Nerthus: a bowel preparation quality video dataset. In: *Proc. ACM Multim. Sys. Conf. (MMSys)*, pp. 170–174.
- Pogorelov, K., et al., 2018. Medico multimedia task at mediaeval 2018. In: *Proc. CEUR Worksh. Multim. Bench. Worksh. (MediaEval)*.
- Poon, C.C., Jiang, Y., Zhang, R., Lo, W.W., Cheung, M.S., Yu, R., Zheng, Y., Wong, J.C., Liu, Q., Wong, S.H., et al., 2020. AI-doscopist: a real-time deep-learning-based algorithm for localising polyps in colonoscopy videos with edge computing devices. *NPJ Digit. Med.* 3 (1), 1–8.
- Poudel, S., Kim, Y.J., Vo, D.M., Lee, S.-W., 2020. Colorectal disease classification using efficiently scaled dilation in convolutional neural network. *IEEE Access*.
- Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In: *Proc. NIPS*, pp. 91–99.
- Riegler, M., et al., 2017. Multimedia for medicine: the medico task at mediaeval 2017. In: *Proc. CEUR Worksh. Multim. Bench. Worksh. (MediaEval)*.
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2018. MobileNetV2: inverted residuals and linear bottlenecks. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4510–4520.
- Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.
- Smedsrud, P.H., Gjestang, H.L., Nedrejord, O.O., Naess, E., Thambawita, V., Hicks, S., Borgli, H., Jha, D., Berstad, T.J.D., Eskeland, S. L., et al., 2020. Kvasir-Capsule, a video capsule endoscopy dataset. *Sci. Data* In press.

- Song, E.M., Park, B., Ha, C.-A., Hwang, S.W., Park, S.H., Yang, D.-H., Ye, B.D., Myung, S.-J., Yang, S.-K., Kim, N., et al., 2020. Endoscopic diagnosis and treatment planning for colorectal polyps using a deep-learning model. *Sci. Rep.* 10 (1), 1–10.
- Steiner, M., Lux, M., Halvorsen, P., 2018. The 2018 medico multimedia task submission of team NOAT using neural network features and search-based classification. In: *Proc. CEUR Worksh. Multim. Bench. Worksh. (MediaEval)*.
- Suzuki, K., 2012. A review of computer-aided diagnosis in thoracic and colonic imaging. *Quant. Imaging Med. Surg.* 2 (3), 163–176.
- Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In: *Proc. AAAI Conf. Artif. Intell.*
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2818–2826.
- Taschwer, M., Primus, M.J., Schoeffmann, K., Marques, O., 2018. Early and late fusion of classifiers for the mediaeval medico task. In: *Proc. CEUR Worksh. Multim. Bench. Worksh. (MediaEval)*.
- Thambawita, V., et al., 2018. The medico-task 2018: disease detection in the gastrointestinal tract using global features and deep learning. In: *Proc. CEUR Worksh. Multim. Bench. Worksh. (MediaEval)*.
- Thambawita, V., et al., 2020. An extensive study on cross-dataset bias and evaluation metrics interpretation for machine learning applied to gastrointestinal tract abnormality classification. *ACM Trans. Comput. Healthca.*
- Tian, Y., Maicas, G., Pu, L.Z.C.T., Singh, R., Verjans, J.W., Carneiro, G., 2020. Few-shot anomaly detection for polyp frames from colonoscopy. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 274–284.
- Wan, J.-J., Chen, B.-L., Kong, Y.-X., Ma, X.-G., Yu, Y.-T., 2019. An early intestinal cancer prediction algorithm based on deep belief network. *Sci. Rep.* 9 (1), 1–13.
- Wang, P., et al., 2018. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nat. Biomed. Eng.* 2 (10), 741.
- Woolhandler, S., Himmelstein, D.U., 2014. Administrative work consumes one-sixth of us physicians' working hours and lowers their career satisfaction. *Int. Journ. Health. Serv.* 44 (4), 635–642.
- Xiao, W.-T., Chang, L.-J., Liu, W.-M., 2018. Semantic segmentation of colorectal polyps with deeplab and LSTM networks. In: *2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, pp. 1–2.
- Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1492–1500.
- Yamada, M., et al., 2019. Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy. *Sci. Rep.* 9 (1), 1–9.
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2921–2929.