UiT The Arctic University of Norway

Faculty of Science and Technology

Department of Computer Science

Machine Learning-based Classification, Detection, and Segmentation of Medical images

Debesh Jha

A dissertation for the degree of Philosophiae Doctor - August 2021

Machine Learning-based Classification, Detection, and Segmentation of Medical Images

Debesh Jha
PhD Programme in Computer Science
Faculty of Science and Technology
UiT - The Arctic University of Norway

To my loving parents who has always been source of inspiration throughout my life.

"Research means that
you don't know, but
are willing to find out."
Charles Franklin Kettering (1876-1958)

Preface

This thesis is submitted as a partial fulfilment of the requirement of the Philosophie Doctor at the Department of Computer Science, Faculty of Science and Technology, UiT The Arctic University of Norway. This work is conducted under the supervision of Associate Prof. Håvard D. Johansen, Prof. Pål Halvorsen, Chief Research Scientist Michael A. Riegler, and Prof. Dag Johansen.

The research was carried at Simula Research Laboratory, Simula School of Research and Innovation, Simula Metropolitan Center for Digital Engineering (SimulaMet), and UiT The Arctic University of Norway in Norway. Thirty credits of course work were taken at Simula Research Laboratory (University of Oslo), UiT The Arctic University of Norway, and OsloMet University, to fulfil the requirement for the PhD degree. This work was carried out between February 2018 to August 2021. The project was partially funded by the PRIVATON project (263248) from the Research Council of Norway (RCN). We performed experiments on the Experimental Infrastructure for Exploration of Exascale Computing (eX3) system, which is financially supported by RCN under contract 270053.

This dissertation is a collection of 21 published papers on automated gastrointestinal tract examination. We put the introductory chapters, background, and highlights of our work and bind them with the papers. I am the first author of 11 papers, equally contributed first author of two papers, second author of four papers and co-authors of the four other papers. We also have two other published works, and three submitted works that are not part of the thesis. I hereby declare that I am the author of this thesis.

Debesh Jha
Oslo, August, 2021

Acknowledgments

First and foremost, I would like to thank all my supervisors. I would like to express my sincere gratitude to my PhD supervisor Prof. Pål Halvorsen for his consistent motivation, encouragement, guidance, and support throughout the PhD journey, without whom this dissertation would not have been possible. He was very patient and ready to help any day, anytime. He has been consistently encouraging and guiding me till the last day of my thesis submission. I am very grateful to him for his scientific advice, knowledge, insightful suggestions, and support in every possible way throughout my PhD journey. I am blessed to have a Pål as my PhD supervisor, and PhD was comparatively easy working under his supervision.

I would like to express my sincere thanks to my supervisor, chief research scientist Michael A. Riegler, for his motivation, continuous guidance and support. He helped me figure out the relevant research direction, was always ready to discuss new ideas, and always remained curious about my research. Apart from the core research, he encouraged and helped me collect and release public datasets and actively organize challenges and competitions. I have benefited a lot from him. He was always ready to review my papers and provided critical feedback and argumentation that was very important.

I extend my sincere thanks to my supervisor, Associate Prof. Håvard D. Johansen, for funding me through the PRIVATON project. He continuously supported me in the research work and paper writing. I have learnt some useful writing tips from him. He has been very positive and kind to me and always ensured that I had all the research resources. He also helped me to take courses at the UiT The Arctic University of Norway and provided me with all other administrative support.

I would also like to express my sincere thanks to my supervisor, Prof. Dag Johansen, for his guidance and supervision. I would like to thank him for his constructive advice to improve the quality of the work. In general, he is very critical about scientific writing, and I have benefited a lot from it. He also helped me with funding at the last stage of my PhD.

I would also like to express my sincere gratitude to our medical collaborator Dr. Thomas de Lange, for his invaluable time and support. Without his support, dataset collection, curation, and annotation would not have been possible. Additionally, I would also like to thank Dr. Peter T. Schmidt for his help in the dataset annotation. I would like to thank all of my other medical collaborators from the different institutions in Europe and abroad for their valuable time and collaboration. Figure 1.1, Figure 1.2, Figure 2.1, Figure 2.2, Figure 2.3, Figure 2.4, and Figure 2.5 in the thesis has been acquired from shutterstock.com under standard licensing.

I would like to thank Dr. Sharib Ali from the University of Oxford, UK, for working as a research partner for more than two years. I would also like to thank all other collaborators from different universities in Norway and abroad for the research collaboration. Moreover, I would also like to particularly thank my wonderful colleagues Vajira Thambawita, Steven Hicks, Hanna Borgli, Konstantin Pogorelov, Pia Smedsrud, and Håkon Kvale Stensland. It was an honour to work with all of them in an inspiring atmosphere.

I would like to extend my special thanks to my great friend Dr. Ashish Rauniyar and Dr. Desta Haileselassie Hagos, for continuous motivation, encouragement, and support. We spent a lot of time together, and PhD journey was much relaxed to have a supportive and interesting friend like them. Likewise, I am truly indebted to my awesome friend Ramesh Pokhrel, Dr. Dipesh Pradhan, Dr. Sabita Maharjan, Anand Dev, Saruar Alam, Nikhil Kumar Tomar, Bineeth Kuriakose, and Sushmita Adhikari Pokhrel. Many other friends have been nice to me, however, it will be too many to list here. I want to sincerely thank every one of them.

Last but not least, I would like to acknowledge and thank my dad, Mr. Shardanand Jha, and my mother, Mrs. Lalita Devi Jha to whom this dissertation is dedicated. I am very thankful to my brother Vabesh Kumar Jha, my sister Manisha Kumari Jha, and Nishi Kumari Jha for their immense love, support, care and encouragement. All of them have supported me unconditionally to pursue my dream. Without their sacrifice and infinite support, it would never have been possible to complete this journey or any other endeavour in life. I also would like to thank my lovely wife, Ritika Kumari Jha, for her patience and sacrifice during the last part of my PhD thesis. Her consistent support has made it possible to finish this thesis. Thank you, all of you.

I am deeply thankful to God for the blessings and encouragement.

Debesh Jha

Oslo, August, 2021

Abstract

Gastrointestinal tract (GI) cancers are among the most common types of cancers worldwide. In particular, colorectal cancer (CRC) is the most lethal in terms of number of incidences and mortality (third most common cause of cancer and the second common cause of cancer-related deaths). Colonoscopy is the gold standard for screening patients for CRC. During the colonoscopy, gastroenterologists examine the large bowel, detect precancerous abnormal tissue growths like polyps and remove them through the scope if necessary. Although colonoscopy is considered the gold standard, it is an operator-dependent procedure. Previous research has shown large missing rates for GI abnormalities, e.g., polyp miss detection is around 22%-28%. Early detection of GI lesions and cancers at the curable stage can help reduce the mortality rate. The development of automated, accurate, and efficient methods for the detection of the GI cancers could benefit both gastroenterologists and patients. In addition, if integrated into screening programs, an automatic analysis could improve overall GI endoscopy quality.

The medical field is becoming more interdisciplinary, and the importance of medical image data is increasing rapidly. Medical image analysis can play a central role in disease detection, diagnosis, and treatment. With the increasing number of medical images, there is enormous potential to improve the screening quality. Deep learning (DL), in particular, convolutional neural network (CNN) based models have tremendous potential to automate and enhance the medical image analysis procedure and provide an accurate diagnosis. The automated analysis of the medical images could reduce the burden of the medical experts and provide quality and accessible healthcare to a larger population. In medical imaging, classification, detection, and semantic segmentation tasks are crucial for clinical practice. The development of accurate and efficient computer aided diagnosis system (CADx) or computer aided detection system (CADe) models can help to identify the abnormalities at an early stage and can act as a third eye for the doctors.

To this end, we have studied and designed machine learning (ML) and DL based architectures for GI tract disease classification, detection, and segmentation. Our designed

architectures can classify different types of GI tract findings and abnormalities accurately with high performance. Our contribution towards the development of CADe models for automated polyp detection showed improved performance. Out of three different medical imaging tasks, semantic segmentation of medical imaging data plays a significant role in extracting meaningful information from images by classifying each pixel and segmenting it by class. Using the GI case scenario, we have mainly worked on polyp segmentation and proposed and evaluated different automated polyp segmentation architectures. We have also built architectures for surgical instrument segmentation that showed high performance and real-time speed.

We have collected, annotated, and released several open-access datasets such as HyperKvasir, KvasirCapsule, PolypGen, Kvasir-SEG, Kvasir-instrument, and KvasirCapsule-SEG in collaboration with hospitals in Norway and abroad to address the lack of datasets in the field. We have devised several medical image segmentation architectures (for example, ResUNet++, DoubleU-Net, and ResUNet + CRF + TTA) that provided improved results with the publicly available datasets. Beside that, we have also designed architectures that have the capability of segmenting polyps in real-time with high frame per second (FPS) (for example, ColonSegNet, NanoNet, PNS-Net, and DDANet). Moreover, we performed extensive studies on the generalizability of our models on public datasets, and by creating a dataset consisting of data from different hospitals, we allow multi-center cross dataset testing. Our results prove that proposed DL based CADx systems might be of great assistance to clinicians in the future.

Contents

List of Figures

List of Tables

Acronyms

ADR adenoma detection rate. 3

AI artificial intelligence. 1, 15, 49, 85

AP average precision. 48, 56, 83

ASPP atrous spatial pyramidal pooling. 57, 58, 64, 65

CAD Computer aided diagnosis. 24, 28

CADe computer aided detection system. vii, viii, 1, 3, 6, 9, 14, 16, 84, 85, 87, 88

CADx computer aided diagnosis system. vii, viii, 1, 3, 5, 6, 9, 14, 16, 26, 28, 31, 33, 84, 87

CNN convolutional neural network. vii, 2, 8, 51, 52, 54, 83, 84, 88, 89

CRC colorectal cancer. vii, 2–5, 15–20, 26, 31, 62, 85

CRF conditional random field. 58, 59, 61, 83

DDANet Dual decoder attention network. 82

DL deep learning. vii, viii, 1, 2, 6, 9–11, 24–26, 33, 37, 49, 51, 74, 75, 77, 84, 87, 89

DNN deep neural network. 13, 80

DSC dice coefficient. 11, 26, 47, 65, 69, 72, 74, 75, 77, 81–83

ESGE European Society of Gastrointestinal Endoscopy. 20, 21

FDA Food and Drug Administration. 1

FIT Fecal immunochemical test. 20, 31

FOBT Fecal occult blood test. 16, 20, 31

FPS frame per second. viii, 11, 47, 48, 53, 56, 69, 72, 74, 75, 83

GF global feature. 11, 25, 51, 52, 82

GI gastrointestinal tract. vii, viii, x, xiii–xv, 2–11, 14–17, 20–26, 28, 30, 31, 33–35, 40–42, 45–49, 51–53, 76, 77, 79–84, 86–89

GLOBOCAN Global Cancer Statistic 2020. 3

MAP mean average precision. 48

MCC Matthews correlation coefficient. 48, 51, 53, 80, 82, 83

mIoU mean intersection over union. 26, 47, 48, 55, 56, 65, 69, 72, 74, 75, 81, 83

ML machine learning. vii, 1, 6–9, 11, 16, 24, 25, 28, 30, 34, 35, 37, 40, 47, 49, 51, 80, 82, 83

PRC precision recall curve. 80

ROBUST-MIS Robust Medical Instrument Segmentation. 73, 82, 84

ROC receiver operating curve. 62, 63, 80

ROI region of interest. 79

SOTA state-of-the-art. 11, 13, 24, 37, 63, 65, 82, 88

TTA test-time augmentation. 58, 59, 61, 63, 83

USPSTF United States Preventive Service Taskforce. 4, 18, 19

VCE video capsule endoscopy. 20, 21, 31, 35, 43, 49, 87

WCE wireless capsule endoscopy. 6, 72

Chapter 1

Introduction

Advancements in computer vision [65, 107, 185, 178] hold a great promise in the area of medical image analysis. Deep learning (DL) has shown promising potential to impact healthcare technology [12, 29, 38, 44, 59, 62, 78, 102, 104, 121, 136]. The US Food and Drug Administration (FDA) has already approved some artificial intelligence (AI) based medical devices and algorithms [16]. The improvement in the data collection procedure, storage capability, and advancement in the computation resources has further increased the potential of machine learning (ML) in medicine. In the field of medicine, ML could be used from diagnosis (mostly), prognosis, therapy, drug development, and epidemiology. Additionally, ML could be leveraged for providing clinical risk prediction to improving safety [15]. The common examples of ML algorithms that act as complementary to the physicians are Lymph Node Assistant (LYNA) [122], and Deep Learning based Automatic Detection (DLAD) [137]. Thus, the recent studies show that the imaging based DL algorithms have the potential to improve the physicians accuracy [29].

With the success of ML in medicine, it is plausible to predict that ML will have a tremendous clinical impact. The ML models, when integrated with the computer aided detection system (CADe) or computer aided diagnosis system (CADx) system, will provide an opportunity to improve the clinicians' decision in medical image interpretation and act as a second observer. The medical experts could double-check their diagnosis and interpretation results. It could help in early diagnosis and also help to reduce mortality, improving the overall performance of the physicians. Thus, it will also enhance the clinical workflow ensuring the patient's safety. Therefore, there is a growing demand for automated systems in clinics to fulfil the need for medical experts. The ML based systems integrated into the clinical setting could provide benefits such as the opportunity of accessibility to all, especially in countries with low and middle income [203, 219]. The

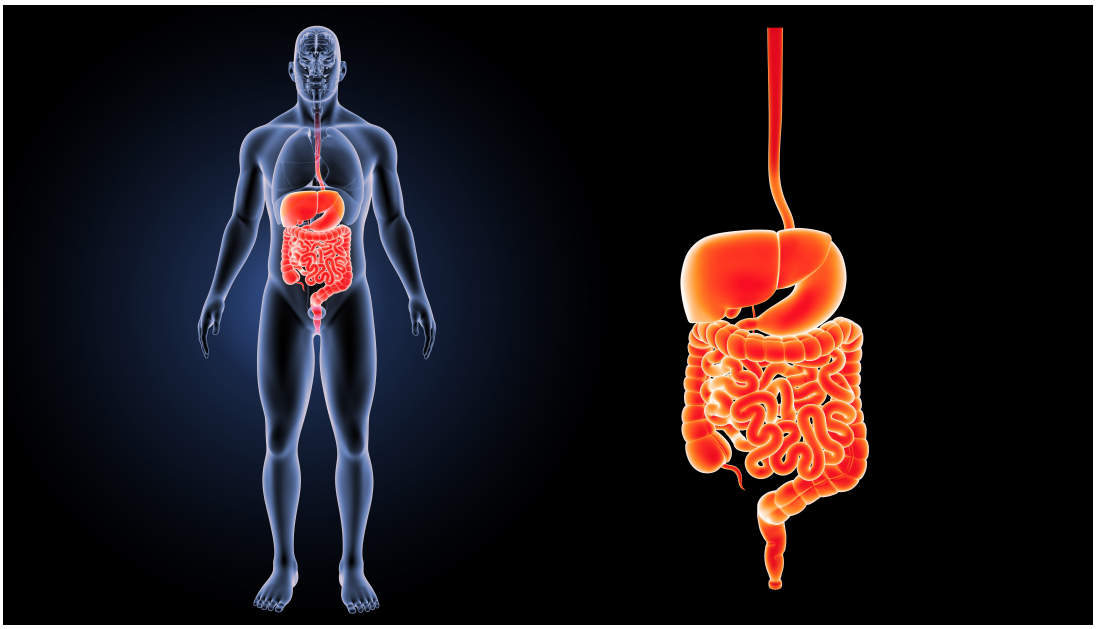


Figure 1.1: Diagram of the human gastrointestinal tract

successful automated systems are faster in image interpretation, easy to scale, efficient, cheaper, attentive, ethical, and produces reproducible results.

We perform research in the intersection between DL (in particular, convolutional neural network (CNN)) and medicine. Our research is mostly focused on the human gastrointestinal tract (GI) tract. Figure 1.1 depicts the human GI tract. The diagnosis of the GI tract is important because it is the potential source of a large number of lethal GI cancers. Early precursors of cancer are often missed during examination until they reach a late stage [198]. Early detection of such GI cancers (esophagus, stomach, colorectal) and diseases (for example, gastroesophageal reflux disease, peptic ulcer disease, inflammatory bowel disease) can provide an opportunity for early treatment and help to reduce the mortality rate. The survival from GI cancers (colorectal cancer (CRC)) have shown a significant impact on quality of life [83, 32]. Therefore, research in gastroenterology is of significant importance.

We have develop automated methods for GI tract findings classification and detection, colorectal polyp segmentation, and surgical instrument segmentation in laparoscopy. Additionally, we created and released GI tract (upper endoscopy and colonoscopy) datasets publicly for academic and research, reproducibility, and commercial purposes. At first, we developed automated methods for GI conditions classification and detection. Then, we focused our research on semantic segmentation, where we designed automated algorithms specifically for colorectal polyp segmentation. We have mostly researched and designed methods for medical image segmentation. The main reason for concentrating our research

on medical image segmentation is that it plays a crucial role in CADx and CADe systems. It is considered as an essential tool in modern clinical applications [160]. We have also researched and built automated methods for surgical instrument segmentation in laparoscopy at the later stage.

In this thesis, we first introduce the medical problem related to the GI tract. Next, we present the related research work and highlight the existing problems and technology gap. Then, we focus on the created datasets in the fields of GI endoscopy and colonoscopy. After that, we present concrete details of our methods to fulfill such technology gaps. Following that, we discuss our contributions based on the objectives and how we achieved the primary goal. We also discuss the strength of our methods, possible limitations and challenges, and progress in the CADe and CADx systems made so far. Finally, we conclude and discuss future directions for the research.

In the first chapter, we briefly describe the GI tract and cancer related to the GI tract. We introduce the current technology used in the diagnosis and treatment of anomalies both in the upper GI tract and lower GI tract. We also highlight the current challenges in the upper GI endoscopy and colonoscopy. After highlighting the problem next, we introduce the basis of our work by introducing our research aim and objectives. We also demonstrate the research methodology used in our study. We sketch the main contribution that will be explained in detail in the latter part of our dissertation.

1.1 Background and motivation

The GI tract is a contributing factor to a large part of cancer-related death worldwide. Global Cancer Statistic 2020 (GLOBOCAN) 2020 [183] estimated that there were around 19.3 million new incidences of cancer cases and approximately 10.0 million cancer-related deaths worldwide. The most frequently diagnosed cancers are female breast cancer (11.7%), lung cancer (11.4%), CRC (10.0%), prostate cancer (7.3%), and stomach cancer (5.6%). The leading cause of death-related cancer is lung cancer (18%), CRC (9.4%), liver cancers (8.3%), stomach cancers (7.7%), and female breast cancers (6.9%). Cancer is globally estimated to increase by 47%, which will be around 28.4 million. From the GLOBOCAN 2020 statistics, we observe that GI tract cancers, which comprise both CRC and stomach cancers, are the leading cause of cancer-related death and is a significant hurdle to the increasing life expectancy. Increasing Adenoma detection rate (ADR) can help to prevent CRC [37]. Additionally, CADe can be safe and effective [161]. Therefore, significant efforts should be applied to build digital infrastructures for the early detection

Figure 1.2: Examination of the upper GI tract using an upper GI Endoscopy (EGD)

and prevention of all types of cancers, including GI cancers, and CRC, for controlling cancer globally.

The examination of the GI tract is essential for investigating and finding abnormalities in the GI tract. Early GI cancer precursors are usually missed during the gastroscopy or colonoscopy. Such miss rate can be relatively high, which at later stages leads to mortality. The main reason for the higher miss rate is the lack of attentiveness of the gastroenterologists while performing the tiring task. Also, the gastroenterologists have to perform several examinations on the same day that may decrease their attention on the individual patient. Additionally, the manual examination of GI tract using gastroscopy or colonoscopy is a challenging process requiring high overall cost, unpleasant experience for the patient while undergoing the test, and consumption of a lot of hospital resources and task force that could have been alternatively beneficial for the other helpful tasks.

Lower GI endoscopy allows the gastroenterologists to observe the lower GI tract. It is used for screening of both colon and rectum. Colonoscopy is the inspection of the whole large bowel. According to the United States Preventive Service Taskforce (USPSTF) for CRC screening, forty-five is considered to be the new fifty for screening of CRC [138]. Previously, many studies including guidelines from USPSTF [117] and European Union [58] recommended 50 years of age for CRC screening. CRC is the third most frequently di-

agnosed cancer type and the second most common cause of cancer-related death in the United States when both male and female populations are combined [176]. Fortunately, CRC is among the preventable malignancies. Early detection of the CRC is possible with the available screening tests. Diet and lifestyle are mostly linked to both CRC occurrence and mortality. Thus, a change in lifestyle and diet could reduce the risk of CRC [138].

To summarize, there is a shortage of qualified gastroenterologists all around the world. A procedure such as a colonoscopy is an operator-dependent procedure. The polyp miss-rate is reported up to 22%—28%, and adenoma miss-rate is reported to be 20%—24% [111]. However, there is variability among polyp miss-rate and adenoma miss-rate in literature because of the nature of study and size of patient cohorts [1, 111, 162, 205, 206], most of the work report similar metrics. Undergoing the colonoscopy requires an adequate amount of time for endoscopists and nurses. Despite being an uncomfortable procedure, finding abnormalities in the bowel depends upon the endoscopist's ability [97]. Additionally, certain factors can influence the quality of the colonoscopy procedure [202]. The variability in the quality can lead to differences in the endoscopist's performance. Moreover, it is an expensive procedure. Therefore, there is a need for the development of automated methods for CADx based support system for upper GI and lower GI tract. The development of such a system will improve patient outcomes. The GI examination will be more accessible to patients, potentially cost-effective, and make public healthcare more effective.

1.2 Research aim and objectives

The primary objective of this dissertation is to develop automated algorithms for the analysis of GI images and videos from GI endoscopy. Most of the works in this dissertation have been targeted towards designing automated methods for segmentation and detection of CRC. We aim to develop real-time automated methods that can provide feedback to the gastroenterologists about the exact location and position in the colon (in the lower GI tract) that can act as a second pair of eyes to the doctors and help reduce polyp miss-rate. Similarly, we also aim to develop automated classification methods for whole GI that can automatically identify and classify each disease of the GI tract. Additionally, we aim to develop automated methods for surgical instrument segmentation in laparoscopy that could assist in accurate tracking of the medical instruments. Our other research aim is to generate and publicly release GI endoscopy and colonoscopy datasets to address the challenges with the lack of dataset in the field.

In this dissertation, we put forward our computer vision and DL approaches to address the research question. Automation in the field of GI endoscopy and colonoscopy can help reduce the miss-detection of the abnormalities in the endoscopic procedure. Additionally, it has several other benefits, such as reducing the mortality rate, cost-effective, faster, accessible, and ensuring better trust and safety. Automated methods for CADe and CADx might have significant societal impact, and it will improve patient care. Based on our research goal, we define the objective of this dissertation. We have considered three main objectives to meet the main research goal. Below we highlight our main research goal and objectives:

Main goal: The main goal of our research is to design automated classification, detection, and segmentation algorithms for CADe and CADx system for examination of GI tract findings. The developed algorithms should automatically identify, detect, and segment the suspicious lesion of the GI tract acquired through standard endoscopy, colonoscopy, or wireless capsule endoscopy (WCE) with high accuracy and real-time processing speed. Additionally, we also aim to develop semantic segmentation architectures for surgical instrument segmentation. Overall, our goal is to design lightweight architectures, achieve high performance across several datasets, extendable, and be easily integrated with endoscopic devices.

- **Objective I**: Research, collect, and construct new datasets in the field of GI endoscopy and colonoscopy to address the lack of dataset problem in the field.

- **Objective II**: Explore, investigate, and design ML and DL methods for GI tract findings classification, and polyp detection.

- **Objective III**: We aim to design new medical image segmentation architectures to address the need for efficient algorithms in colorectal polyp segmentation, surgical instrument segmentation, and general medical image segmentation tasks. Additionally, we aim to design real-time segmentation architectures for polyp and surgical instrument segmentation. Moreover, we aim to explore and improve the generalizability and robustness of the DL models on publicly available independent and multi-centre colonoscopy datasets. Furthermore, our research is focused on designing segmentation algorithms to identify and segment different types of polyps, including flat and sessile polyps that are commonly missed in the colonoscopy examination.

1.3 Scope and limitations

The scope of this thesis lies in three sub-folds. In the first part of the thesis, we limit our research to build automated models for the classification of GI tract findings from both upper GI tract (anatomical landmarks, pathological findings) and lower GI tract (anatomical landmarks, pathological findings, quality of mucosal views, and therapeutic findings). We do not further focus on the classification of certain classes of polyps (for example, non-neoplastic (hyperplastic polyps, inflammatory polyps, and hamartomatous polyps) and neoplastic (adenomas and serrated polyps)) due to the lack of publicly available dataset in the field. In the second part of the thesis, we mainly focus our work on polyp detection. In the final part of the thesis, we limit our research to polyp segmentation, and the same architectures were used for surgical instrument segmentation. We have further used other biomedical datasets to demonstrate that our methods are not only limited to automatic polyp segmentation. However, the developed models could be useful to detect other abnormalities and pathological findings in the GI tract if the ground truth or bounding box information of the dataset is available.

Our methods are only tested on the publicly available datasets, our own datasets, or the dataset shared through the challenges and competitions. There might be presence of inconsistencies in the dataset collection. For example, there might be different lighting conditions and different resolution images. In addition to these, the expertise of gastroenterologists is operator-dependent. Additionally, the hospitals have their own standard for dataset collection. For example, in the Vestre Viken trust, the gastroenterologists save short video clips of meaningful findings along with the frames. However, in Karolinska, the medical experts only capture the most valuable findings in the colonoscopy or upper endoscopy [163]. Although, we have introduced the usefulness of generalizability in the field of colorectal polyp segmentation and GI tract, more research and public datasets are needed to explore in the field of generalizability, robustness, and interpretability of the ML models for building trustworthiness of our models.

1.4 Research methodology

In 1989, the Association for Computing Machinery (ACM) Education Board approved and endorsed a report for release [40]. The final report from the Task Force on the core of computer science puts forward a novel intellectual framework that determines the criteria, discipline, and norms of computing and the basis upon which the computing curricula

can be based. Computing is defined as a crossway between applied math, science, and engineering. All of these three processes are essential in the discipline. The foundation of computer science is built from a wide variety of disciplines. The concept in computer science is extracted from various fields. Thus, computer science combines the processes such as theory, abstraction (in general), and design (specific) [42].

In this dissertation, our work is related to these topics in several ways. In the below section, we will describe each process and discuss how our dissertation covers these topics.

Theory Theory is rooted in mathematical aspects and involves the development of valid theory. The report categorized theory as (i) characterize objects of study (definition), (ii) hypothesize possible relationships among them (theorem), (iii) determine whether the relationships are true (proof), and (iv) interpret results.

Our thesis is related to the theoretical part in the context of different medical image processing techniques in 2D geometry, including image and video processing, 2D vector-based geometric operations, data structures, algorithms, linear algebra, and calculus, statistics. All of these are used for building training and testing of the CNN architectures and ML algorithms.

Abstraction Abstraction process is used for modeling and is directly related to the experimental scientific method. The report describes the abstraction process as (i) form a hypothesis, (ii) conduct a model and make a prediction, (iii) design an experiment and collect data, and (iv) analyze results.

Our thesis has performed several experiments using different existing and new datasets to support the hypothesis. We have performed several experiments within our research group, collaborated with external partners, and performed experiments for relevant open challenges and competitions. We have collected and annotated several datasets with assistance from gastroenterologists that can be useful for the medical image segmentation, detection, and localization tasks in the field of GI endoscopy. We explore image preprocessing, feature extraction, and multi-class classification using ML algorithms. Additionally, we have built semantic image segmentation architectures and classified each pixel of the images into different classes. We have also shown that our proposed architectures can be extended to other medical image segmentation tasks. We have evaluated the performance of our methods using standard computer vision metrics such as accuracy, sensitivity, recall, precision, dice coefficient, Jaccard index, and fps. Furthermore, we analyze the each image carefully and identify the easy, mild, and challenging cases. We

primarily focus on the challenging cases that are medically relevant to address.

Design The report describes the design into four steps: (i) state requirements, (ii) state specifications, (iii) design and implements the system, and (iv) test the system.

We have designed, experimented, and implemented, several ML and DL based architectures for automatic GI tract classification, automatic polyp detection, medical image segmentation architectures for automatic polyp segmentation, and architectures for surgical instrument segmentation. We have used the layers such as strided convolutions, dilated convolutions, transpose convolutions, and bilinear upsampling during the architectural design. Similarly, we have used various spatial and channel-wise attention mechanisms, residual blocks, autoencoders, pre-trained encoders, and multi-scale fusion. Our architectural designs are based on performance and speed. Our classification, detection and segmentation architectures can classify, detect, and segment objects of interest with high accuracy and real-time processing speed. Besides high accuracy on the real-world dataset (GI tract dataset), we have also further explored the generalizability capability of the proposed models with the datasets from different centers. Our proposed architectures are tested on more than one dataset, including public datasets and our own collected datasets in most cases, to show that our end-to-end architectures perform well not only on one dataset but also perform across different imaging datasets.

In summary, we aim to solve a real-world problem related to GI diseases classification, detection and segmentation. The CADe and CADx can solve the current challenges of miss-detection in the field and can create a significant impact in the current healthcare system. Our goal is to provide technology to assist clinicians for improving the health care system based on their requirements of higher accuracy in real-time speed. To achieve these, we have collected many GI tract datasets, curated and annotated the ground truth and bounding boxes and proposed different ML and DL based architectures for each task. Our models showed promising results with the real data obtained from different endoscopic equipment. Thus, our architectures could be tested by clinicians to verify their usefulness in the clinical setting. Moreover, our architectures can be extended to other medical image analysis tasks and natural image segmentation tasks.

1.5 Main contributions

In order to meet the goals and fulfil the objectives of the thesis, we have researched several challenges and addressed various issues in the domain of medical image classification,

detection and segmentation. In particular, the main contributions are as follows:

- We have created and released the Kvasir-SEG [89] dataset. The dataset is publicly available for academic, research, and industrial purposes (with prior consent). In addition, the dataset can be used for automatic polyp detection, localization, and polyp segmentation task (objective I).

- We have identified sessile or flat polyps $\leq 10\,\text{mm}$ from the Kvasir-SEG dataset with the help of a senior gastroenterologist and made it available separately on the same webpage [85]. This was done to test the performance of DL models separately on such polyps that are commonly overlooked during colonoscopy examination. Designing specific models for such datasets can help reduce polyps miss-detection. We have shown that our method had a better polyp segmentation capability for flat or sessile polyps (objective I).

- We have publicly released the HyperKvasir [26] and the KvasirCapsule [180] datasets. To the best of our knowledge, HyperKvasir is the most comprehensive medically verified publicly available GI tract classification dataset and KvasirCapsule is the largest video capsule endoscopy classification dataset in the world (objective I).

- We have created and released the Kvasir-Instrument [88] dataset. The dataset includes images, corresponding ground truth, and bounding box information of the therapeutic and diagnostic tools used in GI endoscopy. To the best of our knowledge, Kvasir-Instrument is the first public dataset of segmented diagnostic and therapeutic tools in the GI endoscopy (objective I).

- We have curated and released a multi-center (6 centers) polyp segmentation and detection dataset [5]. To the best of our knowledge, this is the most comprehensive publicly available polyp detection and segmentation dataset annotated by a team of experts gastroenterologists and computer scientists. We conjecture that the dataset will be useful to address the generalizability issue in polyp segmentation and detection tasks (objective I).

- We have annotated and released a wireless video capsule endoscopy polyp dataset [92] (objective I).

- We have organized competitions such as Medico Automatic Polyp Segmentation [91], Endotect Challenge 2020 [71], and Endocv 2021 Challenge[1]. In these challenges, we

[1] `https://endocv2021.grand-challenge.org/`

have annotated and released complete training and testing datasets or test datasets. Moreover, we have evaluated the participants' methods for each task and provided the ranking. The datasets for each challenge are made publicly available (objective I).

- We have proposed solutions based on global features (GFs) and ML (simple logistic classifier and logistic model tree) for multi-class classification [193]. Additionally, we have done an extensive study on the cross-dataset bias on the GI tract abnormalities classification using GFs and ML based approaches on several publicly available datasets. We have emphasized the use of cross-dataset evaluation to demonstrate the generalizability of DL models before using them in clinical applications [190] (objective II).

- We have done a comprehensive analysis of classification methods in terms of accuracy and efficiency presented in the various competitions, calculated the clinical applicability of the participant's methods, and ranked them based on robustness and speed[84] (objective II).

- We have trained several popular detection algorithms with various backbones and compared them with our method for real-time polyp detection (ColonSegNet [93]). We established a new detection benchmark for the Kvasir-SEG dataset (objective II).

- We have developed new segmentation methods such as ResUNet++ [94], DoubleU-Net [86], and ResUNet++ + TTA + CRF [85], and improved the state-of-the-art (SOTA) results in the publicly available polyp datasets. We have shown improvement over previous SOTA semantic segmentation methods both in terms of dice coefficient (DSC)s and in terms of frame per second (FPS) (objective III).

- We have developed new segmentation models (such as NanoNet [92], ColonSeg-Net [93], DDANet [195], and Progressively Normalized Self-attention Network (PNS-Net) [95]) that can segment polyp in real-time with high FPS and high accuracy (objective III).

- We have explored and introduced generalizability in polyp segmentation to the best of our knowledge [85, 5] and proposed a Ensemble MultiResUNet [197] based method to improve generalizability in polyp segmentation (objective III).

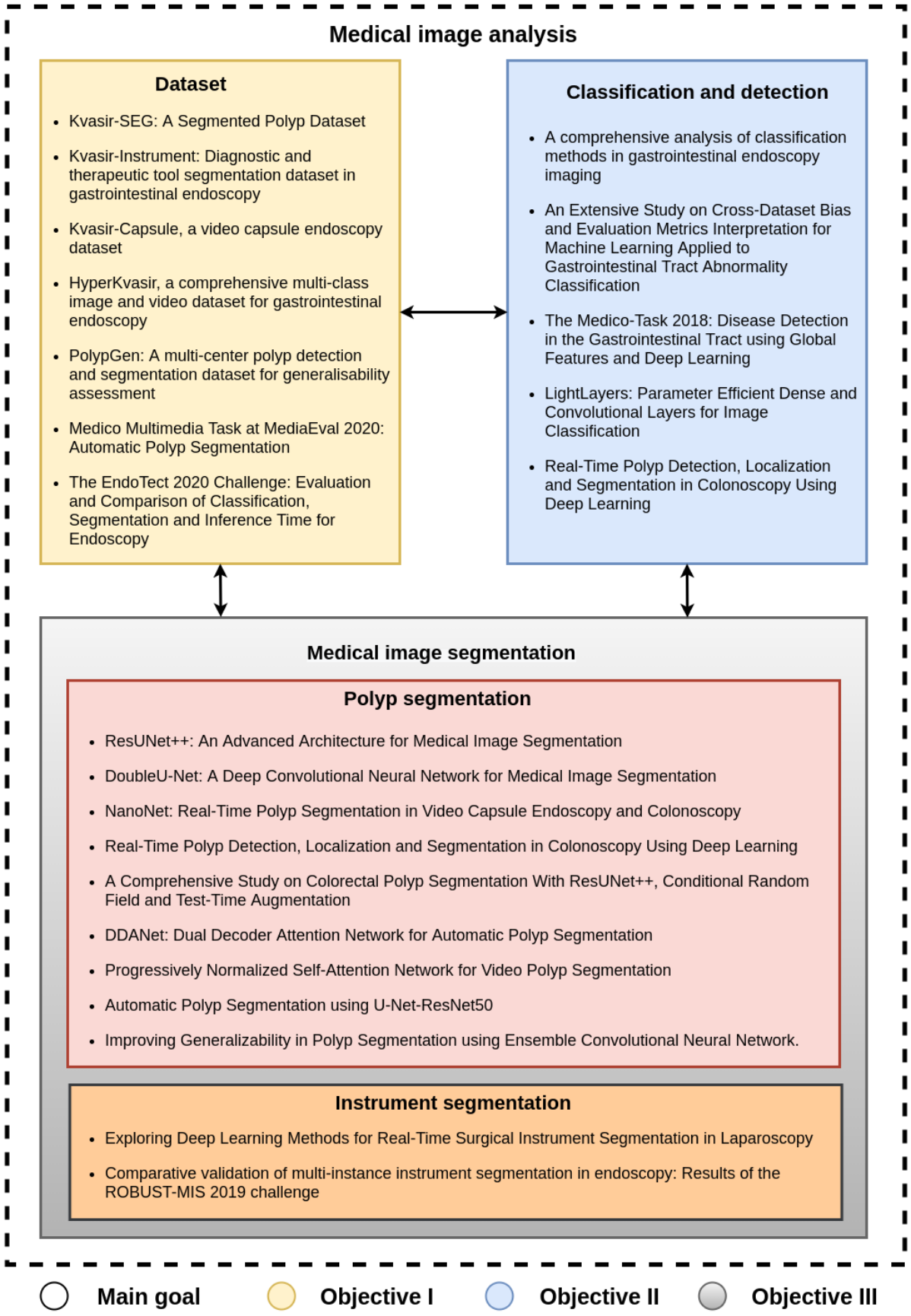


Figure 1.3: Contributions related to objectives

- We have provided a new benchmark on several polyp datasets (for example, Kvasir-SEG [89] (both segmentation and detection), Kvasir-Instrument [88] (segmentation), Kvasir-Sessile (segmentation), KvasirCapsule-SEG (segmentation), HyperKvasir [26] (classification), and Kvasir-Capsule [180] (classification) dataset (objective III, objective II).

- We have studied and proposed solution for surgical instrument segmentation [168]. Additionally, we have also provided a comprehensive study on the SOTA methods for real-time semantic segmentation of surgical instruments in laparoscopy [87] (objective III).

In Figure 1.3, we have listed our papers based on objectives. All of these papers are attached in the appendix. Here, we show that our main goal is medical image analysis, and we aim to achieve it through dataset collection (objective I), development of classification and detection algorithm (objective II), and development of medical image segmentation architectures (objective III). The arrows in the figure, shows that each of the objectives are interconnected.

Additional contributions: Additionally, we have also carried out research related to but outside the main topic of the research. This includes work on sports analytics [192], health monitoring [52], and general medical image segmentation architecture [3]. Additionally, we proposed Lightlayers [90], a method for reducing the number of trainable parameters in deep neural network (DNN). Lightlayers showed promising results for image classification datasets.

In addition to all other contributions, we have supervised master students, bachelor's students, and summer interns based on my thesis project. We have organized workshops such as "Medico automatic polyp segmentation challenge[2]", "Endotect 2020[3]", "Endocv2021[4]". I have been part of several program committees as members, chairs, and reviewers for related conferences and journals. I have reviewed over 75 research articles. Additionally, we initiated collaboration with researchers from the University of Oxford (UK), Inception Institute of Artificial Intelligence (UAE), University of Bergen (Norway), University of Oslo (Norway), Østfold University College (Norway), SINTEF Digital (Norway), Ambroise Paré Hospital (France), Istituto Oncologico Veneto (Italy),

[2] https://multimediaeval.github.io/editions/2020/tasks/medico/
[3] https://endotect.com/
[4] https://endocv2021.grand-challenge.org/

Centro Riferimento Oncologico (Italy), Oslo University Hospital(Norway), John Radcliffe Hospital (UK), Indian Statistical Institute, and University of Alexandria (Egypt).

1.6 Dissertation outline

This dissertation is written as the collection of the 21 research articles. The rest of the dissertation is organized as follows:

- Chapter 2 provides an overview of the different types of the endoscopic procedure and challenges associated with detecting each type of findings. It also discusses the progress and shortcomings of the earlier research that motivates us for further research.
- Chapter 3 presents the design, curation, annotation, and release of the novel dataset to foster research in the field of GI endoscopy.
- Chapter 4 provides an overview of the CADx and CADe based system for the classification, detection, and segmentation of GI findings and for the different types of polyp segmentation problem. In this chapter, we also present our main results. It highlights the benefits of the proposed algorithms and also provides clinical contributions for some of the methods.
- Chapter 5 discusses the contribution based on each objective and shows how it helped achieve our main goal.
- Chapter 6 concludes the dissertation and suggests future work.
- In Appendix A, we include all of the published research articles that are part of this Ph.D. dissertation and highlight the independent contributions of the author.

Chapter 2

Background

AI based automation has the potential to improve the traditional medical systems. With the implication of AI in our medical systems, healthcare can be made more efficient. Automated medical systems could be helpful in the detection of cancer or other abnormalities in the human body at an early stage. In this chapter, we will provide an overview of GI tract and introduce different existing examination procedures. Additionally, we will also briefly highlight the current existing works on GI image classification, polyp detection, polyp segmentation, and surgical instrument segmentation. We also discuss the current challenges and shortcomings and finally close the chapter with a summary.

2.1 Gastrointestinal tract examination

The human GI tract compromises from the mouth to anus. It consists of the digestive system organ, including the mouth, pharynx, esophagus, stomach, small intestine, large intestine, and anus. The GI tract is the source of multiple types of abnormalities and cancers. Global Cancer Statistics 2020 (GLOBOCAN) [184] estimates that out of 19 million new cancer incidence and approximately 10 million cancer related deaths, CRC accounts for 10% of the cancer incidence, and stomach cancer accounts for 5.6% for the new cancer incidence. Similarly, CRC is the second leading cause of death, accounting for 9.4% of the total death, and stomach cancer accounting for 7.7% of the total cancer deaths. CRC, stomach cancer, esophageal cancer, and bladder cancer are the most 10 occurring cancer in 2020 both by incidence and mortality. By 2040, global cancer is estimated to be around 28.4 million, a 47% hike from 2020. From the statistics, it is evident that GI cancer is a threat to human lives. Therefore, significant effort should be applied by the research community, medical doctors, and computer scientists for the early

detection and possible prevention of the GI cancers.

2.1.1 Examination procedures

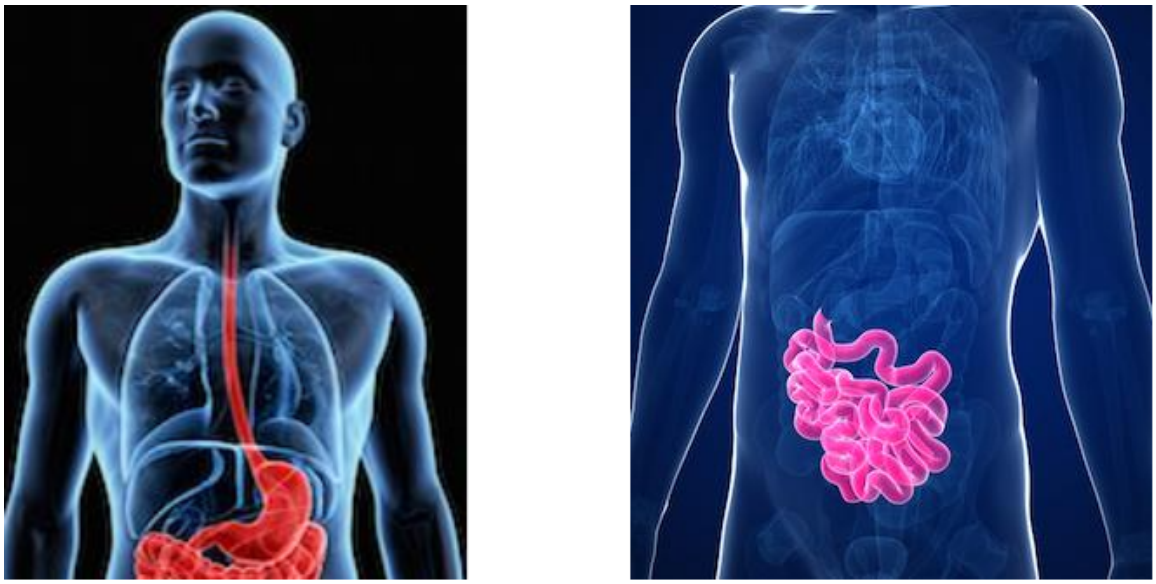
Upper endoscopy

Upper GI endoscopy is sometimes referred to as gastroscopy or esophagogastroduodenoscopy. It is a widely used procedure for diagnosing and treating anomalies in the upper GI tract. The upper GI tract includes the esophagus, stomach, and duodenum. Figure 2.1 (a) shows esophagus and duodenum whereas Figure 2.1 (b) shows small intestine from the upper GI tract. Upper GI endoscopy is carried out with the help of a thin, long, and flexible tubular instrument called an endoscope. The endoscope is used to monitor the esophagus, stomach, and duodenum. The video of the whole endoscopy process can be monitored on the screen.

One of the potential solutions to reduce the miss detection of the GI abnormalities is the development and integration of CADe and CADx system in the clinical workflow. However, a significant effort has to be applied to build ML based methods that work well to detect and evaluate various GI abnormalities. These methods must be clinically relevant. The developed methods should be generalizable to develop clinically relevant systems concerning patient variability and cohort population. Additionally, the system or the software must have real-time processing performance, and the lightweight system should be designed so that it can be integrated into the embedded system or endoscopic tool.

Sigmoidoscopy

Sigmoidoscopy is an invasive test. It is used to examine the inner lining of the rectum and the lower part of the colon. It is one of the options for CRC screening, like a colonoscopy. Sigmoidoscopy has a high sensitivity similar to colonoscopy and can be used for lesion removal. The bowel preparation is less complicated than colonoscopy, and it usually does not require sedation [115]. It is two types: flexible sigmoidoscopy and rigid sigmoidoscopy. The difference between flexible sigmoidoscopy and rigid sigmoidoscopy is that a flexible sigmoidoscopy has a flexible endoscope, whereas the other has a rigid device. Sigmoidoscopy covers only the lower part of the rectum, whereas colonoscopy is used to examine the entire colon. Holme et al. [74] explained the effect of flexible sigmoidoscopy screening on the Norwegian population with respect to CRC incidence and mortality. They concluded that flexible sigmoidoscopy and Fecal occult blood test (FOBT) reduced



(a) Esophagus, duodenum (b) Small intestine

Figure 2.1: Figure shows (a) esophagus, duodenum, and (b) small intestine from upper GI tract

Figure 2.2: Colon

the mortality rate and incidence of CRC as compared to the ones who did not undergo the screening test.

Colonoscopy

Colonoscopy is an invasive medical method where gastroenterologists operate on the colon by using a flexible tube (colonoscope) to examine the presence of abnormalities. It is considered the gold standard for the examination of the colon. An illustration of the colon, colonoscope, and colonoscope being inserted into human colon can be found in Figure 2.2,

Figure 2.3: Example of a colonoscope

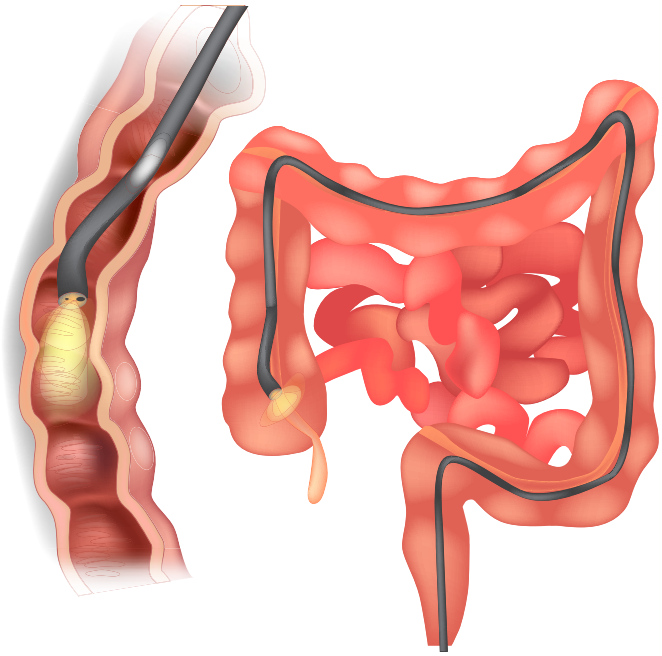


Figure 2.4: A colonoscope being inserted into the colon

Figure 2.3, and Figure 2.4. If the abnormalities such as polyps are detected during a colonoscopy, polypectomy is performed, and the polyp tissues are removed. Figure 2.5 shows an example of colon polyp removal using a colonoscope. Although colonoscopy is a gold standard, the examination procedure is not perfect even when performed carefully [162]. A study showed that, on average, a quarter of polyps were missed [111]. Early detection of CRC is possible, and there is a huge opportunity to decrease the mortality rate due to CRC [96, 175].

CRC is usually diagnosed among the population between 65 to 74 year [81]. USPSTF

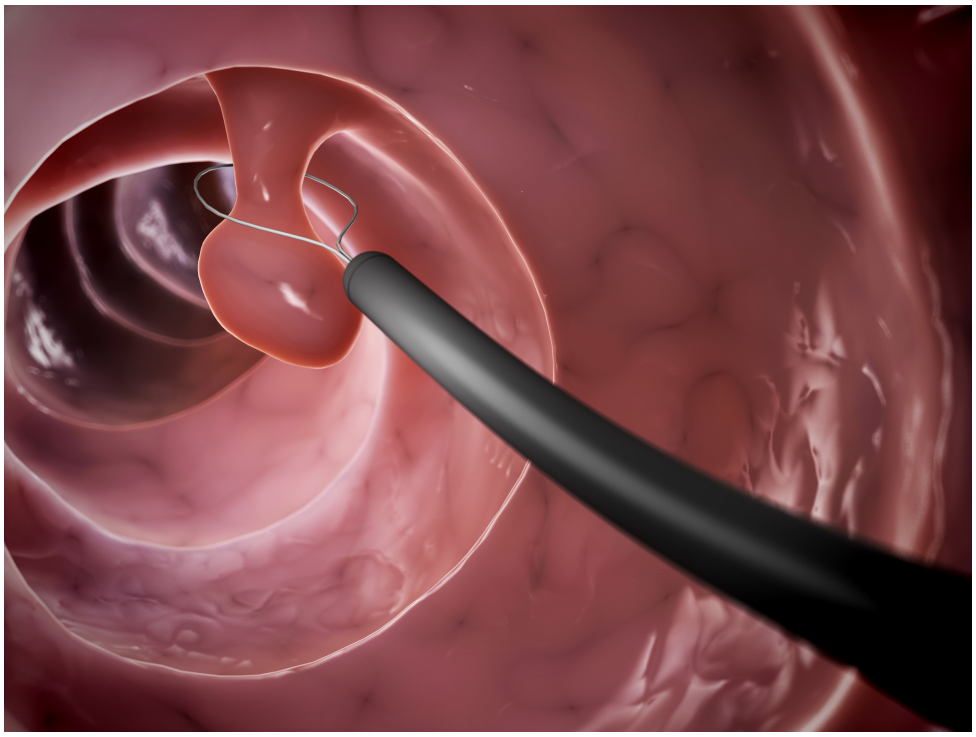


Figure 2.5: An example showing colon polyp removal using a colonoscope

recommends screening of CRC in between 50 to 75 years of age for all the population. However, it is estimated that 10.5% of the CRC incidence will occur in a population younger than 50 years [175]. Another recommendation of USPSTF is screening for CRC for the age group between 45 to 49 years. There is convincing evidence found by the USPSTF that screening for CRC through the different available technologies can detect early stage CRC screening, including adenomatous polyps [23]. Moreover, it is also suggested to undergo colonoscopy at a frequency of 10 years [48].

Colonoscopy is a time-demanding procedure requiring a significant time of the gastroenterologists and nurses. The process is troublesome and can cause significant discomfort for the patient. There is a high cost related to the procedure. Additionally, colonoscopy is an operator-dependent procedure, where the detection of abnormalities such as a polyp lies in the expertise of gastroenterologists. There is a certain bowel preparation procedure for undergoing colonoscopy. The patient has to change their diet and has to take medicine for causing diarrhea. Additionally, during the colonoscopy examination, sedation may be used that can cause minor or major complications [47]. For some complications, there is a higher risk if colonoscopy is performed other than the screening procedure [210]. Despite of the disadvantages, inconveniences, and high costs, colonoscopy still remains the primary screening tool for CRC [158, 63, 214]. We, therefore, aim for a system that identifies colon findings accurately and assists gastroenterologists

Figure 2.6: Olympus EC-S10 endocapsule

during the live examinations.

Fecal Occult Blood Test

The FOBT is a non-invasive test. As colonoscopy is costly, FOBT is preferred for screening CRC, especially in developing countries. Like colonoscopy, it is reported that FOBT has also reduced CRC related mortality and incidence. More detail about the FOBT can be found in this work [113].

Fecal Immunochemical Test

The Fecal immunochemical test (FIT) is a non-invasive test. It is also used for screening CRC. FIT is a low cost solution. FIT has high diagnostic accuracy for CRC screening [101]. A recent study also suggests that for the detection of CRC and advanced adenomas, FIT was higher than sigmoidoscopy when the repeated FIT was performed [179].

Video capsule endoscopy

Capsule endoscopy has transformed small bowel imaging [80]. European Society of Gastrointestinal Endoscopy (ESGE) strongly recommends video capsule endoscopy (VCE) as the first-line test for the patients having obscure GI bleeding. Similarly, it also recommends VCE as a diagnostic modality for investigating small bowel as an alternative test. Using only standard endoscopy to regularly screen a population is impossible due to

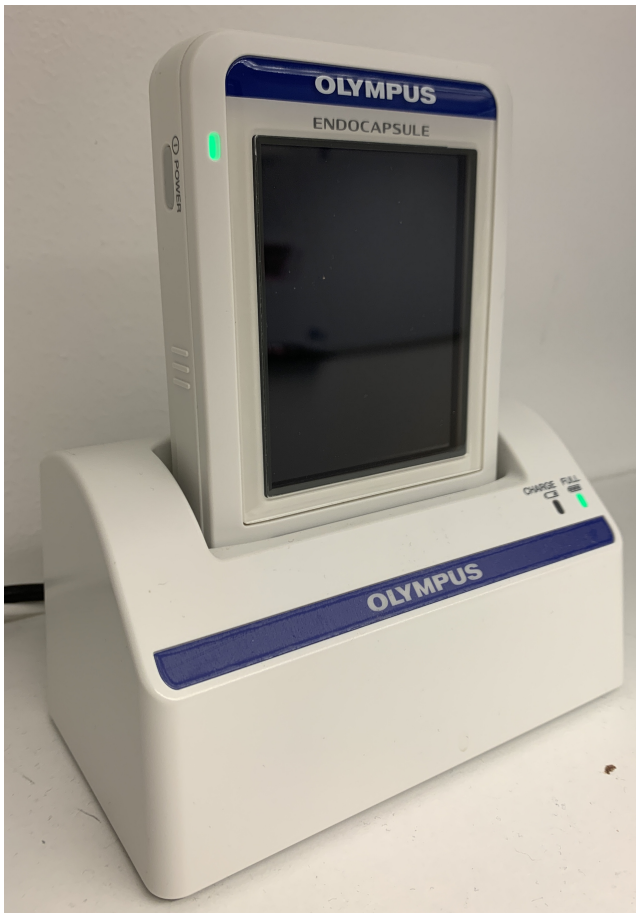


Figure 2.7: Olypmpus RE-10 endocapsule recorder used in our data collection for Kvasir-Capsule [180]

socioeconomic perspective [211]. Furthermore, medical screening used to identify undiagnosed diseases in large populations is debated with known problems like too many false positives, extensive over-diagnosis of diseases that would otherwise clinically not emerge, invasive screening procedures, and high costs. ESGE recognizes VCE as a complementary strategy. A solution that makes it possible to conduct large-scale screening of small bowel diseases in terms of cost-effectiveness and quality is VCE.

A VCE, also often called a camera pill, is a small capsule type device (typically $11mm \times 25mm$) having an image sensor, bleeding sensor, pH-sensor, antenna, battery, light source, and wireless transceiver. Figure 2.6 shows a pill cam. It is swallowed to visualize the GI tract for subsequent diagnosis and detection of GI diseases. Usually, a system such as Olympus Endocapsule 10 System [143] that includes Olympus EC-S10 endocapsule (see Figure 2.6) and Olympus RE-10 endocapsule recorder (see Figure2.7) are used during the examination and for dataset collection (Kvasir-Capsule [180] in our case). A trained clinician analyses the patient's video, and further analysis about the lesion or normal findings are made.

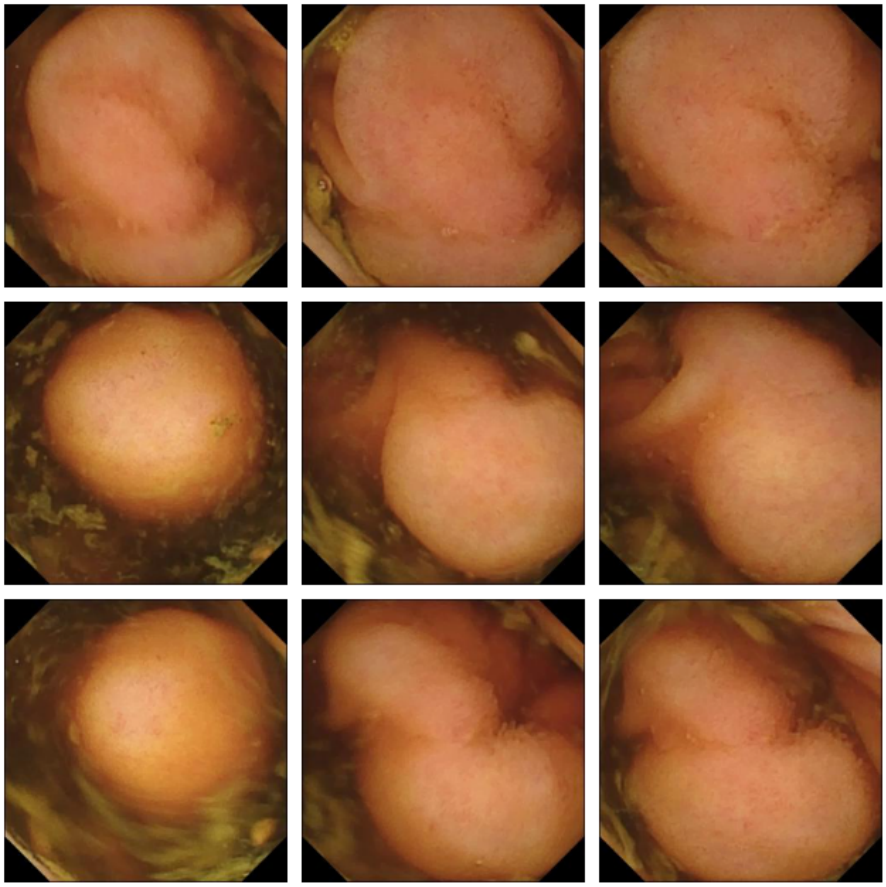


Figure 2.8: Sample polyp images recorded using an Olypmpus RE-10 endocapsule recorder

2.1.2 GI findings

The GI findings include anatomical landmarks, quality of muscoal views, pathological findings, and therapeutic interventions. Figure 2.9 shows example images of the several classes of the GI findings from the Kvasir dataset [148]. In the below section, we briefly describe the classes of GI tract, such as anatomical landmarks, pathological findings, and therapeutic interventions.

- **Anatomical landmarks:** Anatomical landmarks are visible through the endoscope during an endoscopic procedure. Both upper GI tract (esophagus, stomach, and duodenum) and lower tract (terminal ileum, colon, and rectum) have anatomical landmarks. The anatomical landmark is an important characteristic for determining the orientation in an endoscopic procedure [26]. Examples of critical anatomical landmarks are z-line, pylorus, and cecum.

- **Pathological findings:** In the whole GI tract, there is a possibility of occurrences of the abnormalities due to disease. Pathological findings are changes in the in-

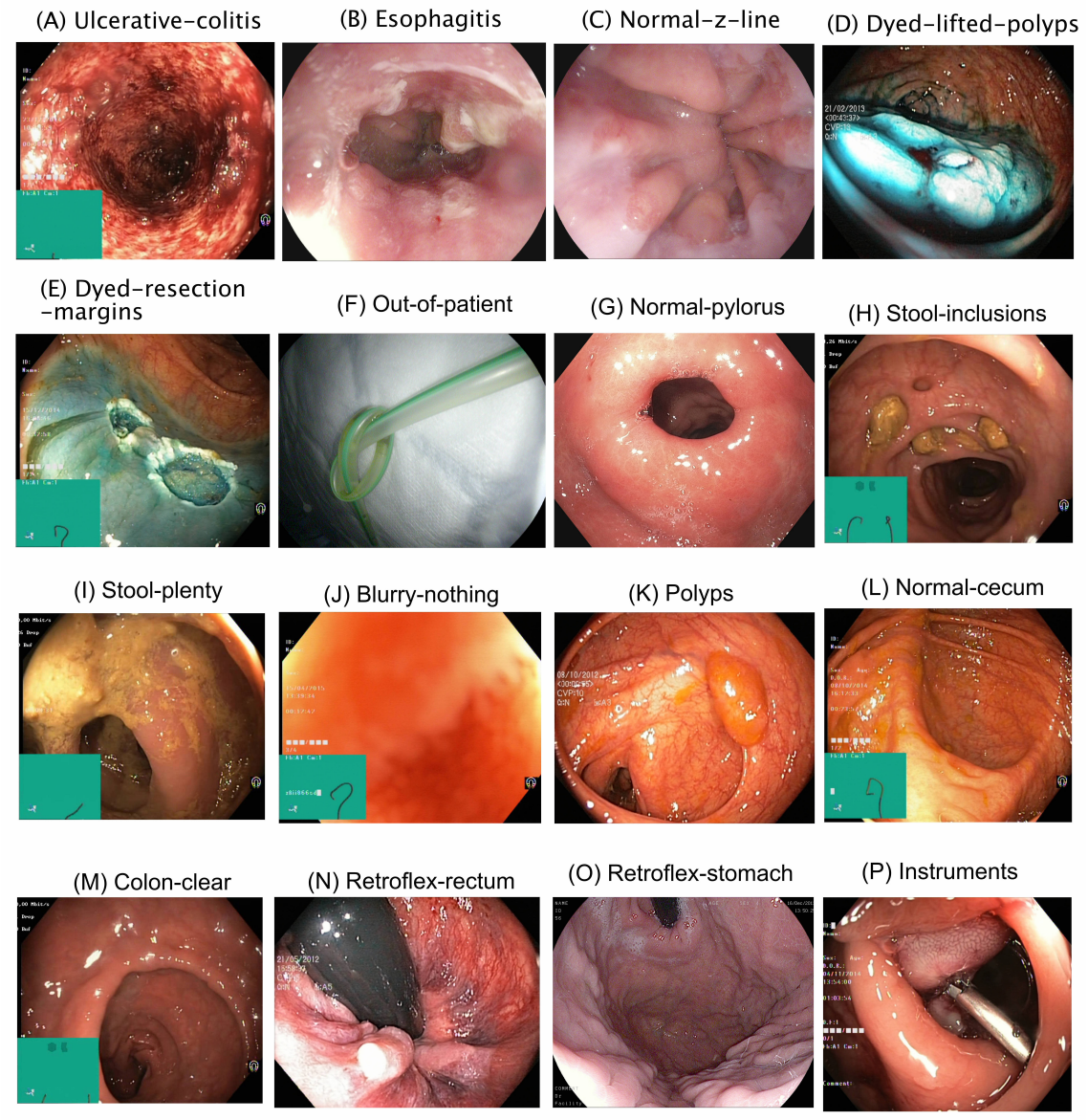


Figure 2.9: Example images from the GI tract [85]

testinal wall mucosa [148]. Examples of pathological findings include esophagitis, polyps, and ulcerative colitis. Early detection of pathological findings could prevent the development of severe GI cancers.

- **Therapeutic interventions:** Therapeutic interventions are interventions required for treatment of the GI tract conditions. A common example is the therapeutic removal of polyps and other suspicious lesions. More information about the therapeutic interventions can be found in [26].

2.1.3 Definition of classification, detection and segmentation

It is essential to build computer vision models that can classify, detect, and segment areas of interest from GI endoscopy frames or videos. In computer vision, to solve the three crucial tasks, the term such as classification, detection and segmentation is used. However, in the community, sometimes the terms are used differently. Therefore, we define these terms and follow the same term in our dissertation.

1. In our work, we use the term *classification* (for example, GI images classification) with the intention to classifying the images to the class it belongs to using ML or DL methods.

2. We use the term *detection* for locating an object or instances of the object of interest in an image. The application of an object detection algorithm could be automated polyp detection in a colonoscopy frame for medical diagnosis.

3. We use the term *segmentation* for identifying each pixel of an image and dividing it into meaningful classes using DL algorithms. One of the applications of segmentation task is automatic polyp segmentation from the colonoscopy images for medical diagnosis.

2.2 GI finding classification

Table 2.1 lists related work in the field of GI image classification. In the table, REC refers to recall, SPEC refers to specificity, ACC refers to accuracy, MCC refers to matthews correlation coefficient, F1 refers to F1 Score, Rk refers to Rk correlation coefficient and FPS refers to Frame Per Second. From the above-related work, we observe that there has been a considerable amount of work done towards the development of an automated model for the classification of the GI tract conditions such as anatomical landmarks, pathological findings, and polyp removal cases, and normal or regular findings from the GI tract [72, 70, 64, 124, 73]. Additionally, there has also been work targeted using a specific dataset [207, 205]. However, for most of the works, the study has been conducted on private datasets [207, 205] which makes it difficult to compare with or reproduce the work in a research context. Additionally, the new SOTA algorithm can not be tested on the same dataset for potential improvements.

An automated Computer aided diagnosis (CAD) system in the clinic has the potential to improve clinical workflow by assisting gastroenterologists during endoscopy or

Table 2.1: An overview of the existing related work on GI tract classification [84]

Reference	Year	REC	PREC	SPEC	ACC	MCC	F1	FPS
Hwang et al. [77]	2007	0.9600	0.8300	-	-	-	-	15
Li et al. [112]	2012	0.8860	-	0.9620	0.9240	-	-	-
Zhou et al. [225]	2014	0.7500	-	0.9592	0.9077	-	-	-
Wang et al. [208]	2014	0.8140	-	-	-	-	-	0.14
Mamonov et al. [127]	2014	0.4700	-	0.9000	-	-	-	-
Wang et al. [209]	2015	0.9770	-	-	0.9570	-	-	10
Riegler et al. [166]	2016	0.9850	0.9388	0.7250	0.8770	-	-	$\sim$ 300
Shin et al. [172]	2017	0.9082	0.9271	0.9176	0.9126	-	-	-
Riegler et al. [165]	2017	0.9850	0.9390	0.7250	0.8770	-	-	$\sim$ 75
Yu et al. [215]	2017	0.5005	0.4917	-	0.9471	-	0.4830	-
Petscharnig et al. [146]	2017	0.7550	0.7550	0.9650	0.9390	0.7200	0.7550	-
Yuan et al. [217]	2018	0.8180	0.7232	-	-	-	0.7431	-
Wang et al. [205]	2018	0.9438	-	0.9592		-	-	-
Mori et al. [134]	2018	>0.9000	-	>0.9000	-	-	-	-
Hoang et al. [72]	2018	0.9281	0.9426	0.9963	0.9932	0.9312	0.9342	23
Hicks et al. [70]	2018	0.9218	0.9378	0.9959	0.9924	0.9228	0.9236	624
Meng et al. [129]	2019	0.8664	0.8664	0.9911	0.9833	0.8542	0.8664	-
Luo et al. [124]	2019	0.9533	0.9533	0.9969	0.9941	0.9480	0.9533	-
Hoang [73]	2019	0.9464	0.9464	0.9964	0.9933	0.9406	0.9464	-
Chang et al. [33]	2019	0.9569	0.9569	0.9971	0.9946	0.9520	0.9569	-
Harzig et al. [64]	2019	0.9490	0.9490	0.9966	0.9936	0.9490	0.9105	-
He et al. [67]	2020	0.9130	0.9130	-	-	0.9030	0.9130	-
Galdran et al. [49]	2020	0.8740	0.8740	-	-	0.8600	0.8740	-

colonoscopy. The ML or DL models could be helpful in findings to unrecognized lesions that were previously missed during the examination. In our research [193, 190], we have done a comprehensive study on the classification of different GI findings using ML and DL techniques. Most of our experiments are done using a public dataset or the dataset that was available to us through competitions (Medico Challenge (Kvasir [148], with some additional images), CVC-ClinicDB [21], CVC-ColonDB [18], and CVC-VideoClinicDB [11, 20])). We have explored and suggested two solutions based on GFs and ML approaches and three DL based models. In addition to the development of improved ML and DL models, the research mostly explored the generalizability of such models and interpretation of the performance metrics used in the GI images classification tasks.

2.3 GI lesion detection

In the entire GI tract, our research is mainly focused on developing DL models for ***colorectal polyp detection***. Automatic polyp detection is an active research area. Table 2.2 shows an overview of some of the related works in the field of automatic polyp detection over time. From the table, we can observe that there has been substantial research in this area. In addition to it, there are also competitions on automated polyp detection [4, 19].

Table 2.2: Overview of the related work on automated polyp detection
.

Reference	Year	Sens	Spec	Prec	mIoU	FPS
Karkanis et al. [99]	2003	0.9000	0.9700	-	-	-
Hwang et al. [77]	2007	0.9600	-	-	-	-
Park et al. [145]	2012	0.5600	-	-	-	-
Tajbaksh et al. [187]	2015	0.8800	-	-	-	-
Mo et al. [132]	2018	0.8733	-	0.8970	-	-
Zhang et al. [220]	2018	0.7160	-	0.8860	-	6.5
Mohammed et al. [133]	2018	0.8440	-	0.8740	-	-
Quadir et al. [173]	2018	0.8030	0.8650	-	0.8330	-
Liu et al. [119]	2019	0.8030	-	0.7360	0.7680	32
Zhang et al. [221]	2019	0.7637	-	0.9392	0.8424	50
Lee et al. [110]	2020	0.9670	-	-	-	67.17

The results from the table and the competitions show that the recent methods achieve decent performance and speed. However, polyp miss-rate are reported up to 27% [1, 126]. The statistics include miss-rate due to the polyp and the operator characteristics. Studies have shown that the polyp detection rate increases when there is assistance from a second observer [13, 30, 108]. Although there are several CADx systems for polyp detection, [131, 135, 187], a system that can be deployed into the clinical setting for detecting and locating polyps accurately in real-time during live examination is still missing. Therefore, we carry out research in the area of development of a real-time polyp detection model that can help reduce the miss-detection of polyps.

2.4 GI lesion segmentation

In the GI tract, CRC is the second most lethal cancer [184]. Therefore, our research is mostly focused on building ***semantic segmentation architectures for colorectal polyp segmentation***. Table 2.3 shows related work in the field of automated polyp segmentation. From the above-related work, we can see that there has been a considerable amount of work done in the field of automated polyp segmentation for more than a decade, and significant progress has been made. Researchers have conducted both retrospective and prospective studies. Researchers have also considered using more than one dataset to demonstrate the effectiveness of their architectures [205]. The previous research mostly focused on obtaining high performance metrics such as DSC, mean intersection over union (mIoU), recall, sensitivity, or precision. Development of the real-time DL models has been mostly neglected. Also, most of the research was done on the non-public dataset, and the

Table 2.3: Overview of the related work on automated polyp segmentation

Reference	**Year**	**DSC**	**mIoU**	**Rec**	**Prec**	**FPS**
Wijk et al. [199]	2010	-	-	0.9500	-	-
Breier et al. [28]	2011	-	-	0.4814		-
Bernel et al. [18]	2012	0.5533	-	-	-	-
Ganz et al. [51]	2012	-	-	0.7100	-	-
Tajbakhsh et al. [188]	2014	-	-	0.8000	-	-
Bernel et al. [21]	2015	-	0.7274	-	-	-
Li et al. [114]	2017	-	-	0.7732	0.8999	-
Vazquez et al. [201]	2017	-	0.5607	-	-	-
Zhang et al. [218]	2017	0.7000	-	0.7566	-	-
Wang et al. [205]	2018	-	-	0.8824	-	-
Akbari et al. [2]	2018	0.8100	-	0.7480	-	-
Brandao et al. [27]	2018	-	0.5695	0.6802	-	-
Nguyen et al. [140]	2018	0.8890	0.8935	-	-	-
Wichakam et al. [213]	2018	0.9594	0.6936	0.7814	0.8848	-
Kang et al. [98]	2019	-	0.6607	0.7437	0.7384	-
Sun et al. [182]	2019	0.8248	0.9611	0.9551	0.9671	-
Fang et al. [46]	2019	0.8308	0.8633	-	-	-
Poorneshwaran et al. [153]	2019	0.8848	0.8127	-	-	-
Thomaz et al. [9]	2019	-	0.9140	0.8980	0.9360	-
Kassani et al. [100]	2019	0.9087	0.8382	-	-	-
Guo et al. [60]	2019	0.8014	-	0.8210	0.8349	-
Qadir et al. [154]	2019	0.7042	0.6124	0.7259	0.8000	-
Dijkstra et al. [41]	2019	0.6906	0.5820	-	-	-
Zhou et al. [226]	2019	-	0.3345	-	-	-
Banik et al. [14]	2020	0.8130	-	0.7860	0.8090	-
Fan et al. [45]	2020	0.8980	0.8400	-	-	-
Nguyen et al. [139]	2020	0.9130	-	0.9530	0.9580	-
Thambawita et al. [191]	2021	0.9720	0.9490	0.9720	0.9720	-
Huang et al. [76]	2021	0.9040	0.8480	0.9230	0.9070	86.7
Zhang et al. [222]	2021	0.9200	0.8700	-	-	-
Ghimire et al. [55]	2021	0.9124	0.8649	-	-	-

Table 2.4: Overview of the related work on surgical instrument segmentation

Reference	Year	DSC	mIoU	Sens	Spec	Acc	FPS
Shvets et al. [174]	2018	0.8887	0.8236	-	-	-	88.87ms
Pakhomov et al. [144]	2019	-	-	0.8570	0.9880	0.9230	-
Yu et al. [216]	2019	0.9220	0.8645	-	-	0.9156	-
Lee et al. [109]	2019	0.9519	0.9085	-	-	-	-
Ni et al. [141]	2020	0.6410	-	-	-	-	-
Isensee et al. [82]	2020	0.8741	-	-	-	-	-

implementation codes were not released publicly. Additionally, a generalizability study has been completely missing in the field. In our research, we have addressed the above issues.

2.5 Surgical instrument segmentation

Several medical examinations and procedures involve medical instruments and tools. In the colonoscopy and gastroscopy examinations from GI tract, we can often see the scopes, snares, scissors, etc. Another example is diagnostic laparoscopy in general, a minimally invasive surgical procedure used to examine the organs inside the abdomen that requires only small incisions. In all such automated analyses of the images or video frames, it is essential to identify and localize the instruments, which gives another requirement for the ML models to be used in CAD systems in the clinic.

Table 2.4 shows the relatively limited related work in the field of surgical instrument segmentation. In addition to the related works, there are also challenges and competitions organized [25, 8, 168]. Despite significant progress achieved, the goal of developing a successful model that can segment and different instruments from the different categories to meet the requirement of the clinical needs has not been achieved. The development of such models or systems could assist in surgery. Such a system could be efficient and reduce the overload of the surgeons and the hospitals.

2.6 Current challenges

There are several challenges associated with developing a CADx system for GI tract disease detection. A key challenge for the development of a CADx system is the availability of datasets for training the ML algorithms. There is a high cost associated with the collection of the medical dataset. It requires a team of experienced medical experts

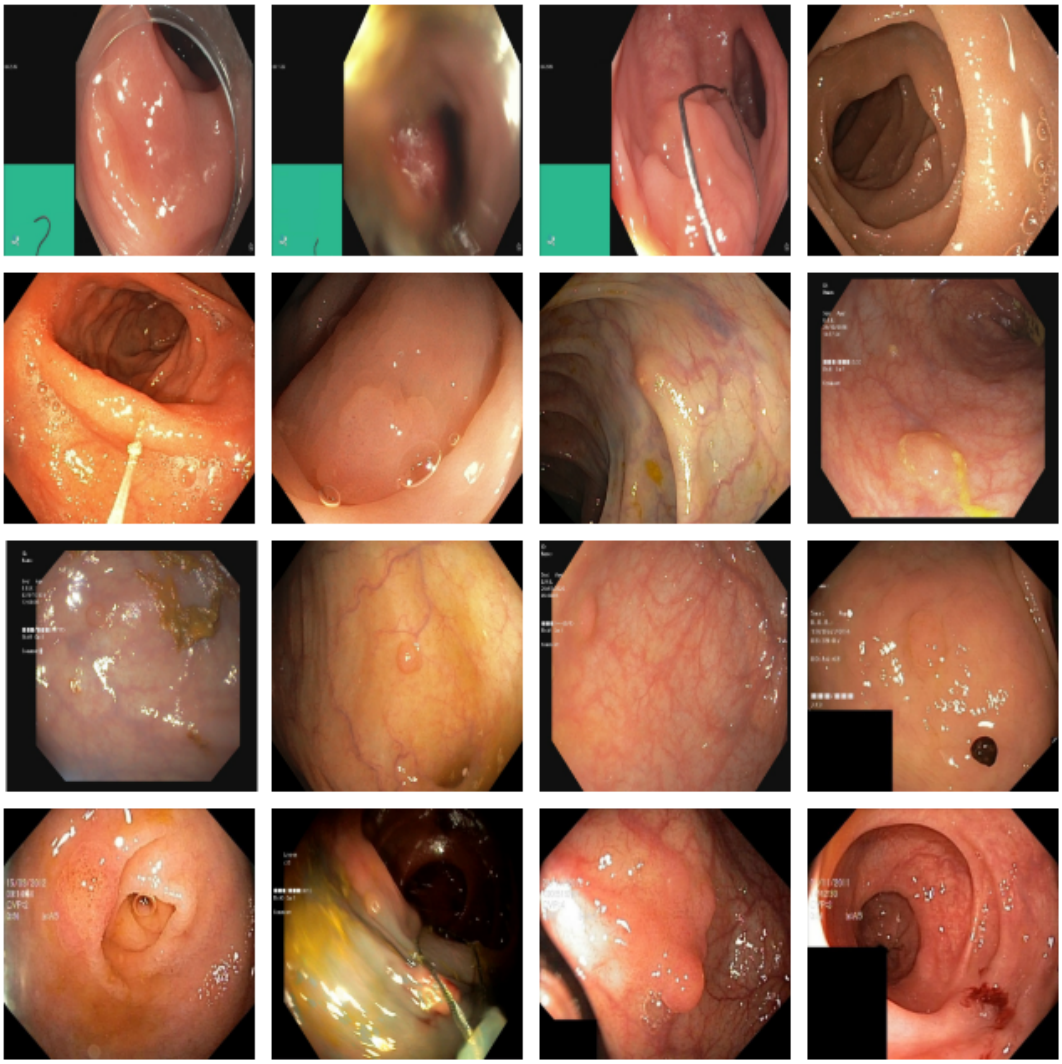


Figure 2.10: Example of challenging polyps with different shapes, sizes, and appearances from our Kvasir-SEG and PolypGen datasets

(gastroenterologists in our case). Additionally, it requires consent from the hospital and the patients. An agreement must be signed depending on country regulations. The privacy concerns associated with the dataset should be addressed. The most important and tedious task is dataset labeling, annotation, and cleaning for training and testing the ML algorithms.

GI tract diseases, for example, in colonoscopy polyps, the shape, size, color, and appearance varies largely. The example of colonoscopy polyps can be seen in Figure 2.10. There is also a change in the structure and characteristics over time [85]. During a colonoscopy, the color of the same polyp may vary because of lightning conditions and scope movement. Additionally, there are polyp-like structures in the colon that may look similar to actual polyps. There might be certain artifacts such as blurriness in the images, surgical instruments, flares, intestinal content, etc. Moreover, there is high inter-class similarity inside the GI tract and intra-class variations. Despite the progress in colonoscopy, the adenoma miss-rate is 6-27% [1]. In the pooled analysis of 8 randomized tandem colonoscopy studies, the frequently missed polyps were sessile polyps, flat polyps, and the polyps whose size were $\leq$ 10 mm [227]. One of the reasons for polyp miss-rate during colonoscopy is that it was either not in the visual field of the colonoscope or hard to reach by gastroenterologists. The other reasons may be that polyps in the colon were not identified but were overlooked. This is either because polyps were too small and it has a resemblance with the surrounding tissue. Additionally, the gastroenterologists might miss polyps that were under the visual field but might have missed due to inadequate diligence [171, 186]. Despite several efforts from the community [155, 147, 22, 163, 10], more research is required to address this challenges.

2.7 Summary

In this chapter, we have detailed some statistics about the GI tract and also highlighted the examination procedure. We have presented the terminology in the domain such as used classification, segmentation, and detection that will be used throughout the work. We have also listed some relevant related work in GI tract classification, polyp segmentation, polyp detection, and instrument segmentation. We have listed the current challenges in the field of GI endoscopy.

In general, endoscopy activity has been increasing all over the world due to improved healthcare and awareness. However, the competing demands for endoscopy have yet to be met. This is mainly due to lack of medical expertise, hospitals, and high cost.

New initiatives (for example, use of FIT, FOBT, wireless VCE) or different types of endoscopy (therapeutic endoscopic ultrasound, endoscopic submucosal dissection) and guidelines have helped to improve the quality of the colonoscopy and endoscopy. However, new technology raises new challenges to the physicians (for example, the need for dedicated endoscopy training) rather than the prior endoscopy practice. Although there has been various several screening options and supporting evidence for reducing one of the leading cause of cancer in the GI tract, CRC, there is still need of additional research [116, 117].

The CADx system has the potential to accelerate the diagnosis process and can potentially reduce the miss-detection. Moreover, it could also save clinicians time. CADx model should have the capability to provide at least real-time feedback and have the potential to be integrated into the clinical workflow. Additionally, the development of these systems could incur minimum additional costs. Among different types of challenges, one of the main challenges in the field of GI endoscopy is the lack of dataset availability. For the dataset collection and annotation, there is the requirement of a substantial amount of time and effort from medical experts. Although different research groups and hospitals collecting datasets and working toward the development of the CADx system, public sharing of datasets remains a challenging issue. Although there are works that have shown promising results [31, 36, 110, 130, 152, 205], the algorithms are mostly tested on non-public datasets, making it difficult for reproducibility and comparison. In the next chapter, we will present details about datasets design that were publicly released as a part of our work to accelerate research in the field of GI endoscopy.

Chapter 3

Dataset design and curation

Dataset is a prerequisite for the DL models. Despite the potential of the CADx system to automatically detect the disease, the GI tract field remains relatively unexplored. The lack of a publicly available dataset for research and development is one of the contributing factors to the missing development of automated systems for GI disease. One of the reasons behind the lack of dataset in the field is because it is challenging to obtain. In this chapter, we present our datasets that were released to tackle the lack of dataset in the field of GI endoscopy.

3.1 Dataset collection

Table 3.1 shows an overview of the GI datasets. From the table, we can observe that there is a lack of datasets in the field. Even with the available datasets, some of them are not usually used in the research. They are atlas (for example, atlas of gastrointestinal endoscope [212]) or mostly used for education purposes. Most of the datasets have very few samples. For example, ETIS-Larib [177] has 196, CVC-ColonDB [18] has only 380 images. In addition, most of the dataset is only available through request or participation through challenge (for example, GIANA 2017 [17], GIANA 2018 [11, 20]) and proper licensing are required to use the datasets (for example, ASU-Mayo polyp database [187]). In this respect, we have released few publicly available datasets with large samples, diversity, and classes that do not require any licensing or restriction for academic and research use.

Data collection, curation, annotation, and public release remained vital parts of our research. We achieve this goal by collecting several GI endoscopy datasets, cleaning, annotating, extracting the ground truth, and releasing them through our webpage and various dataset platforms for the research community. We encourage open and repro-

Table 3.1: An overview of existing GI datasets [26]

Dataset	Findings	Size	Availability
ETIS-Larib Polyp DB [177]	Polyps	196 images†	Public
CVC-ColonDB [18]	Polyps	380 images† $^{\psi}$	By request•
CVC-ClinicDB [21]	Polyps	612 images†	Public
Endoscopy Disease Detection challenge (EDD 2020) [6]	Polyp, barrett's esophagus, high-grade dysplasia, suspicious (low-grade), cancer	386 images	Public
GIANA 2017 [17]◇	Polyps & angiodysplasia	3462 images and 38 videos	By request
GIANA 2018 [11, 20]◇	Polyps & small bowel lesions	8262 images and 38 videos	By request
ASU-Mayo polyp database [187]	Polyps	18,781 images†	By request•
CVC-VideoClinicDB [11, 20]	Polyps	11954 images†	By request•
KID [106]◇	Angiectasia, bleeding, inflammations, polyps	2371 images and 47 videos	public•
GASTROLAB [54]	GI lesions	Some 100s of images and few videos	Public♣
WEO Clinical Endoscopy Atlas [212]	GI lesions	152 images	By request♣
GI Lesions in Regular Colonoscopy Data Set [53]	GI lesions	76 images†	By request
Atlas of Gastrointestinal Endoscope [194]	GI lesions	1295 images	Unknown•
El salvador atlas of gastrointestinal video endoscopy [43]	GI lesions	5071 video clips	Public♣
Kvasir [148]	Polyps, esophagitis, ulcerative colitis, z-line, pylorus, cecum, dyed polyp, dyed resection margins, stool	8000 images	Public
Nerthus [150]	Stool - categorization of bowel cleanliness	21 videos	Public
Our datasets			
HyperKvasir [26]	Upper GI and lower GI findings	110,079 GI images and 373 videos	Public
KvasirCapsule [180]	GI anomalies	4,741,504 images	Public
Kvasir-SEG [89]	Polyps	1000 images†	Public
Kvasir-instrument [88]	Therapeutic tools	590 images†	Public
Medico Automatic polyp segmentation challenge [91]	Polyps	160 images†	Public
EndoTect Challenge [71]	Upper (U) and lower (L) GI findings, and Polyps	800 images (U & L GI) & 200 images†	Public
KvasirCapsule-SEG [92]	Polyps	55 images†	Public
PolypGen [5]	Polyps	3446 images†	Public

†Including ground truth segmentation masks ◇Video capsule endoscopy •Not available anymore $^{\psi}$ Contour ♣Not really a dataset usable for ML

ducible science, so we make our dataset accessible for non-commercial, research, and academic purposes. We also encourage the community to use our research for commercial purposes after the agreement. Additionally, we also publish research articles on the dataset and establish standard benchmarks so that other researchers in the community

will be encouraged to use and improve the results. Making the dataset public and the official splits enable comparing the methods on the same dataset and related analysis. The public release of the dataset helps in the reproducibility of the ML methods that remains one of the significant challenges in the field.

Toward solving the availability of the dataset in the public domain, our team has already released the Kvasir [148] dataset for multi-class classification of the GI disease and the Nerthus [150] dataset for evaluating the quality of bowel cleansing. These datasets were useful for the classification of the GI disease. However, the Kvasir dataset is only a multi-class classificationGI tract dataset. Similarly, the Nerthus dataset contains image frames from the bowel-preparation quality. Both of them are very useful datasets, but these datasets could not be useful for the segmentation and detection task. So, we started collecting a large number of the samples for the Kvasir dataset and released HyperKvasir [26]. At the same time, we also focused on VCE datasets, and released Kvasir-Capsule [180]. Moreover, we started collecting images and videos from a single centre and multi-centre, annotating and generating ground truth, bounding boxes for the corresponding images and videos, and releasing it publicly. So, far we have released Kvasir-SEG [89], HyperKvasir [26], Kvasir-Capsule [180], PolypGen [5], Kvasir-instrument [88], KvasirCapsule-SEG [92], Medico automatic polyp segmentation challenge dataset [91] and Endotect dataset [71].

3.2 Dataset annotation protocol

The annotation protocols for different datasets were different. For the Kvasir-SEG [89], there was a team of a PhD student (computer science), a medical doctor and a senior gastroenterologist. All the images were uploaded to the Labelbox[1]. Our aim was to label the region of interest from each image. Each of the images was manually annotated by the computational scientist with the help of a medical doctor. Later on, each image was verified by an expert gastroenterologist. Figure 3.1 shows the images, annotated corresponding ground truth, and the bounding boxes surrounding polyp.

Similarly, for PolypGen [5] collection, construction and annotation, we had a team of two post-doctoral researchers, a PhD student from a computer science background and a team of 6 senior gastroenterologists. Each image was annotated by the computational scientist and reviewed and verified by a team of senior gastroenterologists. More details about the dataset collection and annotation protocol can be found in the paper [5].

[1] `https://labelbox.com/`

Similarly, for Kvasir-instrument [88], we had a team of two senior gastroenterologists, a PhD student, and a summer intern for dataset collection and annotation. Automatic polyp segmentation challenge [91], Endotect challenge [71], and KvasirCapsule-SEG [92] dataset were annotated by a team of a PhD student and an expert gastroenterologists. The details about HyperKvasir[26] and Kvasircapsule [180] curation can be found in their respective paper.

3.3 Collected datasets

In this section, we introduce each of the collected, curated, and annotated datasets.

3.3.1 Kvasir-SEG

Colorectal polyp segmentation is a demanding task in medical image segmentation. Research in the field of polyp segmentation using computer vision techniques has the potential to improve examination procedures and reduce the polyp miss rate during colonoscopy. However, it is challenging to find the publicly available dataset. Even if the images are available, it is challenging to obtain ground truth that shows the pixel-precise region covered by polyps. In this context, we selected the polyp class of the Kvasir dataset and annotated it with the help of a medical doctor and an expert gastroenterologist. The information about the Kvasir dataset collection procedure and dataset detail can be found on this webpage[2]. By adding ground truth and bounding box information to the polyp class of the Kvasir dataset, we enable and encourage the computer vision and multimedia community researchers to develop methods that can contribute to automated polyp segmentation.

Figure 3.1 shows the example images from Kvasir-SEG. The white mask shows the area covered by the polyp region, and the background regions show that it contains non-polyp tissue pixels. Few image samples contain the endoscope position marking probe that shows the position from where the images were captured. The images in the Kvasir-SEG were captured using ScopeGuide (Olympus). It is to be noted that in the Kvasir-SEG, we have replaced 13 images from the polyp class to enhance the dataset quality. We have put images and masks into a separate folder. The information about the bounding box is stored in JSON file format. The image name and its corresponding ground truth are the same.

[2] `https://datasets.simula.no/kvasir/`

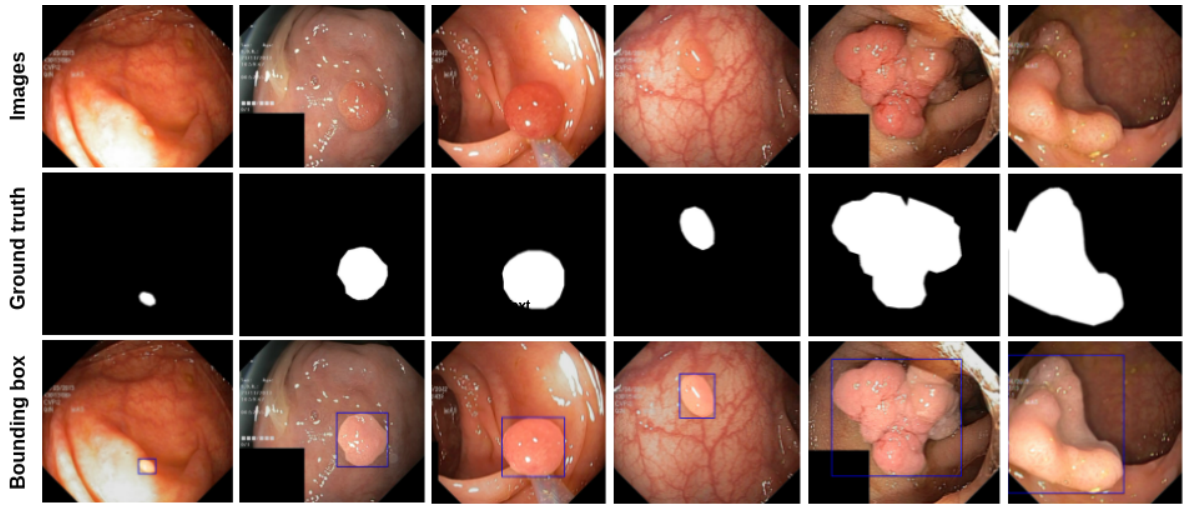


Figure 3.1: Polyps, corresponding ground truth, and bounding boxes from Kvasir-SEG dataset

Kvasir-SEG can be downloaded from here[3]. No prior permission is required for downloading when using the dataset for academic and research purposes. For commercial purposes, prior permission is required. We have received few requests to use the dataset for commercial purposes and have approved them all. The dataset has also been used in a few competitions and challenges. It was recently used at a training dataset at "Medico automatic polyp segmentation challenge" and "EndoTect Challenge" in 2020. In addition to providing the dataset, we also provide baseline results. The baseline results can be found in [89]. Through these baseline results, we also invite other medical image analysis and multimedia research to develop and improve the current SOTA. Portions of the research community has adopted the Kvasir-SEG for benchmarking their new ML algorithms and developing novel methods on the dataset.

3.3.2 Endocv2021 Challenge dataset

In 2021, we have hosted the EndoCV2021 Challenge in collaboration with the University of Oxford. The challenge aimed to develop DL methods to address the generalizability in polyp detection and segmentation task on a multi-centered dataset. We collected the dataset from 6 different clinical centers. The example of sample dataset from each center can be found in Figure 3.2.

These datasets were collected from medical centers from Norway, France, Italy, UK, and Egypt. We have provided both still image frames and video sequence frames as part of the dataset. Along with the image and video sequence, we provided: (1) The ground

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

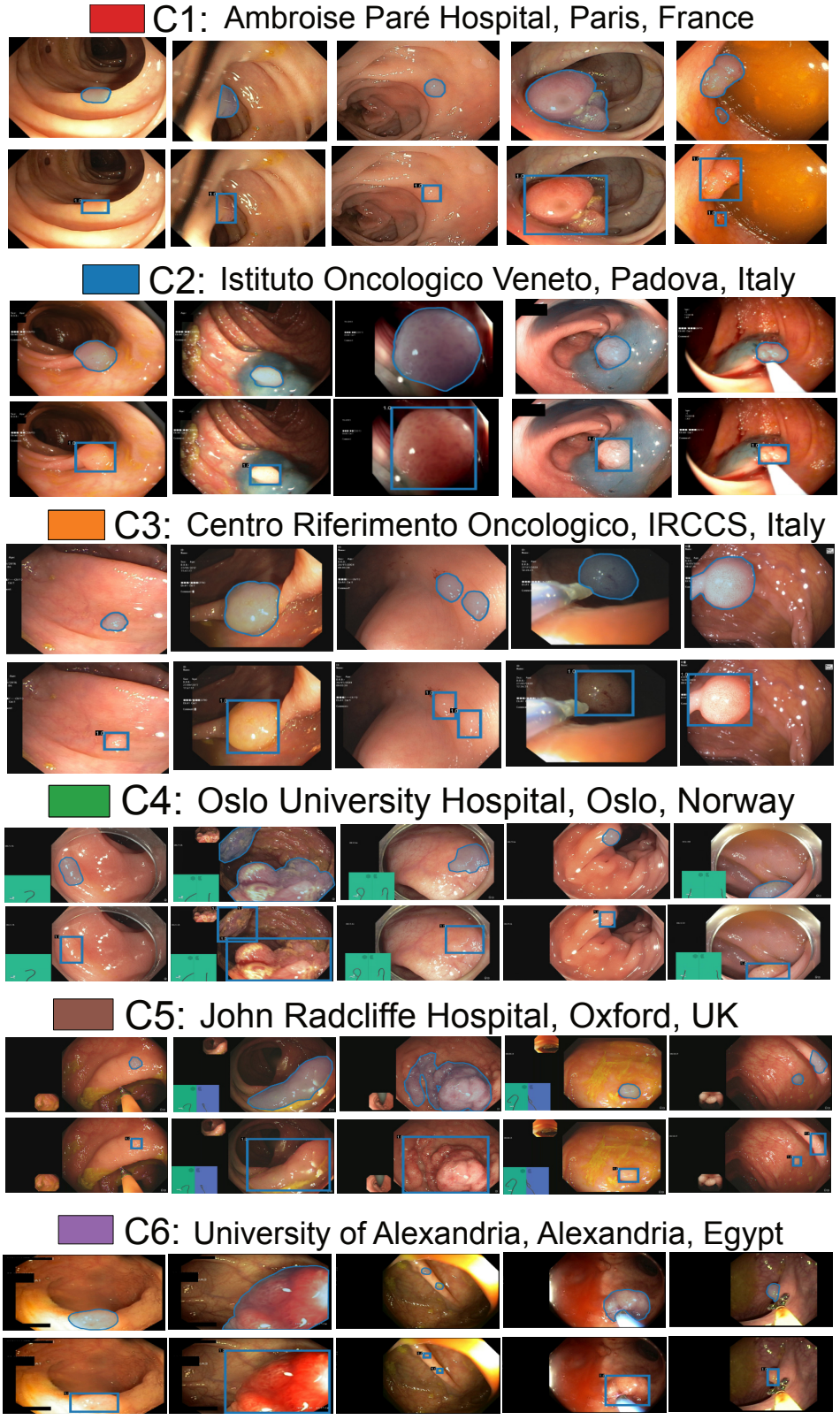


Figure 3.2: Example of polyp annotations from six different centers [5]

truth mask; (2) Image with bounding box showing the area of interest; (3) Bounding box information for each image. When releasing the challenge dataset, we organized the datasets contributed by the different institutions in separate folders. Overall, we have

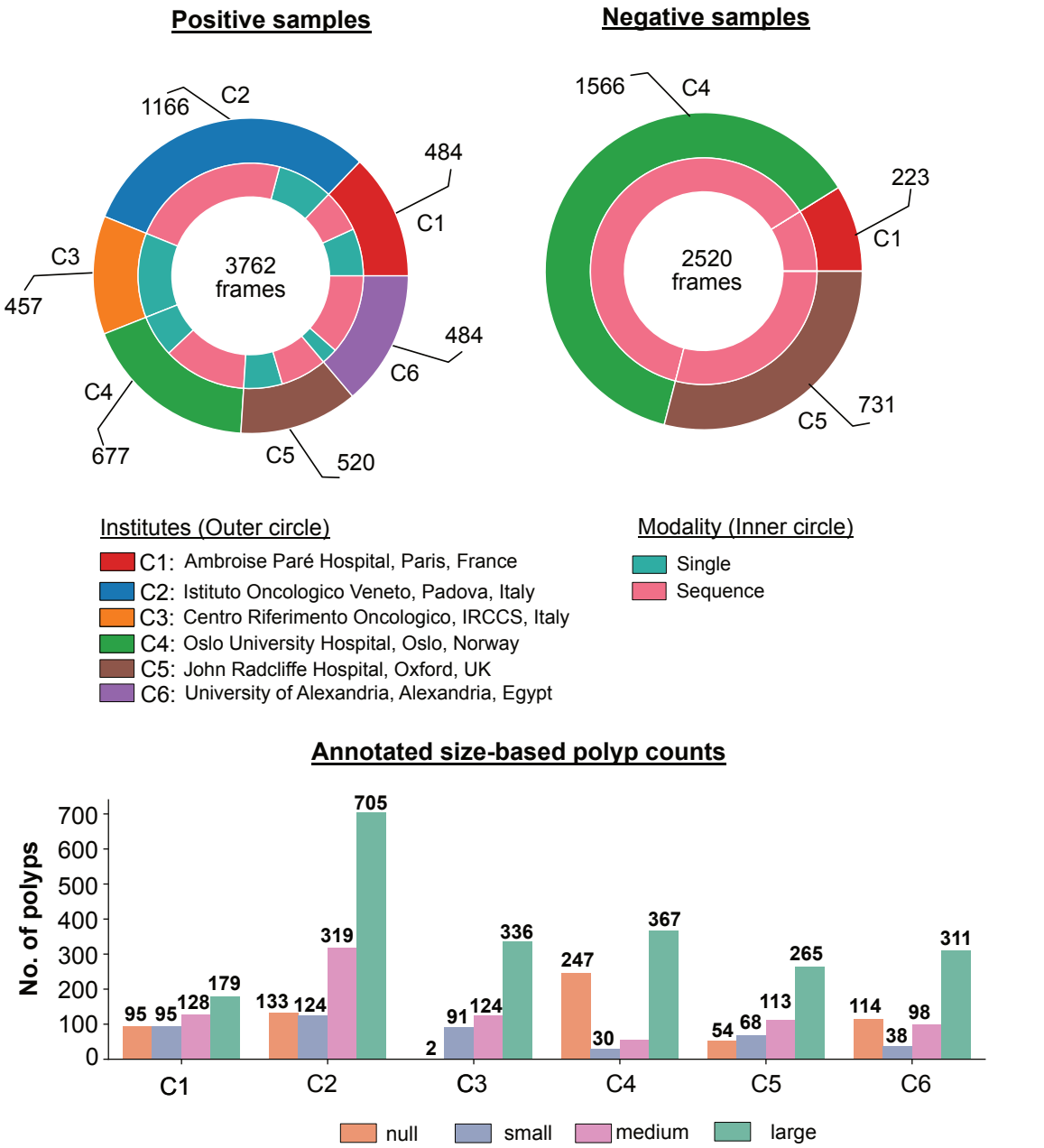


Figure 3.3: An overview of the positive and negative samples from PolypGen dataset [5]

released 1,452 still images. We have also released a sequence dataset, which consists of 655 images (490 positive samples comprised of a polyp and 165 negative samples consisting of normal image frames). Later on, the size of the polyp samples of the same dataset was increased to 3762 frames, and negative samples were to 2520 frames. Figure 3.3 shows an overview of polyp samples from each dataset. The details about the challenge set up, information about the winning solution for both detection and segmentation can be found in `https://endocv2021.grand-challenge.org/`.

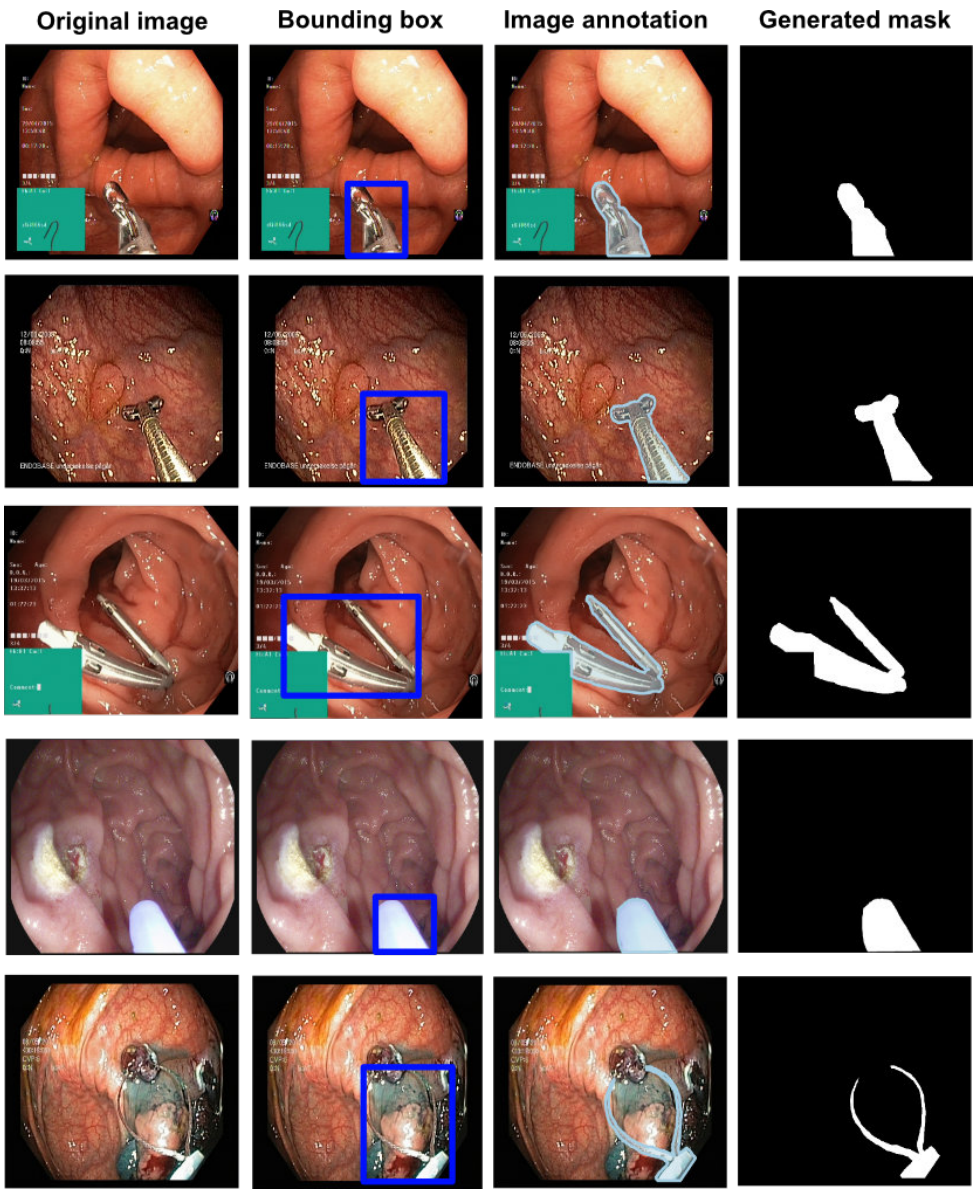


Figure 3.4: Examples from Kvasir-Instrument dataset [88]

3.3.3 Kvasir-Instrument dataset

GI endoscopy is a well-researched topic. Previous studies have considered the development of new ML methods for classification of GI tract findings, including regular findings, anatomical landmarks, pathological findings, and instruments in challenges such as Medico Task 2018 [149] and BioMedia 2019 grand challenge [69]. We have provided a detailed comprehensive analysis on these challenges [85]. An automated method for the segmentation, detection, and localization of endoscopic tools is essential in laparoscopy [168] and robotic-assisted surgery [7, 8]. However, because of the limited availability of publicly available datasets, it is difficult for research and development. This was the main motivation for publishing the *Kvasir-Instrument* dataset.

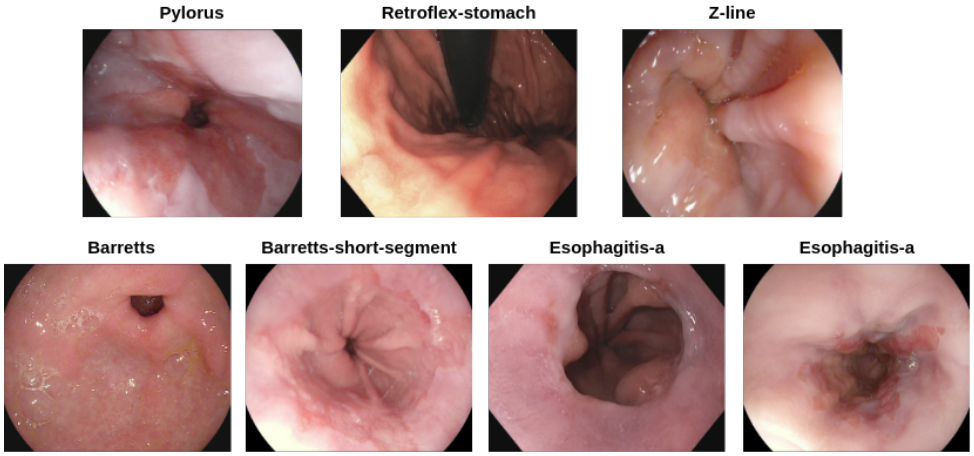


Figure 3.5: Example images from upper GI findings of HyperKvasir dataset

The Kvasir-instrument [88] dataset comprises 590 instrument images from GI endoscopic procedures, their corresponding ground truth, and the bounding box information. It is provided open access and can be downloaded from our webpage[4]. The image resolution varies from 720×576 to $1,280 \times 1,024$. Figure 3.4 shows sample images from Kvasir-instrument. From the figure, we can see different types of instruments used in endoscopic surgery. In the GI endoscopy and laparoscopy, instruments are used surgery. For the input images, we have first annotated the image using Labelbox. After annotation, we can extract the ground truth. We have generated the bounding boxes from the ground truth. In the Figure, we have shown examples of bounding boxes surrounding the instrument, image annotation, and generated masks. Detailed descriptions about the Kvasir-instrument dataset, such as annotation strategy and benchmark results, can be found in our paper [88].

3.3.4 HyperKvasir

Hyper-Kvasir [26] is the world's largest publicly available dataset in the field of GI. It consists of 110,079 images and 374 videos. Out of 110,079 images, the total number of labelled images from 23 classes is 10,662. The total number of unlabelled images in HyperKvasir is 99,417. It is likely that the research community might further classify the unlabelled images into different classes. GI tract is classified into upper GI and lower GI. Upper GI consists of anatomical landmarks such as pylorus, retroflex-stomach, z-line, and pathological findings such as barretts, barretts-short-segment, and esophagitis. Figure 3.5 shows example images from upper GI findings from HyperKvasir dataset.

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

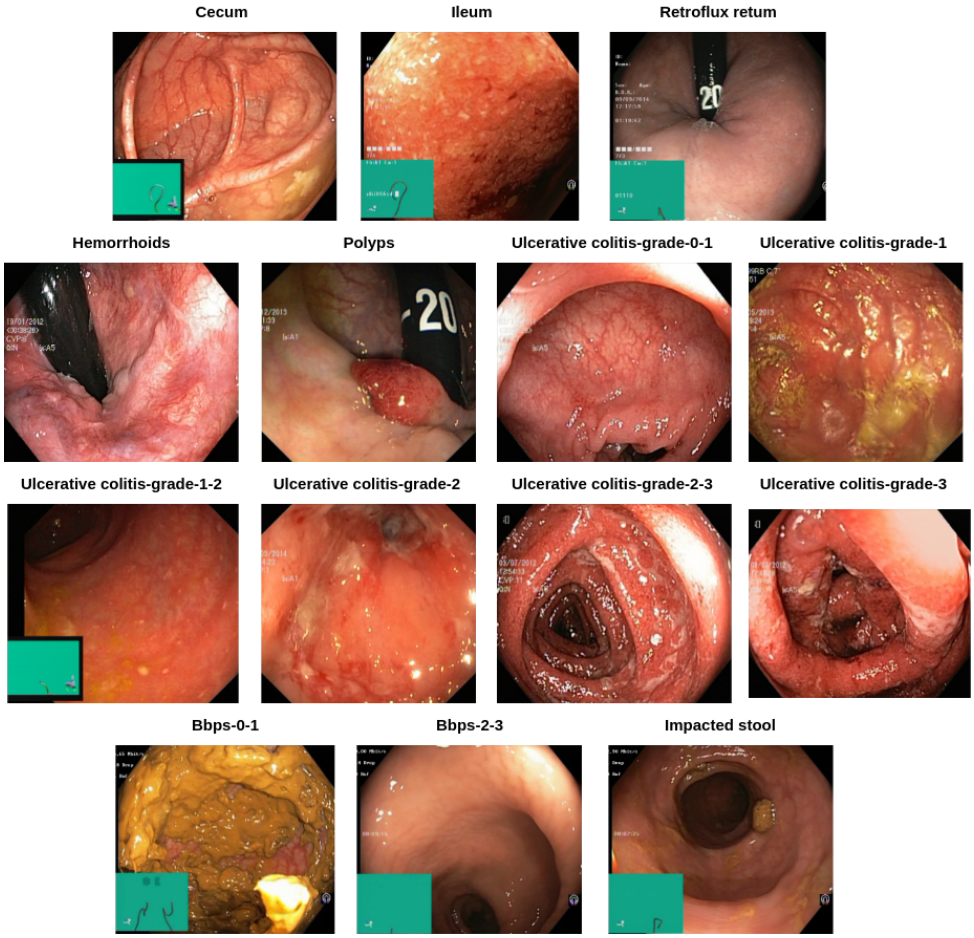


Figure 3.6: Example images from lower GI findings of HyperKvasir dataset

Similarly, the lower GI consists of anatomical landmarks (cecum, ileum, and retroflex-rectum), pathological findings (hemorrhoids, polyps, ulcerative colitis), quality of mucosal view (impacted stool, Boston bowel preparation scale), and therapeutic interventions (dyed-lifted polyps and dyed resection margins). Figure 3.6 shows sample images from lower GI tract. This dataset was used for organizing the Endotect Challenge [71] recently. The dataset can be downloaded from our web pages[5]. A more detailed explanation about the HyperKvasir dataset and baseline experiments can be found in our paper [26].

3.3.5 Kvasir-Capsule

Kvasir-Capsule [180] is the world's largest publicly available video-capsule endoscopy dataset. The images were collected from Norwegian hospitals and contains 118 videos. The total number of extracted image frames is 4,820,739, out of which 44,228 frames are medically verified. These datasets are categorized into 13 different classes. Still, there are

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

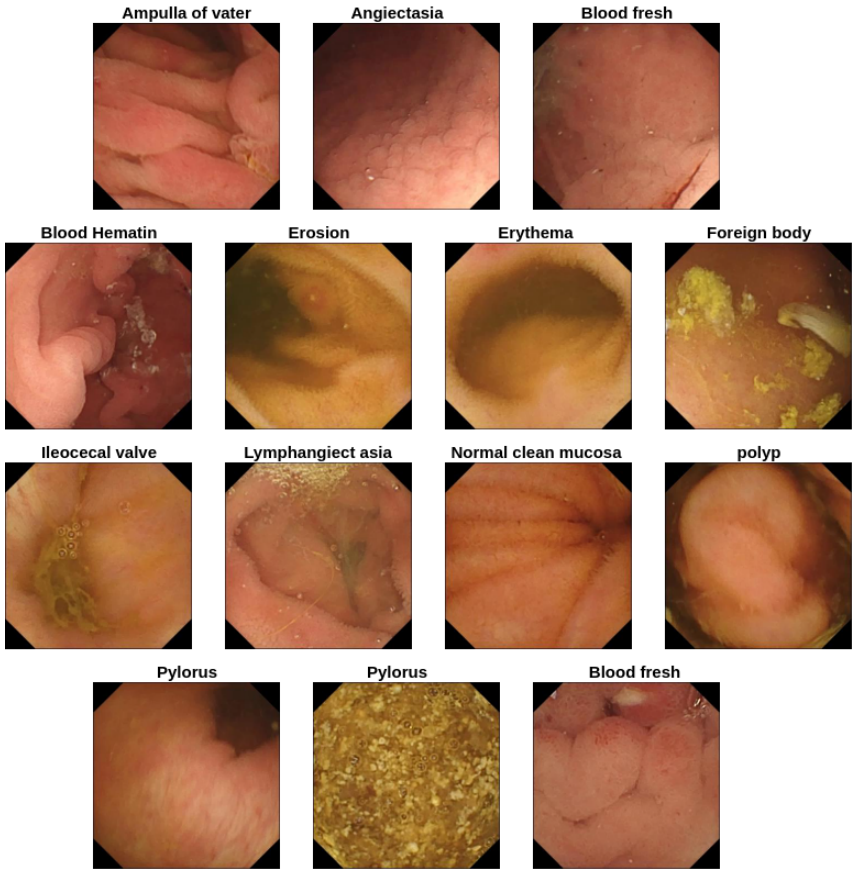


Figure 3.7: Example images from the labelled image classes of Kvasir-Capsule dataset

around 4,776,479 unlabelled frames. It could be of interest to the multimedia and medical community to further classify and annotate these datasets. The dataset sample can be seen in Figure 3.7. Kvasir-Capsule can be downloaded from[6]. More detailed information about the dataset, baseline experiments and results can be found in our paper [180].

3.3.6 KvasirCapsule-SEG

KvasirCapsule-SEG [92] is a VCE polyp segmented dataset. It contains polyp images from the KvasirCapsule [180]. The polyp class of KvasirCapsule has only 55 images. Figure 3.8 shows example images from KvasirCapsule-SEG dataset. We have annotated and generated both ground truth and bounding box information with separate folders for images, ground truth, and images with bounding boxes. The dataset can be downloaded from `https://www.kaggle.com/debeshjha1/kvasircapsuleseg`. More details about the dataset and baseline experiment can be found in [92].

[6]https://osf.io/dv2ag/

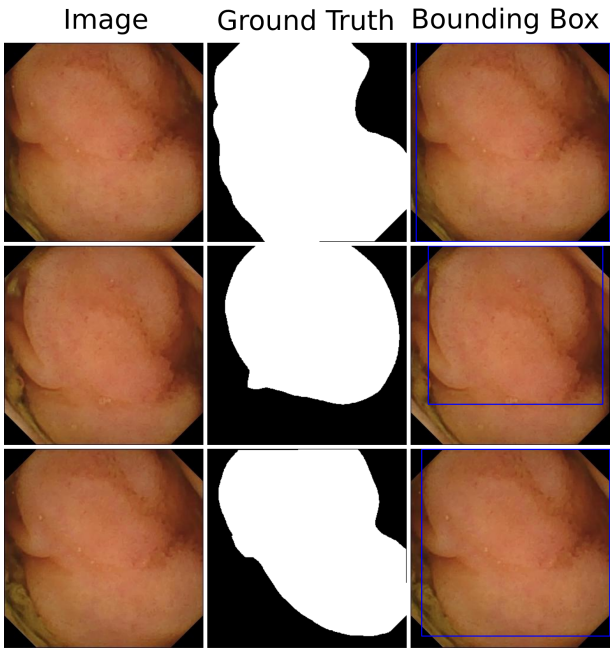


Figure 3.8: Polyps, their corresponding ground truth, and bounding box information from the KvasirCapsule-SEG dataset

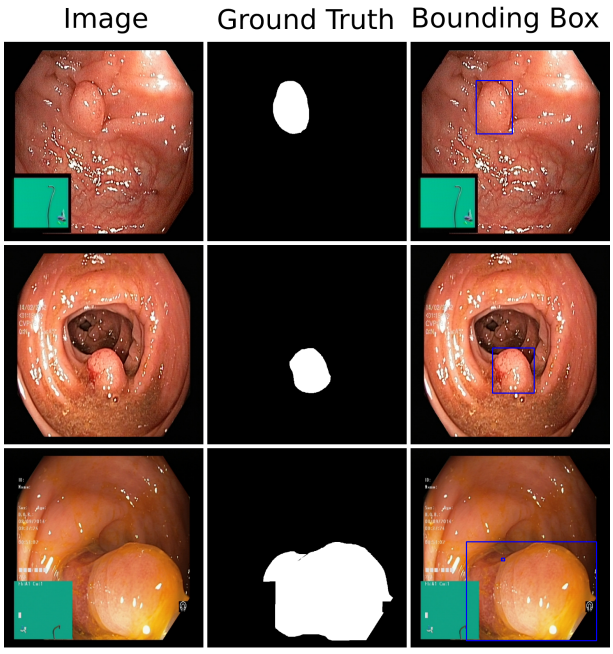


Figure 3.9: Polyps, their corresponding ground truth, and bounding box information from *Medico automatic polyp segmentation challenge* dataset

3.3.7 Medico automatic polyp segmentation Challenge dataset

The Medico automatic polyp segmentation dataset was released as part of the 2020 Medico challenge. The participants were asked to use Kvasir-SEG [89] as the training dataset.

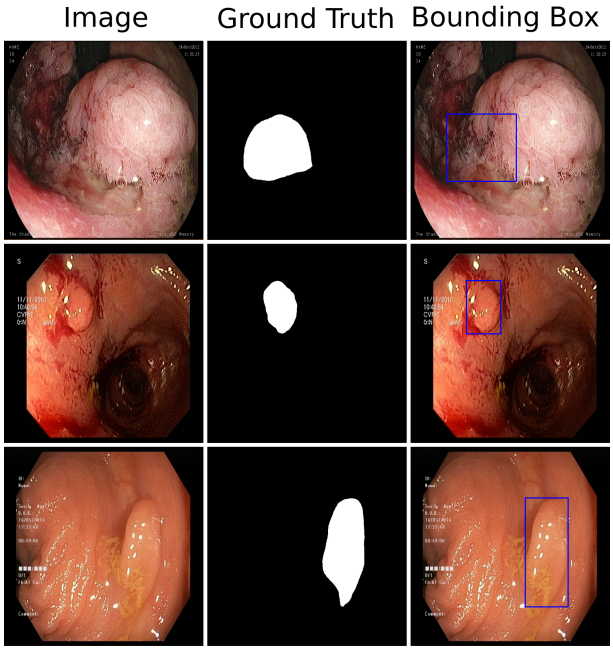


Figure 3.10: Polyps, their corresponding ground truth, and bounding box information from *endotect challenge* dataset

We released separately 160 polyp images as the test dataset. The example images from "Medico automatic polyp segmentation challenge dataset" can be found in Figure 3.9. The participants were asked to test their algorithm and send us the ground truth, where we compared the predicted ground truth with our ground truth and provided the scores to the participants. The detailed information about the challenge can be found in `https://multimediaeval.github.io/editions/2020/tasks/medico/`. The test dataset can be found in the [7]. We have also introduced the challenge and shown the results of each task in Chapter 4.

3.3.8 Endotect Challenge dataset

The Endotect Challenge dataset was released as part of Endotect challenge[8]. The challenge is divided into two parts: (1) GI tract findings classification and (2) polyp segmentation. The development dataset contains 110,079 images and 373 videos. It contains labeled images, unlabeled images, segmented images, and videos. HyperKvasir [26] was used for training the algorithms. Kvasir-SEG [89] was used for training the polyp segmentation task.

As the test dataset, we have released 721 images from 11 different classes for the

[7] `https://www.kaggle.com/debeshjha1/medico-automatic-polyp-segmentation-challenge`
[8] `https://endotect.com/`

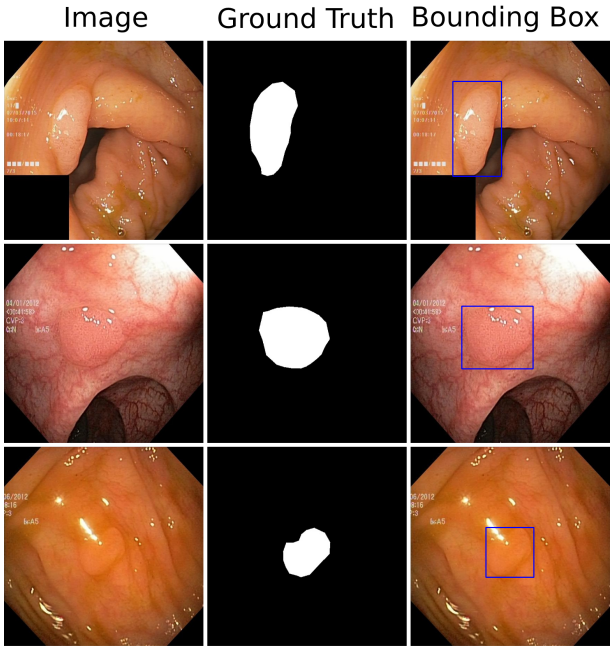


Figure 3.11: Polyps, their corresponding ground truth, and bounding box information from *kvasir-sessile* dataset

classification task. For the segmentation task, we have released 200 images. The dataset is made public for research and academic purposes. These classification datasets can be downloaded from [9] and the segmentation dataset can be downloaded from [10]. Figure 3.10 illustrates example frames for Endotect challenge dataset. We have created the ground truth with the help of expert gastroenterologists from our team. In our overview paper [71], we have presented comprehensive results of all the teams for all GI findings classification, efficiency and polyp segmentation task.

3.3.9 Kvasir-Sessile

Kvasir-Sessile [85] is not a novel dataset. It is a subset of Kvasir-SEG [89]. Flat adenomas, smaller diminutive, and sessile polyps are usually missed during colonoscopy examination [105]. Therefore, it was important to select such polyps and train model and benchmark algorithms. In this regard, we selected 196 such polyp frames from Kvasir-SEG[89]. After selection, we have put such images in the separate folder. The folder contains images, ground truth and bounding box. Therefore, the Kvasir-Sessile contains 196 polyp images, ground truth, and images showing the bounding boxes. The example of Kvasir-Sessile can be found in Figure 3.11. It consists of a polyp with a size of less than 10

[9] https://drive.google.com/file/d/19cBAyQuEBMfydKZIV0N1q8StJLmuQHWM/view
[10] https://drive.google.com/file/d/1LNpLkv5ZlEUzr_RPN5rdOHaqk0SkZa3m/view

mm. More information about the Kvasir-Sessile dataset can be found on the webpage[11]. The baseline experiments on the Kvasir-Sessile can be found in this paper [85].

3.4 Evaluation metrics

We have performed GI findings classification, polyp detection and polyp segmentation tasks. Here, we present the most commonly used computer vision metrics for polyp segmentation, GI image classification, and polyp detection task. Let us assume that for all of our tasks, *tp*, *fp*, *tn*, and *fn* represent true positives, false positives, true negatives, and false negatives, respectively.

In ML, the trained model is evaluated by metrics such as *tp*, *fp*, *tn*, and *fn*. In the context of polyp segmentation, *tp* means that the model has correctly predicted the pixel of the polyp classes. Similarly, *tn* in polyp segmentation means that the model has correctly predicted the background class. Likewise, *fp* means that the model has incorrectly predicted the polyp pixel as the background pixel and *fn* means that the outcome of the model incorrectly classified the background pixel as the polyp pixel.

3.4.1 Polyp segmentation

The most commonly accepted metrics for the polyp segmentation task are DSC, and mIoU. Additionally, we also calculated metrics such as recall, precision, overall accuracy and FPS.

$$\text{Dice coefficient } (DSC) = \frac{2 \cdot tp}{2 \cdot tp + fp + fn} \tag{3.1}$$

$$\text{Intersection over union } (IoU) = \frac{tp}{tp + fp + fn} \tag{3.2}$$

$$\text{Recall } (Rec) = \frac{tp}{tp + fn} \tag{3.3}$$

$$\text{Precision } (Prec) = \frac{tp}{tp + fp} \tag{3.4}$$

$$\text{F2} = \frac{5p \times r}{4p + r} \tag{3.5}$$

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

$$\text{Accuracy } (Acc) = \frac{tp + tn}{tp + tn + fp + fn} \tag{3.6}$$

$$\text{Frame per second } (FPS) = \frac{\#frames}{sec} \tag{3.7}$$

3.4.2 GI findings classification

We use metrics such as recall, precision, specificity, accuracy, *Matthews correlation coefficient (MCC)* and F1-score for the GI tract classification tasks. The definition of recall, precision, specificity and accuracy are the same as above. Here, we define the specificity, *MCC* and F1-score for the classification task. For the multi-class classification problem of the GI endoscopy image, we consider *MCC* as the most relevant metrics.

$$\text{Specificity } (Spec) = \frac{tn}{tn + fp} \tag{3.8}$$

$$\text{Matthews correlation coefficient } (MCC) = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}} \tag{3.9}$$

$$\text{F1-score } (F1) = \frac{precision \times recall}{precision + recall} \tag{3.10}$$

3.4.3 Polyp detection

The preferred metrics for the polyp detection task are average precision (AP) and mIoU. The definition of mIoU remains the same as the segmentation task. The AP or mean average precision (MAP) is used to evaluate the comparison of the predicted ground truth bounding boxes of the area of interest with the original ground truth of the area of interest.

$$\text{AP} = \sum_{n} \{(r_{n+1} - r_n)\, p_{\text{interp}}(r_{n+1})\}, \tag{3.11}$$

with $p_{\text{interp}}(r_{n+1}) = \max_{\tilde{r} \geq r_{n+1}} p(\tilde{r})$. In the equation, $p(r_n)$ represents the precision value at a given recall value. Additionally, we calculate APs for mIoU with threshold of 0.25, 0.5, 0.75. Moreover, we also calculate FPS.

3.5 Summary

In this chapter, we have shown the existing dataset in the field of GI endoscopy and pointed out the lack of public datasets in the field. We have discussed the data collection and annotation procedure. Additionally, we have briefly introduced our collected and annotated datasets. Moreover, we have also presented the evaluation metrics used for evaluating the model's performance for the classification, detection and segmentation tasks.

Lack of medical dataset has restricted the research and development in the field of GI endoscopy despite the potential of the ML algorithm to improve the clinical procedure. Because of the lack of datasets, there is no possibility of testing an existing algorithm or developing new solutions. One of the reasons why the dataset is not available is that medical datasets are challenging to collect requiring a team of medical experts from the hospitals and computer scientists. Moreover, it requires prior approval from the hospital and patient consent. Medical experts (gastroenterologists) are usually less accessible for research purposes. Thus, data collection may not be possible for all medical image analysis researchers. Even if the dataset is collected in the community, the research community often does not share their dataset publically, due to their agreement with the hospital or some other reasons.

Towards this end, we have collected, and publicly released world's largest datasets for GI conditions classification (HyperKvasir[26]) and VCE (Kvasir-Capsule [180]. Additionally, our effort has led to the collection and public release of Kvasir-SEG [89], Kvasir-Instrument [88], KvasirCapsule-SEG [92], Kvasir-Sessile [85]), PolypGen [5], Medico automatic polyp segmentation dataset [91], and Endotect challenge dataset [71]). Much of the research community has already adopted our datasets for the research and development of novel DL methods. We conjecture that the open access and publicly available datasets will be helpful to accelerate research in both academia and industry. Moreover, our datasets could play a crucial role in the development of AI technology for the VCE, colonoscopy, and GI endoscopy.

In the next chapter, we will present our designed algorithms used for the GI tract classification, polyp segmentation and detection, and surgical instrument segmentation. We will show the results on datasets presented in this chapter. Additionally, we will also introduce the challenges and competitions organized, which inspired us to generate and develop the presented datasets.

Chapter 4

Classification, detection, and segmentation

In this chapter, we briefly introduce our proposed DL models to address the research objectives. Here, we will introduce some of the proposed works and CNN based architectures for automatic GI tract findings classification, polyp detection, polyp segmentation, and surgical instrument segmentation in laparoscopy. For the experimentation, we have mainly used our collected datasets presented in the earlier chapter. More details about each architecture and dataset can be found in the respective papers listed in the appendix.

4.1 Classification models for GI findings

We approach classification of GI tract findings from both upper and lower GI tract using ML techniques. The findings include images from the anatomical landmark (z-line, pylorus, cecum), pathological findings (esophagitis, polyps, ulcerative colitis), polyp removal cases (dyed and lifted polyps, dyed resection margins), and normality and regular findings (normal colon mucosa, stool). We have used images from the Medico challenge [149], and extracted GFs using Lire [125]. We have used GFs such as joint composite descriptor (JCD), tamura, color layout (CL), edge histogram, and pyramid histogram of oriented gradients (PHOG). The extracted features were sent to the simple logistic classifier and logistic model tree classifier. We have used Weka [61]. Weka has the collection of ML algorithms for classification, regression, clustering etc. Later on, the trained models were used to compute the prediction on the new test datasets provided by the task organizers. The predictions were sent to the challenge organizers. The official results showed that our model achieved MCC score of 0.8353 with the simple logistic classifier and 0.8350 with the

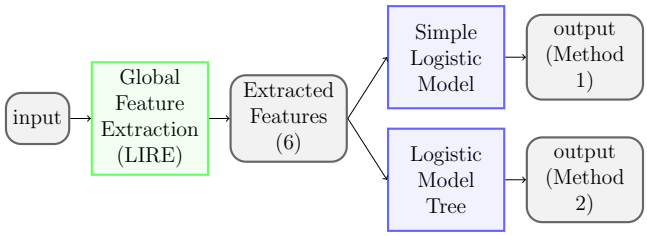


Figure 4.1: Block diagram of the proposed methods for multi-class GI tract findings classification [84]

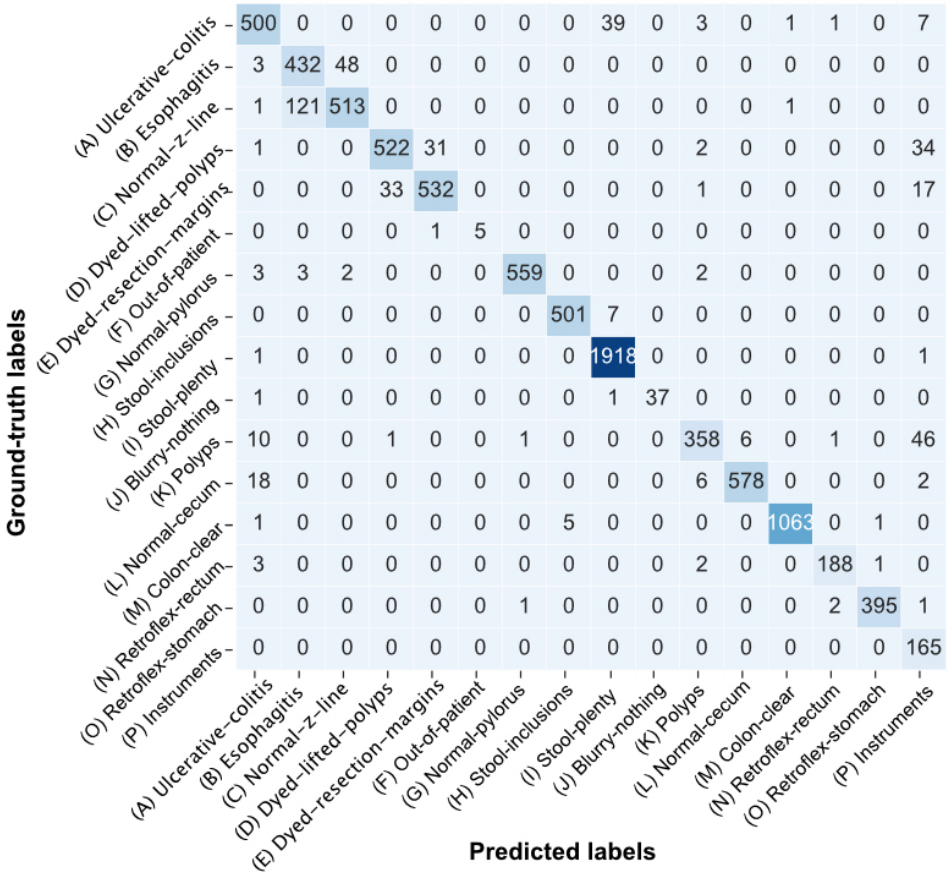


Figure 4.2: Confusion matrix plot of the best results. A-P represents class labels [84]

logistic model tree classifier [193]. We also presented CNN and a transfer learning based approach. We obtained the best results of 0.9421 and ranked 2^{nd} in the competitions. The confusion matrix of the best results can be shown in Figure 4.2. All of these five models were tested across various datasets. The explanation about the GFs, datasets, classifiers, experimental setup, and results can be found in our study [190].

Additionally, we have presented a comprehensive study on the automated classification methods for multi-class GI tract findings. The study is based on Medico GI challenges

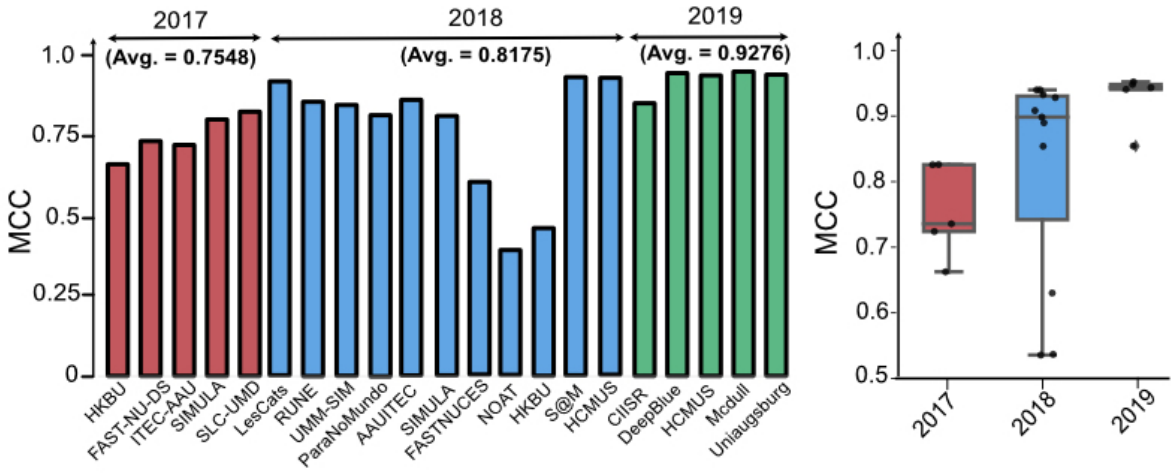


Figure 4.3: MCC comparison of the 21 participating teams in Medico 2017, 2018, and BioMedia 2019 challenge [84]

(Medical Multimedia Task at MediaEval 2017, Medico Multimedia Task at MediaEval 2018, and BioMedia ACM MM Grand Challenge 2019). In this study, we have dissected the three challenges and highlighted the strength and weaknesses of the 21 methods presented in 3 consecutive years. The average MCC score for 2017 is 0.7548, 2018 is 0.8175, and 2019 is 0.9276. The individual and the average MCC can also be observed in Figure 4.3.

We have also dissected the methods based on their credibility to be used in clinical settings. For this, we have calculated the clinical ranking score for each method. The clinical applicability rank is presented in Table 4.1. The clinical applicability rank is based on MCC, efficiency, and processing speed. We have calculated the clinical applicability rank based on the algorithm performance through frame-level classification score and efficiency (FPS). In the table, we refer 'Clas.' as MCC classification, 'AR' as MCC Algorithm Robustness, 'RR' as Robustness-rank, 'SR' as Speed-rank, 'Rank' as MCC Rank, 'CAR' as Clinical applicability rank, and 'na' as not available. 10 is the imputed rank for speed and robustness ranking. More detailed information about the ranking can be found in our paper [84].

In addition to the classification and efficiency task, we have also evaluated the "automatic report generation task", where the participants were asked to generate the endoscopic procedure's text report. Moreover, we have also evaluated the results for the hardware task, where the participants were asked to submit docker images for their submission to benchmark the proposed algorithms on the same hardware. The detailed information about all the Medico challenges can be found here [164, 149, 68]. Detailed information about each teams' method, automatically generated report, their experimental

Table 4.1: Clinical applicability of the participants methods [84]

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**
2019	**CIISR**	**0.8542**	**0.8542**	**98.9**	**1**	**1**	**12**	**3**
	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**

setup, results, analysis of failed classes, clinical applicability, and critical discussion can be found in our paper [84].

4.2 Polyp detection

In this work, we aim to accurately locate the instances of polyps in the colonoscopy image datasets using different popular detection methods. We have used object detection algorithms to predict bounding boxes of the object classes. There are two-stage detectors and one-stage detectors. In the two stage detectors (for example, R-CNN [57], Fast R-CNN [56], Faster R-CNN [159] or Mask R-CNN [66], (i) the region proposals are identified, (ii) in the second stage of the network, the objects are classified within the region proposals along with the bounding box regressions. In one-stage detectors, (for example, you-only-look-once (YOLO)v2 [156], single shot multiBox detector (SSD) [120]), anchor boxes are used for predictions across the entire image that for the regions that the CNN network predicts. Figure 4.4 shows the example of one-stage object detection and localization methods. The one-stage detectors are usually faster than two-stage detectors. Although one-stage detector models have lower accuracy, these methods have attracted

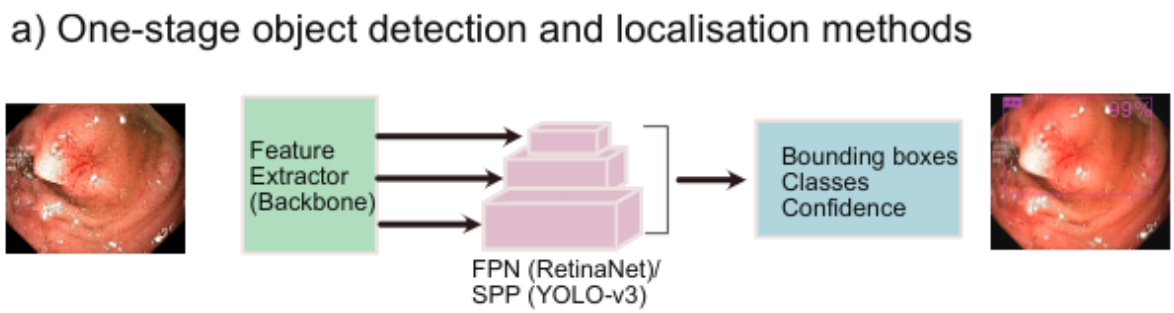


Figure 4.4: One stage object detection methods [93]

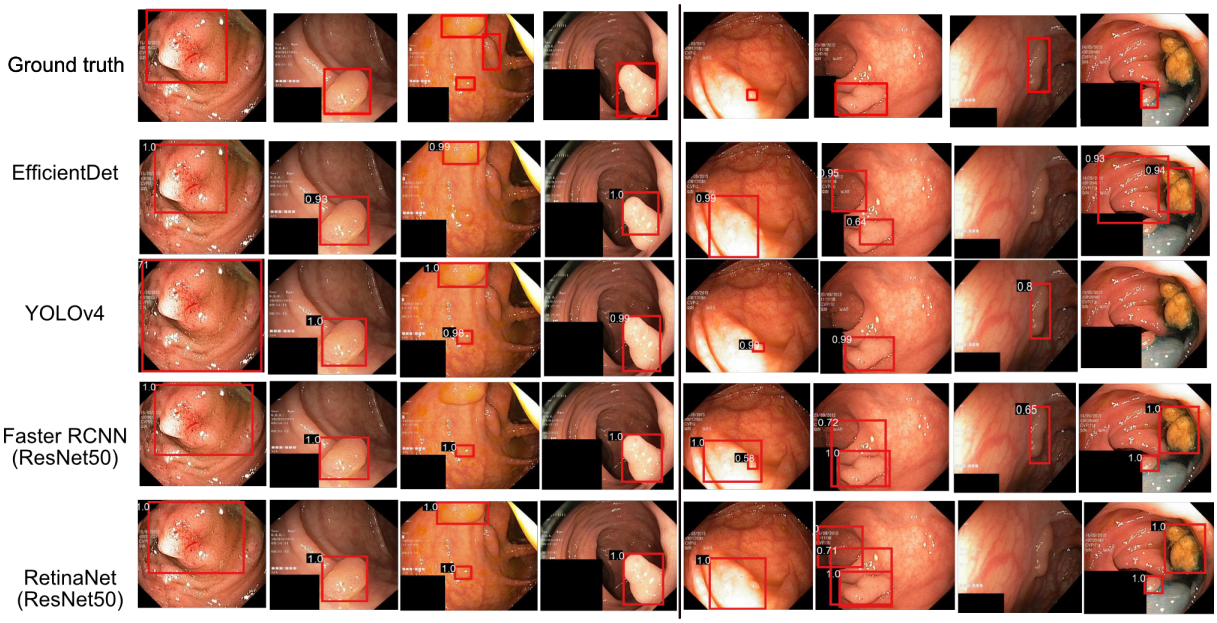


Figure 4.5: Detection results on the test set of Kvasir-SEG dataset [93].

attention in the community because of their ability to achieve higher speed and optimal accuracy.

In our work, we have used both two-stage detectors such as (Faster R-CNN [159]) and one-stage detectors (for example, RetinaNet [118], YOLOv3 [157], YOLOv4 [24]) for the comparison. Figure 4.5 shows the results of the detection methods of EfficientDet-D0, YOLOv4, Faster R-CNN and RetinaNet (with ResNet50 backbone). The confidence score of each method on each image is provided. The figure is divided into two parts. The left side of the border shows the best results obtained by the presented methods. Here, the methods obtained the highest mIoU and have a similar type of results in all of the cases. On the right side, the results show the example of the failure cases on the presented algorithm.

Table 4.2: Performance comparison on polyp detection task on the Kvasir-SEG. We have highlighted two best scores [93]

Method	Backbone	AP	IoU	AP_{25}	AP_{50}	AP_{75}	FPS
EfficientDet-D0 [189]	EfficientNet-b0, biFPN	0.4756	0.4322	0.6846	0.5047	0.2280	35.00
Faster R-CNN [159]	ResNet50	0.7866	0.5621	0.8947	0.8418	0.5660	8.00
RetinaNet [118]	ResNet50	**0.8697**	0.7313	**0.9395**	**0.9095**	0.6967	16.20
RetinaNet [118]	ResNet101	**0.8745**	0.7579	**0.9483**	**0.9095**	0.7132	16.80
YOLOv3+spp [157]	Darknet53	0.8105	**0.8248**	0.8856	**0.8532**	**0.7586**	45.01
YOLOv4 [24]	Darknet53, CSP	0.8513	0.8025	0.9123	0.8234	**0.7594**	**48.00**
ColonSegNet (Proposed)	-	0.8000	**0.8100**	0.9000	0.8166	0.6706	**180.00**

The proposed ColonSegNet [93] is basically an end-to-end segmentation network. We have converted the predicted mask from the ColonSegNet into the bounding boxes and evaluated the prediction score for the polyp detection task. Table 4.2 shows the quantitative results on polyp detection tasks. From the results, we observe that ColonSegNet produces a competitive AP of 0.8000, and mIoU of 0.8100. However, ColonSegNet achieves the highest processing speed of 180 FPS, which is nearly 3.5 times that of the baseline YOLOv4 [24]. A detailed explanation about the hyperparameters, architectural design, and datasets can be found in our paper [93].

4.3 Polyp segmentation

In the last section, we have explained our automatic polyp detection method. The polyp detection method can indicate the area of interest for the potential polyp or polyps. However, it does not provide any information about each pixel of the polyp region, which is crucial for gastroenterologists. Polyp segmentation algorithms can predict each pixel of the polyp from an input image. Therefore, most of our research is focused on automatic polyp segmentation. We have designed architectures such as ResUNet++ [94], DoubleUNet [86], ResUNet++ + CRF + TTA [85], ColonSegNet [93], NanoNet [92], DDANet [195], and PNS-Net [95]. All of these architectures are endcoder-decoder networks.

4.3.1 ResUNet++

We have designed the ResUNet++ [94] architecture, especially for polyp segmentation. It is an encoder-decoder network. Our architecture is inspired by ResUNet [223]. The

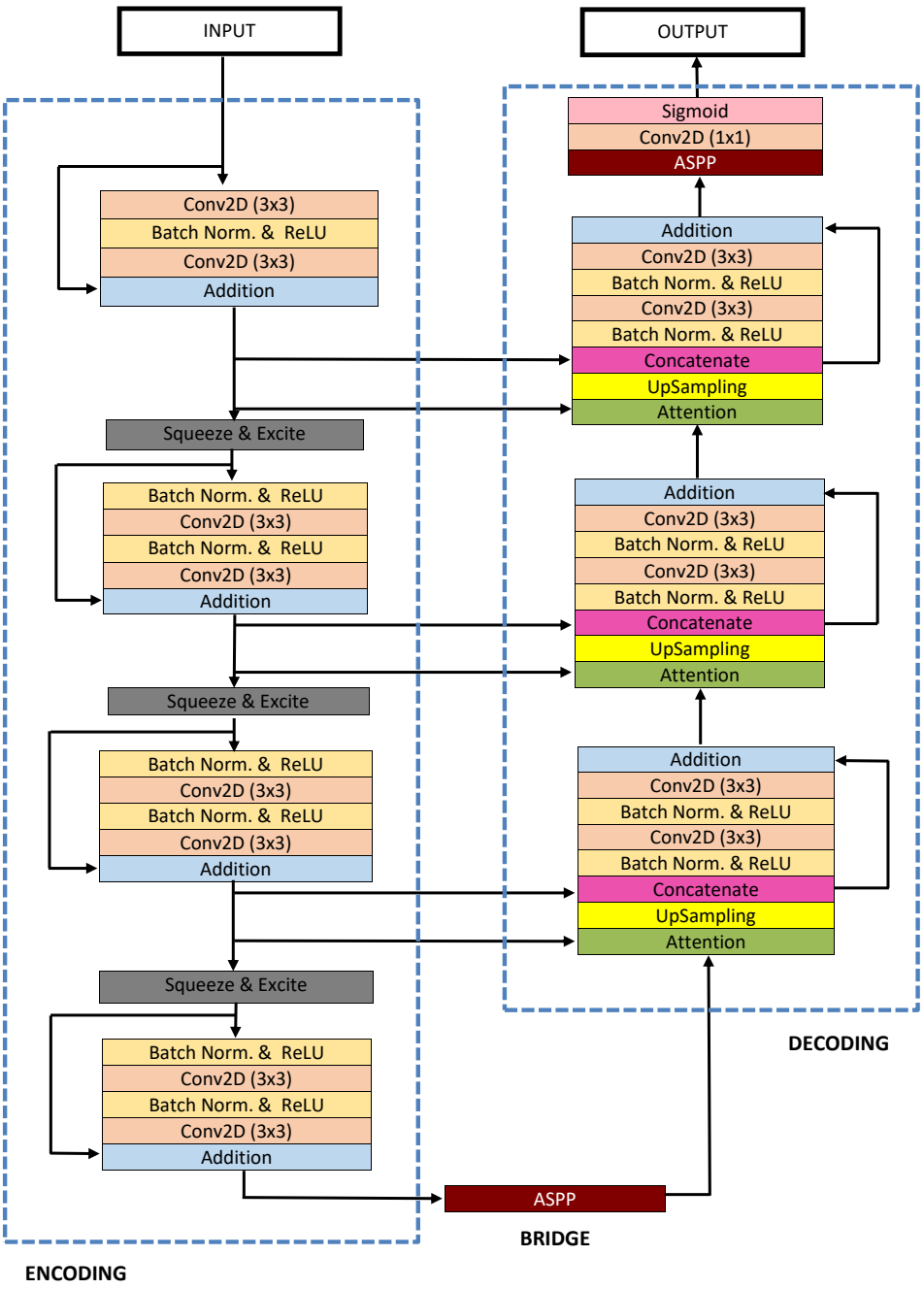


Figure 4.6: ResUNet++ architecture [94]

designed architecture uses residual block, squeeze and excitation block [75], atrous spatial pyramid pooling [35], and attention block [200]. The block diagram of the ResUNet++ architecture is shown in Figure 4.6. In the proposed architecture, we have introduced a sequence of squeeze and excitation blocks as shown in Figure 4.6. We have also used atrous spatial pyramidal pooling (ASPP) to connect the encoder and decoder part of the network.

In the decoder part of the network, we introduce a novel attention block that is used

Table 4.3: Performance comparison of proposed models on Kvasir-SEG [85]

Method	DSC	mIoU	Recall	Precision
UNet [167]	0.7147	0.4334	0.6306	**0.9222**
ResUNet [223]	0.5144	0.4364	0.5041	0.7292
ResUNet-mod [223]	0.7909	0.4287	0.6909	0.8713
ResUNet++ [94]	0.8119	0.8068	0.8578	0.7742
ResUNet++ + CRF	0.8129	0.8080	0.8574	0.7775
ResUNet++ + TTA	0.8496	0.8318	**0.8760**	0.8203
ResUNet++ +TTA + CRF	**0.8508**	**0.8329**	0.8756	0.8228

Table 4.4: Performance comparison of proposed models on CVC-VideoClinicDB [85]

Method	DSC	mIoU	Recall	Precision
ResUNet++	0.8798	0.8730	**0.7749**	0.6702
ResUNet++ + CRF	**0.8811**	**0.8739**	0.7743	**0.6706**
ResUNet++ + TTA	0.8125	0.8467	0.6896	0.6421
ResUNet++ + TTA + CRF	0.8130	0.8477	0.6875	0.6276

in each decoder block. We have used nearest-neighbour up-sampling. The output of the decoder block is concatenated with the feature map of the residual block from the encoder through skip connection. The residual unit with identity mapping follows this process. We have also used a series of skip connections from the residual unit of the encoder section to the attention block of the decoder section. We specify the number of filters to [32, 64, 128, 256, 512] in the encoder section. The sequence is reversed in the decoder part of the network, and the order ultimately becomes [512, 256, 128, 64, 32].

The proposed architecture consists of one stem block and three encoder blocks in the encoder part of the network. ASPP lies between encoder and decoder, and there are three decoder blocks. Skip connections are used for information propagation between encoder and decoder. The output of the last decoder block is passed to the ASPP. The output is now passed to the 1×1 convolution followed by the sigmoid activation function. Batch normalization is used for all the convolution layers except the output layer. The final output is a binary segmentation mask. A detailed explanation about the residual block, squeeze, and excitation blocks, ASPP, attention units can be found in our work [94, 85].

4.3.2 Extension of the ResUNet++

ResUNet++ showed promising results and was adopted as the baseline method for results comparison in other research articles. With this motivation, we further explored the possibility of improvement of ResUNet++ and evaluated it with several publicly available datasets. For this, we used conditional random field (CRF) and test-time augmentation

Table 4.5: Performance comparison of the models on ASUMayo Clinic database [85]

Method	DSC	mIoU	Recall	Precision
ResUNet++ [94]	0.8743	0.8569	**0.6534**	0.4896
ResUNet++ + CRF	**0.8850**	**0.8635**	0.6504	0.4858
ResUNet++ + TTA	0.8553	0.8535	0.6162	**0.4912**
ResUNet++ + TTA + CRF	0.8550	0.8551	0.6107	0.4743

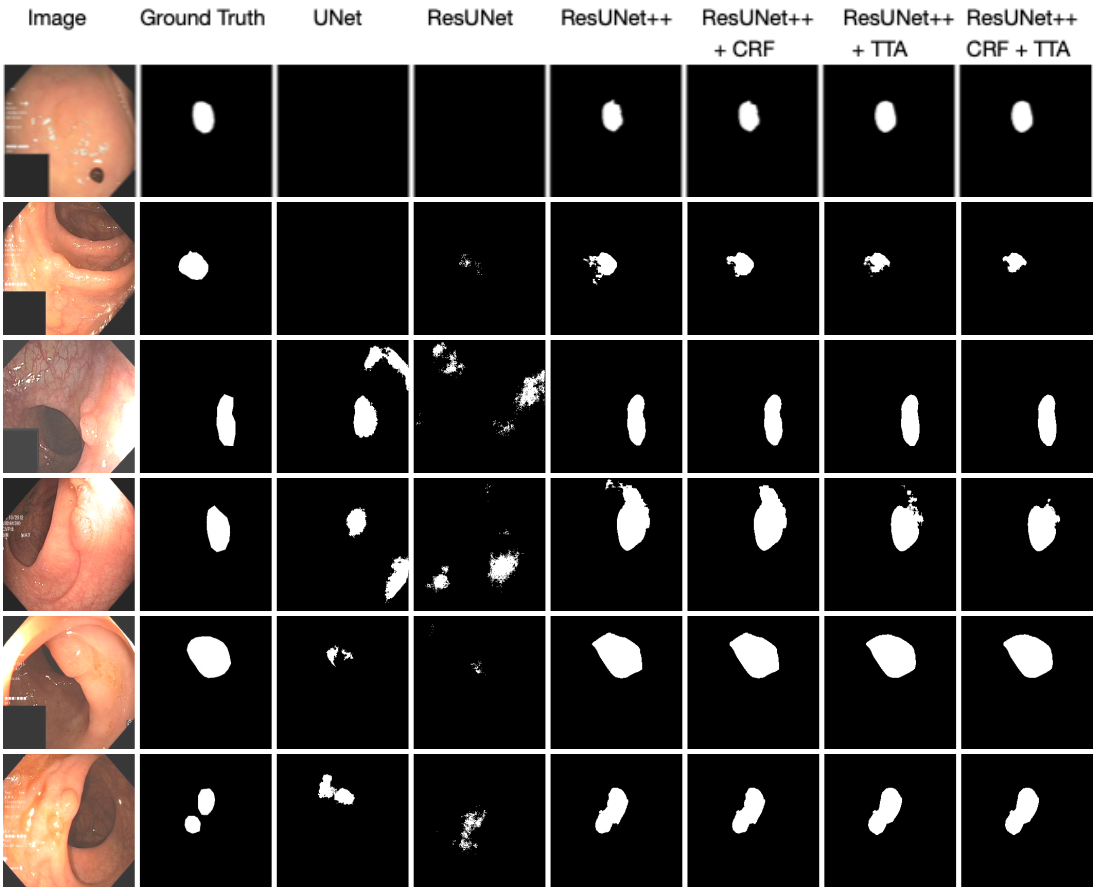


Figure 4.7: Qualitative comparison of our proposed method with the baseline methods [85]

(TTA) for further results improvement. We performed comprehensive experiments using four publicly available still image datasets and two video datasets. The still images dataset have at least one polyp contained in an image, whereas the video dataset has both positive samples containing polyp and negative samples without polyps. Additionally, to increase the training samples, we have performed data augmentation such as horizontal flip, vertical flip, center crop, compose, transpose, elastic transform, grid distortion, optical distortion, random brightness, random contrast, gaussian blur, gauss noise, channel shuffle, coarse dropout etc. An extensive hyperparameter search mechanism was applied to find the optimal hyperparameter. Moreover, we also evaluate standard computer vision metrics for the ResUNet++, CRF, and TTA used in our work [85].

We perform a cross-dataset test to explore the generalizability of the polyp segmenta-

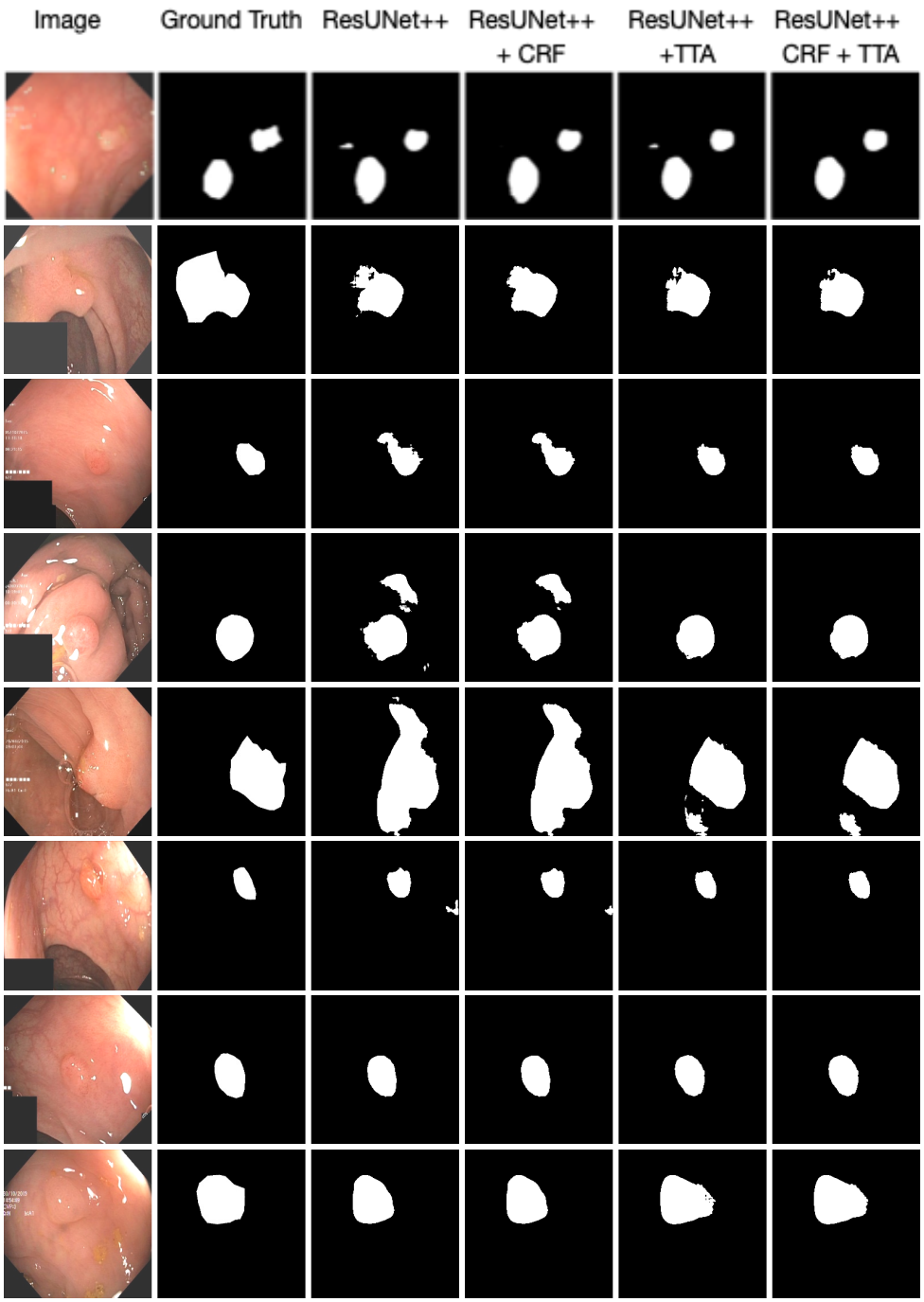


Figure 4.8: Qualitative result comparison of models that are trained on CVC-ClinicDB dataset and tested on Kvasir-SEG [85]

tion. For this, we have trained on polyp datasets from one centre and tested them across other publicly available datasets. Additionally, we mixed the datasets from two different center and tested it across the new datasets. Moreover, we experimented with the Kvasir-Sessile dataset, a subset of Kvasir-SEG, that contained flat or sessile polyps to evaluate the effectiveness of our method on the polyps that are usually overlooked by endoscopists during endoscopic examination. Table 4.3, Table 4.4 and Table 4.5 show the results after

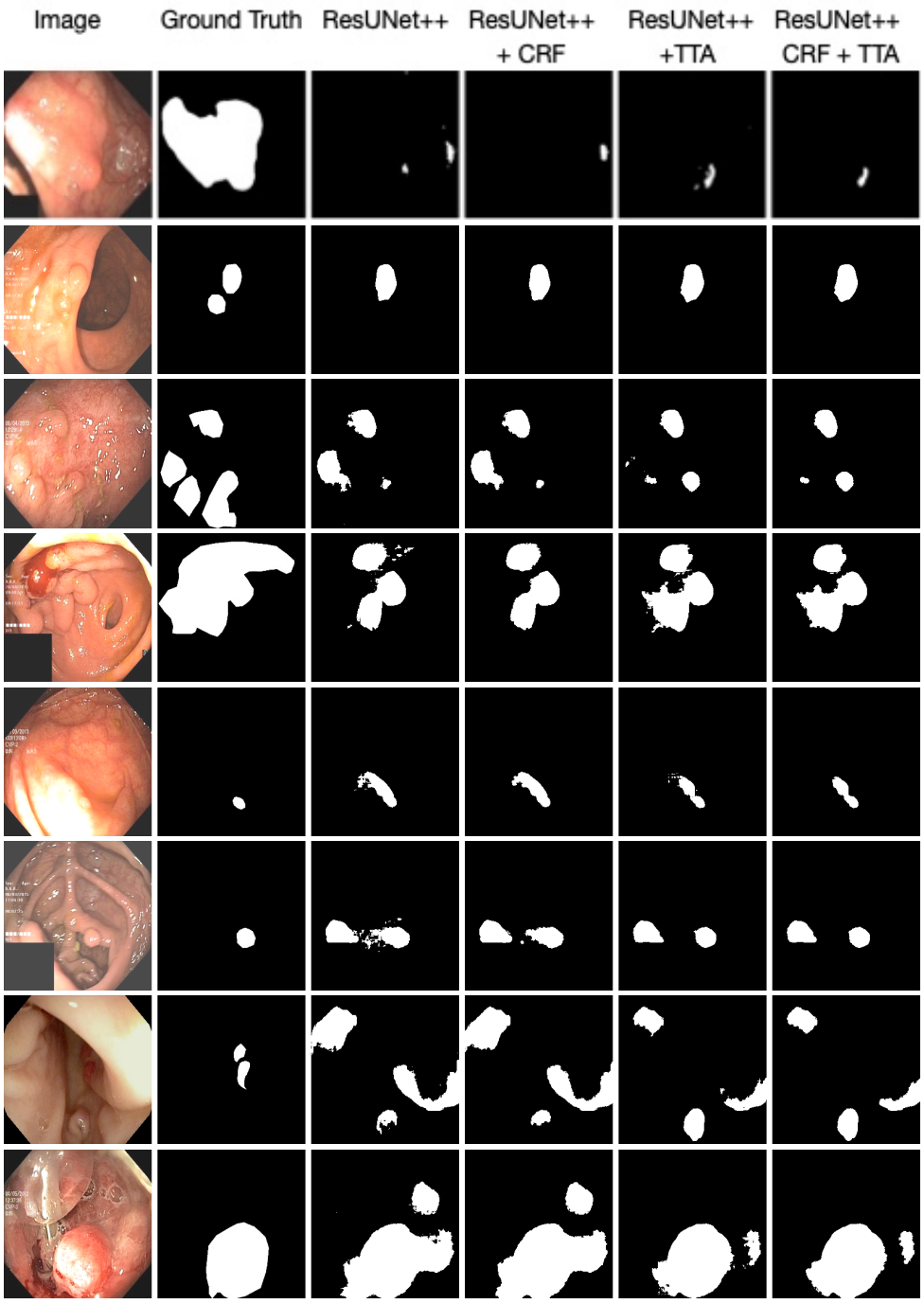


Figure 4.9: Example shows the failing cases on Kvasir-SEG, where ResUNet++, and its extension fails [85]

using CRF and TTA. More results on different individual datasets and cross-dataset can be found in our paper [85]. We have also presented the qualitative results in addition to the quantitative results.

We present the qualitative results comparison of the proposed ResUNet++, CRF, and TTA with it competitor networks such as UNet, ResUNet, and ResUNet++ architecture in Figure 4.7. Here, we have selected flat or sessile polyps as the use case. Flat or sessile

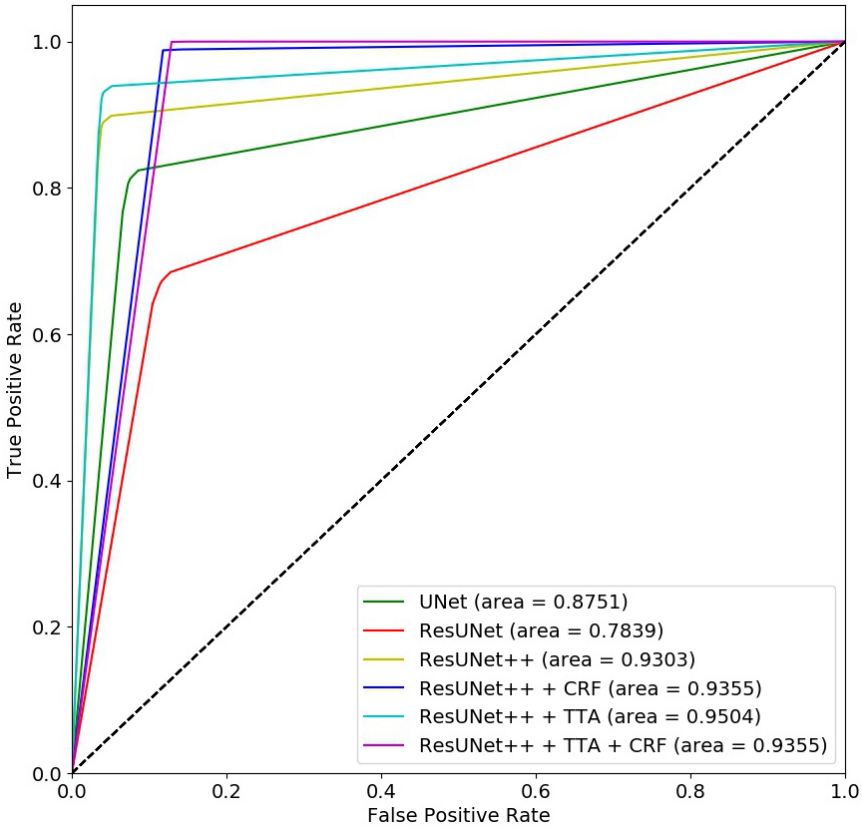


Figure 4.10: ROC curve of proposed and baseline models on the Kvasir-SEG [85]

polyps are usually overlooked during the colonoscopy examination, which can turn out into CRC at the later stage. Our proposed method "ResUNet++ + CRF + TTA" shows better results as compared to UNet, ResUNet, and ResUNet++. Additionally, from the qualitative results, we observe that our method has a high similarity between the ground truth and predicted masks of the polyp sample.

Similarly, Figure 4.8, shows qualitative comparison of the cross-dataset test results. From the qualitative results, we can observe that even for the cross-dataset test (i.e., model trained on CVC-ClinicDB and tested on Kvasir-SEG), the proposed method shows promising results with different types polyp samples. However, it is to be noted that these are among the best examples of the predicted masks. The quantitative results of the model trained on CVC-ClinicDB and tested on the Kvasir-SEG dataset can be found in our paper. Moreover, we also show the cases, where our algorithm fails in Figure 4.9.

From the Figure, we can see that the models usually fail when there are multiple polyps in a single frame or fails when there is a polyp like structure but not a polyp. More training data will be required to develop accurate and robust models. Figure 4.10 shows the receiver operating curve (ROC) curve of different models trained on the Kvasir-SEG dataset. From the ROC curve, we observe that our proposed model "ResUNet++

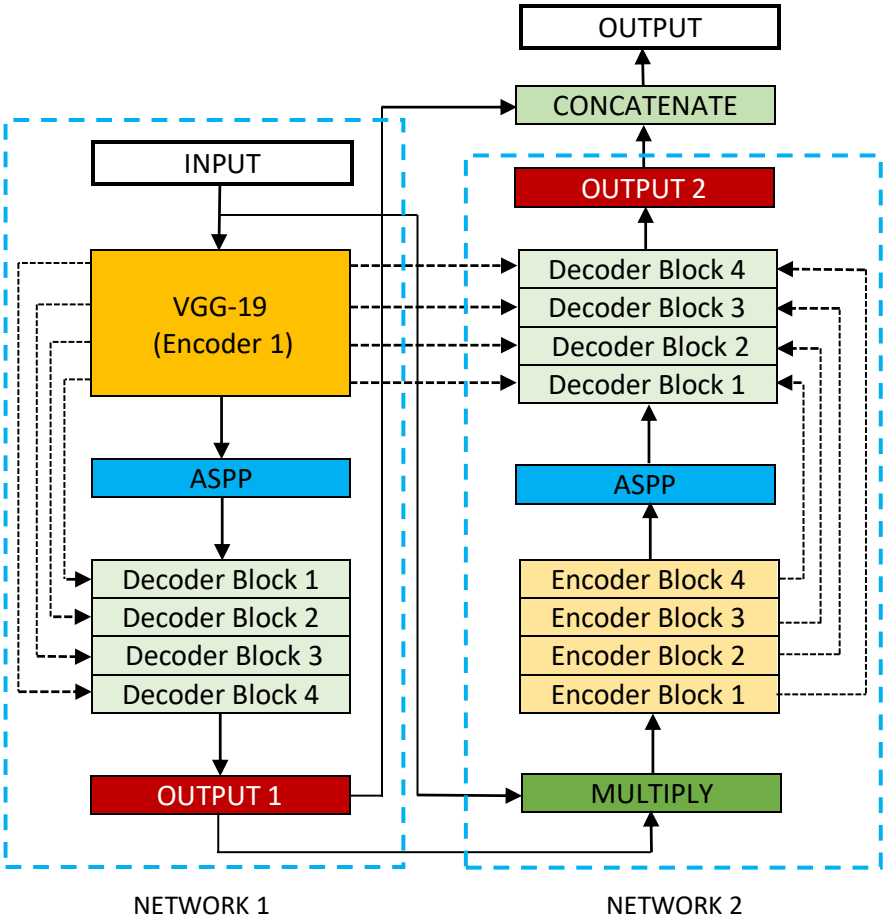


Figure 4.11: Block diagram of the proposed DoubleU-Net architecture [86]

+ TTA", has the value of 0.9504, which is 2.01% more than ResUNet++. From all of the above results (qualitative, quantitative, and ROC curve), we have shown that our proposed model can segment polyp of different sizes and shapes. Additionally, our models showed better performance on flat or sessile polyps. Our model outperformed other SOTA methods on the various publicly available datasets.

4.3.3 DoubleUNet

Inspired by the success of UNet [167], for medical image segmentation tasks, we have proposed DoubleUNet [86]. In our proposed network design, we have used two modified UNet architectures. Thus, it is termed as DoubleUNet. In the first encoder, we have used VGG-19 [178] as the pretrained encoder. VGG-19 is selected because it is a lightweight model compared to other pre-trained Imagenet [39] models. Also, the concatenation between UNet and VGG-19 is more straightforward. Additionally, a deeper segmentation network can potentially produce better segmentation output.

Figure 4.11 shows the block diagram of DoubleUNet. DoubleUNet has two networks,

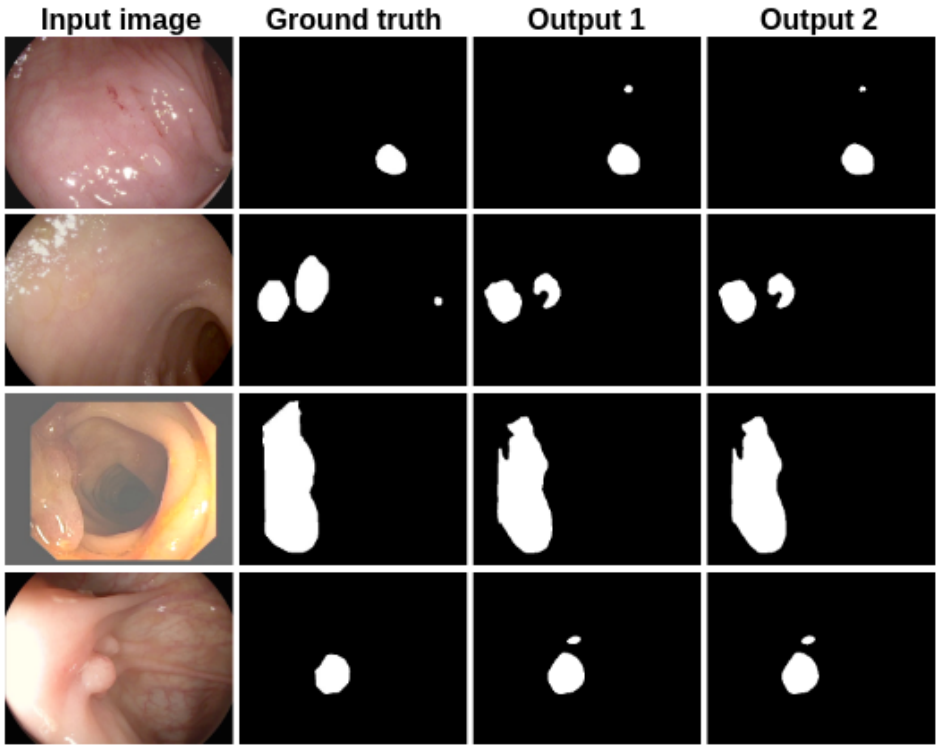


Figure 4.12: Qualitative result comparison between initial output and final output of DoubleU-Net on CVC-ClinicDB [86]

Table 4.6: Result comparison on CVC-ClinicDB [86]

Method	DSC	mIoU	Recall	Precision
Fully Convoutional Network [114]	-	-	0.7732	0.8999
CNN [140]	(0.62-0.87)	-	-	-
SegNet [205]	-	-	**0.8824**	-
Multi-scale patch-based CNN [14]	0.8130	-	0.7860	0.8090
MultiResUNet with data augmentation [79]	-	0.8497	-	-
Conditional generative adversarial network [151]	0.8848	0.8127	-	-
U-Net	0.8781	0.7881	0.7865	0.9329
DoubleU-Net	**0.9239**	**0.8611**	0.8457	**0.9592**

namely NETWORK 1 and NETWORK 2. In NETWORK 1, input is passed through VGG-19 (encoder 1), ASPP, and decoder blocks, which generates the predicted output masks (*Output 1*). The *Output 1* acts as an input to NETWORK 2, where element-wise multiplication is performed between original input and *Output 1*. After passing through the different encoder blocks, ASPP and decoder blocks, we get *Output 2*. Finally, the *Output 1* and *Output 2* are concatenated to observe the qualitative difference between the intermediate mask (*Output 1*) and final predicted output mask (*Output 2*).

We assumed that the generated output feature map from NETWORK 1 could be im-

proved further by fetching the input image along with the corresponding masks and its concatenation with the output of network 2 (i.e., *Output 2*). The concatenation can potentially produce better segmentation output masks utilizing two modified U-Net based architectures. This is the main idea behind choosing the architectural design of DoubleUNet. An explanation about the ASPP, squeeze and-excitation block, encoder, and decoder blocks used in DoubleUNet can be found in our paper [86].

Table 4.6 shows the comparison of the quantitative results of DoubleUNet with other recent methods. From the quantitative results, we can observe that DoubleUNet produces DSC of 0.9239, and mIoU of 0.8611, which is almost 4% higher than SOTA, Poomeshwaran et al. [151] and around 1.5% in mIoU than in Ibtehaz et al. [79]. We have experimented with four different datasets. The other results can be found in our paper[86]. Similarly, in Figure 4.12, we have shown the comparison of the results of the outputs from NETWORK 1 and the final output (*output 2*). From the comparisons of the qualitative results between *output 1* and *output 2*, we observe that both of the networks produce over-segmentation on challenging images. However, careful observation can show that the results of final output masks are better than the intermediate masks (i.e., *output 1*).

4.3.4 ColonSegNet

ColonSegNet [93] is an encoder-decoder based network like ResUNet++ [94] and DoubleUNet [86]. It mainly uses residual block [65], and squeeze and excitation network [75]. While designing the network, we have considered the number of parameters. Real-time performance can be achieved with the lightweight model, so we aim to maintain a low number of parameters. ColonSegNet has only a few trainable parameters as compared to the other popular semantic segmentation architectures.

Figure 4.13 illustrates the block diagram of ColonSegNet. Our proposed architecture consists of two encoder blocks. The encoder block extracts all the essential information from the provided input image. This necessary information is passed to the decoder block. Each decoder block is connected with two skip connections from the encoder block. The first skip connection is simple concatenation. The second skip connection from the encoder is passed via the transpose convolution for incorporating multiscale features with the decoder block. The multiscale features acquired through the encoder helps the decoder in generating better semantic and meaningful information that can be observed in the form of segmentation masks.

At first, the input image of size (512, 512, 3) is fed into the first encoder. The first

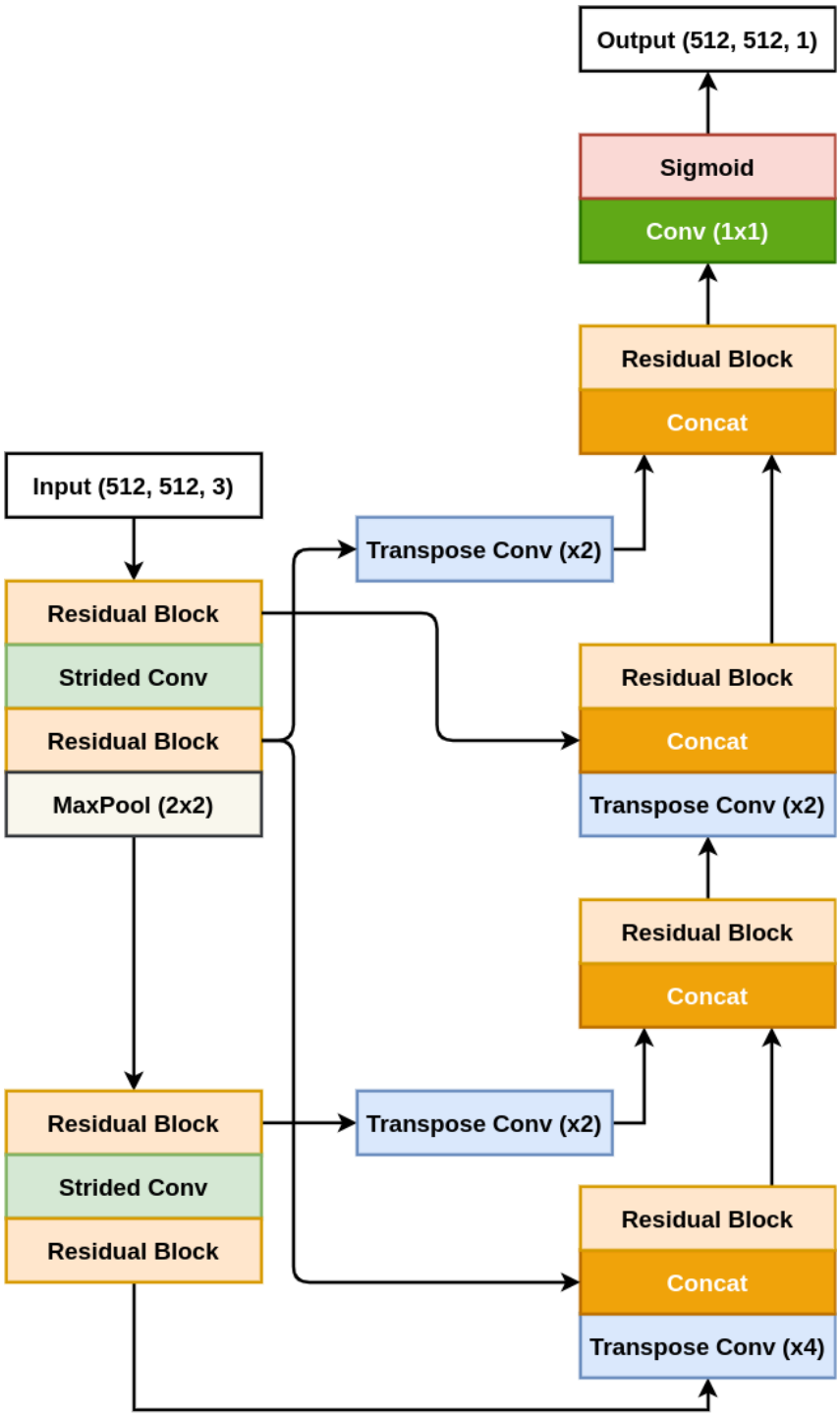


Figure 4.13: Block diagram of ColonSegNet [86]

encoder consists of two residual blocks and a 3×3 strided convolution in between them. A 2×2 maxpool layer follows the residual block. After the maxpool layer, the output feature map spatial dimension is contracted to $\frac{1}{4}$ to that of the input image. The components in the second encoder block are two residual blocks and strided convolution in between them.

The decoder block starts with transpose convolution. Here, the decoder uses a stride

Table 4.7: Performance comparison on Kvasir-SEG [93]

Method	Backbone	Jaccard C.	DSC	F2-score	Precision	Recall	Overall Acc.	FPS
UNet [167]	-	0.4713	0.5969	0.5980	0.6722	0.6171	0.8936	11.01
ResUNet [223]	-	0.5721	0.6902	0.6986	0.7454	0.7248	0.9169	14.82
ResUNet++ [94]	-	0.6126	0.7143	0.7198	0.7836	0.7419	0.9172	7.01
FCN8 [123]	VGG 16	0.7365	0.8310	0.8248	0.8817	0.8346	0.9524	24.91
HRNet [204]	-	0.7592	0.8446	0.8467	0.8778	0.8588	0.9524	11.69
DoubleUNet [86]	VGG 19	0.7332	0.8129	0.8207	0.8611	0.8402	0.9489	7.46
PSPNet [224]	ResNet50	0.7444	0.8406	0.8314	0.8901	0.8357	0.9525	16.80
DeepLabv3+ [34]	ResNet50	0.7759	0.8572	0.8545	0.8907	**0.8616**	**0.9614**	27.90
DeepLabv3+ [34]	ResNet101	**0.7862**	**0.8643**	**0.8570**	**0.9064**	0.8592	0.9608	16.75
UNet [167]	ResNet34	**0.8100**	**0.8757**	**0.8622**	**0.9435**	**0.8597**	**0.9681**	**35.00**
ColonSegNet (Proposed)	-	0.7239	0.8206	0.8206	0.8435	0.8496	0.9493	**182.38**

of 4. Now, the feature map spatial dimension is increased by 4. Likewise, the second decoder of our network utilizes a stride value of 2, which increases the spatial dimensions

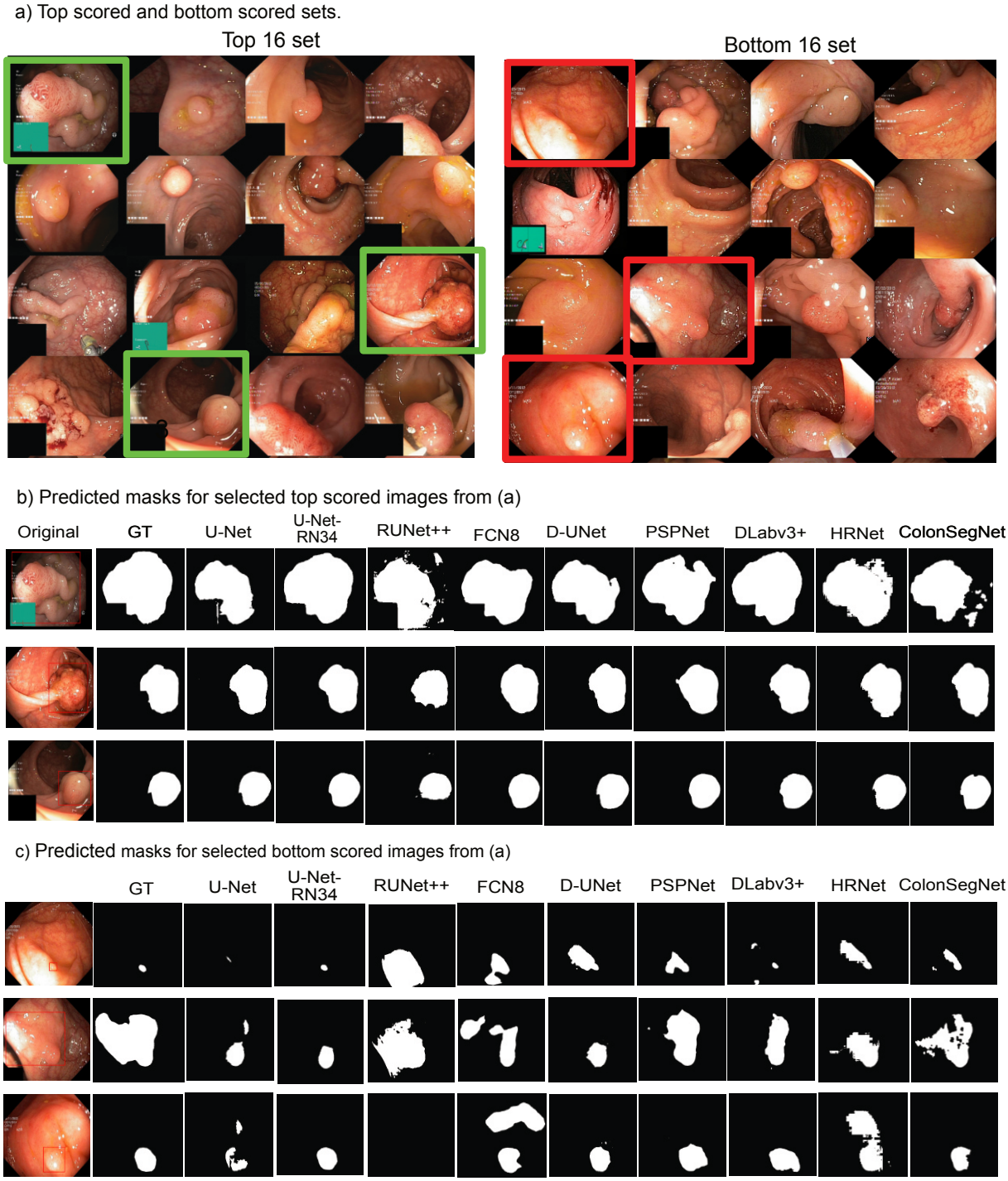


Figure 4.14: Example images show the best and worst-performing polyp samples : a) 16 top set (left) and bottom 16scored sets (right), b) example images, corresponding ground truth and their predictions for top-scored images and c) example images, its ground truth and their prediction for the bottom scored images. Green rectangles highlight the selected images sample from the top-scored set, and red rectangle highlights those from the bottom scored set. Here, we assume, UNet-RN34 = UNet-ResNet34, RUNet++ = ResUNet++, D-UNet = Double UNet, DLabv3+ = DeepLabv3+ (ResNet50) [93]

by the factor of 2. It is followed by simple concatenation and a residual block. After this, the output is concatenated with the second skip connection and are passed through the residual block. The output of the final decoder block is passed through a 1×1 convolution and sigmoid activation function, which produces binary segmentation masks [86].

In Table 4.7, we present the quantitative results of ColonSegNet and its competitors. We have comapared ColonSegNet, from baseline architectures (example, FCN [123] and UNet[167]) to the widely used computer vision architectures such as DeepLabv3+ [34], HRNet [204], PSPNet [224], along with the different backbones. From the Table, we can observe that ColonSegNet achieves competitive DSC of 0.8206, mIoU of 0.7239, F2-score of 0.8206, precision of 0.8496, recall of 0.8496, accuracy of 0.9493, and a FPS of 182.38. Here, it is to be noted that ColonSegNet does not use any pre-trained encoders, whereas UNet, which has a maximum DSC of 0.8747, uses ResNet34 as pretrained encoder. The results show that our architecture can produce decent performance metrics and a high processing speed of 182.38, which is more than 5 times compared with UNet-ResNet34.

Figure 4.14 shows the example images from the best- and worst-performing polyp samples. This figure can be divided into three sub-figures. The first sub-figures shows top and bottom scored sets. Sub-figure (b) shows the qualitative results on the top-scored images. The results show the efficiency of ColonSegNet. From the qualitative results, we can observe that even DeepLabv3+ performs well. Similarly, in sub-figure (c), we can observe the qualitative results on the bottom scored images. ColonSegNet shows over-segmentation for some examples, whereas under-segmentation for some samples. However, the qualitative comparison shows ColonSegNet as the best option.

4.3.5 NanoNet

NanoNet [92] is an encoder-decoder architecture. We have presented the block diagram of NanoNet in Figure 4.15. While designing NanoNet, we aimed to design a lightweight network architecture. The proposed network architecture uses MobileNetV2 [170], a pre-trained ImageNet model as the encoder. There is a modified residual block in between the encoder (MobileNetV2) and decoder. There are three decoder blocks in the architecture.

At first, the input images are fed into the pre-trained MobileNetV2 encoder. The main reason for using MobileNetV2 as a pre-trained encoder is that it is a lightweight model. It has shown high accuracy despite of utilizing very few parameters. The pre-trained encoder (MobileNetV2) begin with standard convolution having a number of 32 feature channels. This is followed by the bottleneck layer and ReLU6 activation function. In

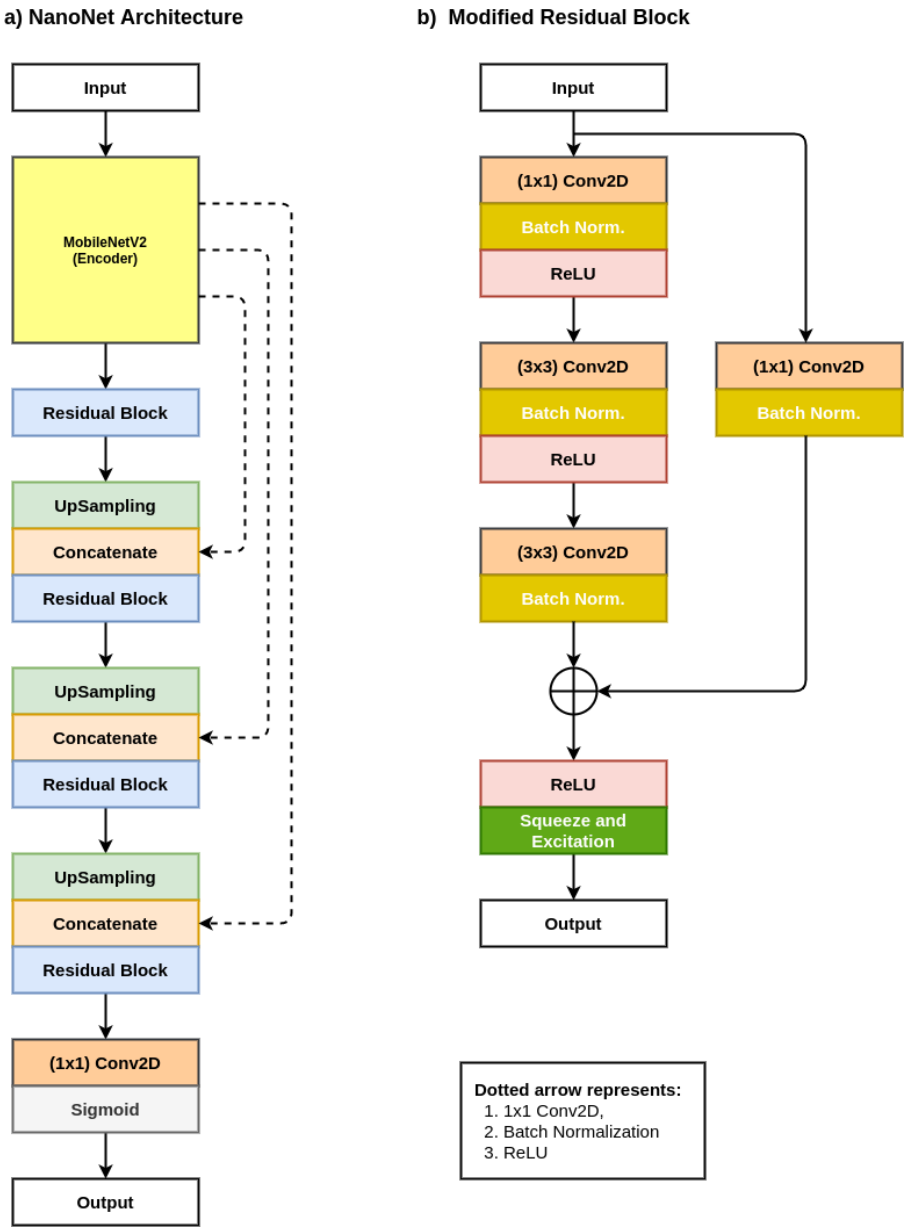


Figure 4.15: Block diagram of (a) NanoNet architecture, and (b) Modified Residual block [92]

the entire encoder network, the feature maps are gradually downsampled by utilizing the strided convolution [92].

The output from the encoder block is passed through the modified residual block, which is passed to the decoder block. The block diagram of the modified residual block can be observed in Figure 4.15. In each decoder block, bilinear upsampling is used for increasing the spatial dimension of the input feature maps. Next, the concatenation is done between the appropriate feature maps from the pre-trained MoibleNetV2 encoder to the decoders using skip connections. The final concatenated features are passed through the modified residual block as shown in the Figure. It was done to improve the gener-

Table 4.8: Qualitative results comparison of NanoNet with recent baseline methods on KvasirCapsule-SEG [92]

Method	Parameters	DSC	mIoU	Recall	Precision	F2	Accuracy	FPS
ResUNet (GRSL'18) [223]	8,227,393	**0.9532**	**0.9137**	**0.9785**	**0.9325**	**0.9677**	**0.9386**	17.96
ResUNet++ (ISM'19)[94]	4,070,385	0.9499	0.9087	0.9762	0.9296	0.9648	0.9334	15.39
NanoNet-A (Ours)	235,425	0.9493	0.9059	0.9693	**0.9325**	0.9609	0.9351	28.35
NanoNet-B (Ours)	132,049	0.9474	0.9028	0.9682	0.9308	0.9593	0.9324	27.39
NanoNet-C (Ours)	36,561	0.9465	0.9021	0.9754	0.9238	0.9629	0.9297	**29.48**

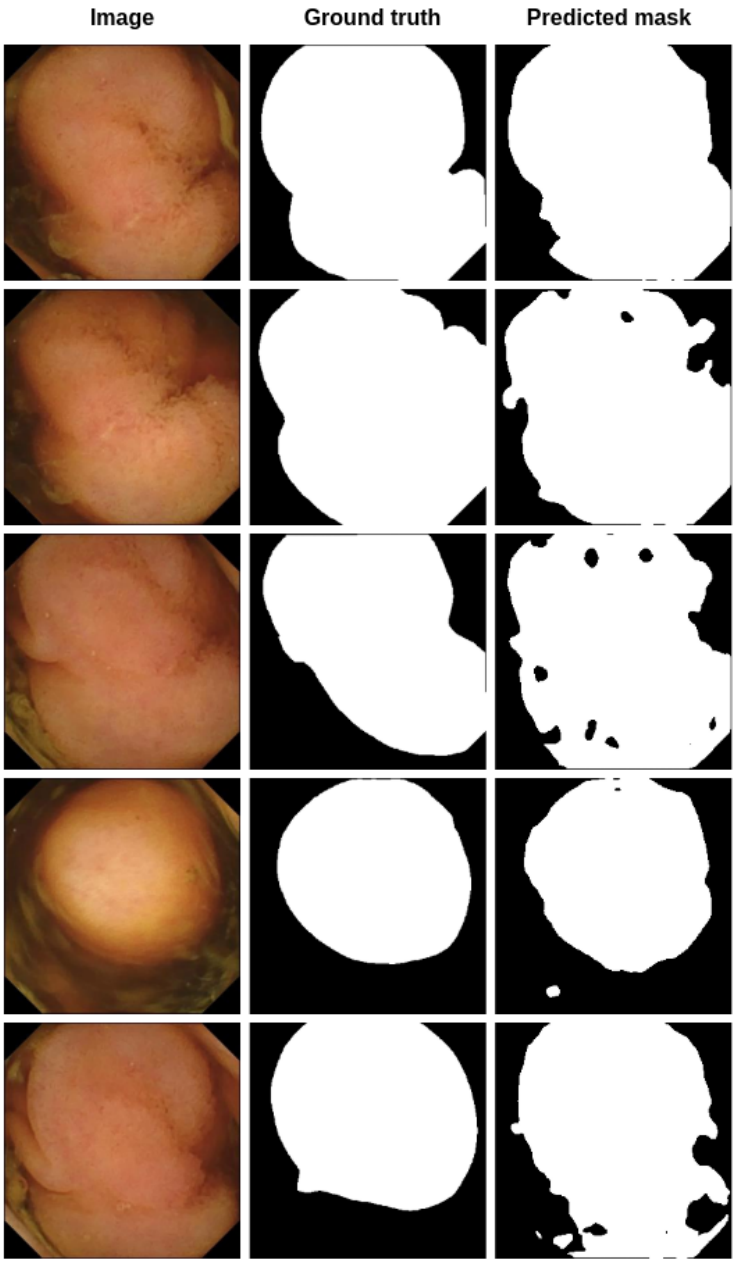


Figure 4.16: Qualitative results analysis of NanoNet-A on KvasirCapsule-SEG dataset

alizability capability of the decoder block. Finally, the output is passed through 1×1 convolution and sigmoid activation function for the binary class segmentation [92].

We have designed three NanoNets, namely NanoNet-A, NanoNet-B, and NanoNet-

C. The only difference in the three architectures is the number of feature channels. In NanoNet-A, 32, 64, and 128 feature channels are used. We have used feature channels of 32, 64, and 96 for NanoNet-B. In NanoNet-C, the feature channels are again reduced to 16, 24, and 32. As the feature channels are reduced, the number of parameters in the network is also reduced, evidenced by 235,425 trainable parameters in NanoNet-A, 132,049 trainable parameters in NanoNet-B, and 36,561 trainable parameters in NanoNet-C [92].

Table 4.8 shows the qualitative results comparison of different version of NanoNet with ResUNet [223] and ResUNet++ [94]. A result comparison shows that our lightweight model, NanoNet-C achieves a DSC of 0.9465 and mIoU of 0.9021, which is very competitive to the DSC and mIoU of ResUNet, that has a maximum DSC of 0.9532 and mIoU of 0.9137. However, when we take the processing speed into consideration and compare, NanoNet-C almost produces a near real-time speed of 29.48 FPS, which is nearly twice of the ResUNet. Similarly, Figure 4.16 shows the qualitative results of NanoNet-A on KvasirCapsule-SEG dataset. From the qualitative results, we can observe that for some WCE frames, our proposed method NanoNet-A, produces a good segmentation mask. However, for some of the images, it shows over-segmentation. It is to be noted that KvasirCapsule-SEG, has only 55 samples. Therefore, in the future, we plan to collect and annotate more WCE datasets and test our algorithm to make any generalizable remarks. Additionally, we have also trained and tested our architectures with three other different datasets. A detailed explanation about the architecture, datasets, implementation details, and results can be found in our paper [92].

4.4 Other segmentation architectures

We have also designed other medical image segmentation architectures such as Dual Decoder Attention Network (DDANet) [195], Progressively Normalized Self-Attention Network (PNS-Net) [95], and UNet-ResNet50 [3]. The works such as Feedback Attention Network (FANet) [196], Multi-scale Residual Fusion Network (MSRFNet) [181], and Metalearning under few shot setting [103] are under review.

The architecture design of all of these segmentation models can be found in their respective papers. DDANet [195] was especially designed for "2020 Endotect Challenge". The model produces a DSC of 0.7874 and FPS of 70.23 on the Endotect challenge segmentation dataset. PNS-Net [95] achieves a DSC of 0.8400, mIoU of 0.7450 with CVC-300-TV dataset. UNet-ResNet50 [3] achieves a DSC of 0.8154 and mIoU of 0.7396. We have highlighted some of the results in Table 4.9. We have tested our architectures on more than

Table 4.9: Results of the proposed segmentation algorithms on experimented datasets

Reference	Year	Dataset	DSC	mIoU	Rec	Prec	Acc	F2	FPS
ResUNet++ [94]	2019	Kvasir-SEG[89]	0.8133	0.7927	0.7064	0.8774	-	-	-
DoubleUNet++ [86]	2020	CVC-ClinicDB [21] Medico automatic polyp	0.9239	0.8611	0.8457	0.9592	-	-	-
UNet-ResNet50 [3]	2020	segmentation challenge	0.8154	0.7396	0.8533	0.8532	0.9506	0.8272	-
ResUNet++ + CRF [85]	2021	ASU-Mayo [187]	0.8850	0.8635	0.6504	0.4858	-	-	-
ColonSegNet [93]	2021	Kvasir-SEG [89]	0.8206	0.7239	0.8496	0.8435	0.9493	0.8206	182.38
NanoNet [92]	2021	KvasirCapsule-SEG [92]	0.9532	0.9137	0.9785	0.9325	0.9386	0.9677	17.96
DDANet [195]	2021	Endotect 2020 [71]	0.7874	0.7010	0.7987	0.8577	-	-	70.23
PNS-Net [95]	2021	CVC-ColonDB [18]	0.8400	0.7450	-	-	-	-	140.00

Table 4.10: Codes for our segmentation architectures

Method	Code
ResUNet++ [94]	`https://github.com/DebeshJha/ResUNetPlusPlus`
ResUNet++ + CRF + TTA [85]	`https://github.com/DebeshJha/ResUNetPlusPlus-with-CRF-and-TTA`
DoubleUNet [86]	`https://github.com/DebeshJha/2020-CBMS-DoubleU-Net`
ColonSegNet [93]	`https://github.com/DebeshJha/ColonSegNet`
NanoNet [92]	`https://github.com/DebeshJha/NanoNet`
DDANet[195]	`https://github.com/nikhilroxtomar/DDANet`
PNS-Net [95]	`https://github.com/GewelsJI/PNS-Net`

one dataset but only highlighted some of the results here. The source code of the proposed architectures can be found in Table 4.10.

4.5 Surgical instrument segmentation

Intra-operative tracking of the surgical instrument is vital in computer and robotic-assisted interventions [168]. Computer and robotic-assisted systems have the potential to improve the clinical workflow in laparoscopy. Tracking surgical instruments is difficult because of challenging conditions (for example, specularity, blood, smoke, reflections, and motion artifacts). There are existing automated methods for surgical instrument detection and segmentation. However, significant limitations lie in algorithms robustness and generalizability. Robustness is the ability of the automated methods to perform consistently even on challenging images. Generalizability is the ability of the algorithms trained on one specific dataset from one hospital should generalize across new datasets from other interventions.

We participated in the Robust Medical Instrument Segmentation (ROBUST-MIS) challenge and provided our solutions for binary instrument segmentation. Our solution was inspired by Refined Attention Segmentation Network (RASNet) [142]. RASNet is an encoder-decoder architecture. The main motivation behind choosing the RasNet based architecture was its ability to capture both higher-level and lower-level features.

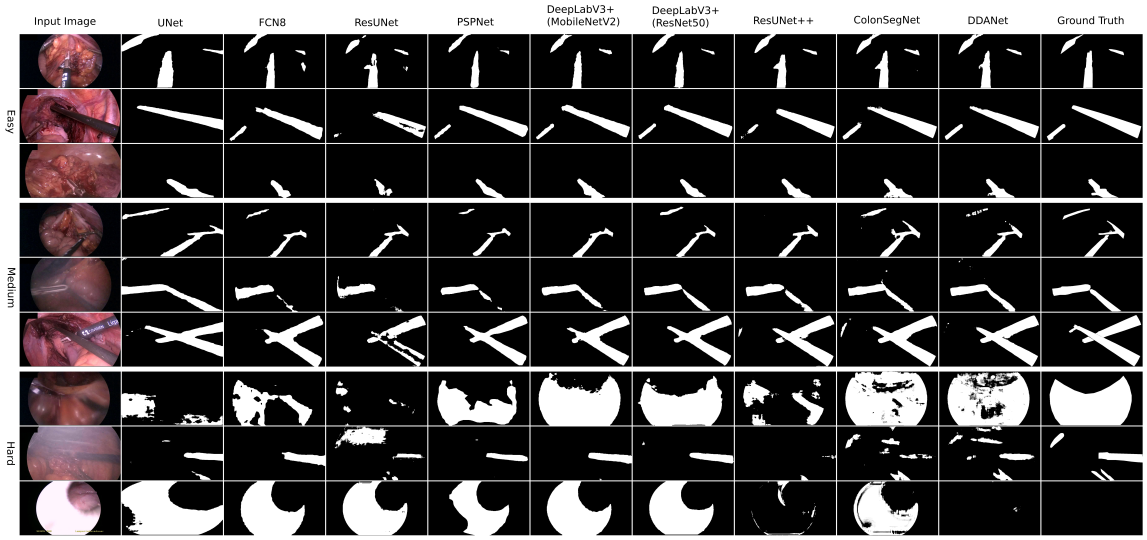


Figure 4.17: Qualitative results of nine DL algorithm on the ROBUST-MIS datasets [87]

Our special focus was on data augmentation and hyperparameter tuning. RASNet uses ResNet50 [65] as an encoder, and the decoder consists of an attention fusion module and decoder block [142]. ResNet50 is pre-trained on ImageNet[39], and therefore, it can capture deep semantic features efficiently. The attention fusion module fuses the low-level and high-level features efficiently and significantly improve segmentation accuracy by selecting precise position information. Our team was ranked 8th in the generalization ranking and 7th in the robustness ranking for the binary segmentation task. Both of them considered DSC as the evaluation metrics. A detailed score of our team and the challenge can be found in our paper [168].

In our last work [87], we have further evaluated and compared some of the recent DL architectures to improve the performance on binary surgical instrument segmentation tasks. For this we have compared algorithms such as UNet [167], FCN8 [123], ResUNet [223], PSPNet [224], DeepLabv3+ [34], ResUNet++[94], ColonSegNet [93], and DDANet [195] with different backbones. Figure 4.17 shows the qualitative results of nine segmentation methods on easy, medium, and hard cases. The qualitative and quantitative results showed that the architectures such as ColonSegNet, and DDANet performed reasonably well with different types of instruments. ColonSegNet produces DSC of 0.8495, mIoU of 0.7943, and FPS of 185.54. Similarly, DDANet produces DSC of 0.8739, mIoU of 0.8183 and FPS of 101.36. The result showed DDANet as the most efficient method among nine methods with a real-time speed of 101.36 FPS. A detailed description of experimental setup and configuration, strengths and weaknesses of the methods can be found in our paper [87].

4.6 Challenges and competitions

We have introduced three challenges and competitions, namely, "2020 Medico Automatic Polyp Segmentation Challenge [91]", "Endotect challenge [71]", and "EndoCV2021 challenge [5]". The Medico and Endotect challenges used the HyperKvasir [26] and Kvasir-SEG [89] datasets. For the EndoCV2021 challenge, we have created and used the PolypGen [5] dataset. Common datasets and benchmarking of the methods are essential for developing automated methods for clinical translation. The goal of the challenge was to benchmark several proposed classification and semantic segmentation DL algorithms on the same dataset. Here, we briefly describe each challenge.

4.6.1 Medico automatic polyp segmentation challenge

The "Medico automatic polyp segmentation challenge[1]" is an international benchmarking challenge hosted through Mediaeval[2] platform. The "Medico automatic polyp segmentation challenge" aimed to benchmark automated polyp segmentation algorithms on the same dataset. The challenge also aims to develop a method that can detect challenging polyps (for example, flat polyps, sessile polyps and small or diminutive polyps). In this challenge, we invite multimedia and computer vision researchers to submit the results on two tasks, namely, (i) polyp segmentation task and (ii) algorithm efficiency tasks. In the first task, the participants were asked to design and submit the best automated methods to automatically segment polyps. In the second task, participants were asked to submit the results taking speed into account. The evaluation metrics for the first task and second task was mIoU and FPS respectively. A detailed description of the challenge, tasks, and evaluation metrics can be found in our paper [91].

Table 4.11 shows the results of the "polyp segmentation task (task i: required)" of the 17 participating teams in the challenge. Table 4.12 shows the results of the "algorithm efficiency task (task ii: optional)" from the 9 participating teams in the challenge. Team "PRML2020GU" obtained the highest mIoU of 0.7897 and highest DSC of 0.8607. Similarly, team "GeorgeBatch"obtained the highest processing speed of 196.89 FPS. However, we consider only the teams with a minimum of mIoU of 0.70 and FPS of 30. Therefore, the algorithm proposed by team "HCMUS" was considered as the best solution for the algorithm efficiency task.

[1] https://multimediaeval.github.io/editions/2020/tasks/medico/
[2] http://www.multimediaeval.org/

Table 4.11: Polyp segmentation task

Team Name	Jaccard	DSC	Recall	Precision	Accuracy	F2
PRML2020GU	**0.7897**	**0.8607**	0.9031	0.8673	0.9546	0.8748
HBKU_UNITN_SIMULA	0.7773	0.8476	0.8503	0.8897	**0.9630**	0.8448
AI-TCE	0.7770	0.8503	**0.9164**	0.8389	0.9566	**0.8790**
HGV-HCMUS	0.7659	0.8405	0.8943	0.8445	0.9465	0.8576
IIAI-Med	0.7619	0.8385	0.8304	**0.9012**	0.9602	0.8283
SBS	0.7550	0.8316	0.8316	0.8851	0.9582	0.8249
ML-MMIV Saruar	0.7516	0.8228	0.8390	0.8822	0.9564	0.8249
AI-JMU	0.7374	0.8143	0.8266	0.8743	0.9463	0.8103
MedSeg_JU	0.7133	0.8019	0.8354	0.8286	0.9446	0.8124
VT	0.7057	0.7926	0.8830	0.7878	0.9331	0.8236
NKT	0.6847	0.7801	0.8077	0.8126	0.9404	0.7854
UNITRK	0.6437	0.7287	0.7098	0.8572	0.9432	0.7131
GeorgeBatch	0.6351	0.7327	0.7500	0.8229	0.9422	0.7361
AMI Lab	0.6195	0.7088	0.7286	0.7914	0.9325	0.7122
IRIS-NSYSU	0.5035	0.6417	0.8791	0.5849	0.8726	0.7508
UiO-Zero	0.4381	0.5618	0.6972	0.5558	0.8806	0.6110
FAST-NU-DS	0.1834	0.2669	0.2744	0.2918	0.8272	0.2676

Table 4.12: Algorithm efficiency task

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

4.6.2 Endotect 2020 challenge

Endotect 2020 challenge[3] is an international challenge that aimed to develop automated methods to detect GI tract abnormalities and findings that could potentially aid gastroenterologists in the clinic. In this challenge, we offered three tasks (i) GI findings detection, (ii) efficient detection, and (iii) polyp segmentation. HyperKvasir [26] was provided as the development dataset to the participants. New datasets were provided as the test dataset. Five teams submitted results for the detection tasks. Two teams submitted results for efficient detection tasks, and five teams submitted the results for the segmentation tasks.

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

Team "howard [67]" provided the winning solution for detection and efficient detection tasks, and team "aggcmab [49]" provided the winning solution with DSC of 0.920. A detailed description of the task, each teams method, and results can be found in our paper [71].

4.6.3 EndoCV2021 challenge

EndoCV2021[4] is an international competition aimed at benchmarking and development of new computer vision and DL algorithms on polyp detection and polyp segmentation tasks. The main aim of the challenge is to address the generalizability issue in polyp segmentation and detection. For this, we have created a new dataset called polypGen [5]. PolypGen is a comprehensive dataset that consists of data from 6 different centres, comprising more than 300 patients. The dataset includes both still frames and sequence frames with more than 3,446 annotated polyps and corresponding ground truth labels and bounding box labels. Our dataset was curated by a team of three computer scientists and six expert gastroenterologists. In addition to the polyps frames in the sequence dataset, we also released 2,520 negative samples that are normal frames captured during the endoscopic examination and that does not consist of any polyp samples. Therefore, we released 6,282 frames, including both still frames and sequence frames, through the challenge. More information about the dataset can be found in our paper [5]. Thambawita et al. [191] provided the winning solution for the segmentation challenge, and Gan et al. [50] provided a winning solution for the polyp detection task.

4.7 Summary

In this chapter, we have briefly presented our designed algorithms for automated GI tract findings classification, polyp detection, polyp segmentation, and instrument segmentation. We have shown the example results obtained by few methods for each objective. We have also pointed out our other valuable works that have been done as part of our research. The codes of the proposed segmentation architectures have been made publicly available and are highlighted. Additionally, we have presented the scores of the proposed architectures. Some of the results achieved from our core research (i.e., polyp segmentation) show that the proposed methods can be used in real procedures. Moreover, we have also briefly explained the challenges and competitions organized as part of the main research. We

[4] `https://endocv2021.grand-challenge.org/`

have presented the results of the "Medico automatic polyp segmentation challenge" in the thesis. The results of Endotect 2020 is presented in our paper [71]. The results of the "EndoCV2021" is found on the challenge webpage.

In the next chapter, we will discuss objective-wise contributions and how it helps to achieve the study's main goal. Additionally, we will also discuss the other contributions that were not part of the main research but contributed to other areas of multimedia research. Moreover, we also highlight a few commercial systems and devices that were recently developed and can be integrated into hospitals to aid gastroenterologists. Furthermore, we also highlight the challenges and potential limitations in the field.

Chapter 5

Discussion

The main research aim of the thesis is to design automated methods for GI abnormalities classification, detection, and segmentation. We have proposed different architectures and datasets to achieve the main goal. In this chapter, we will discuss the contribution based on each objective, and finally, we show how each objective full fill our main research goal.

5.1 Summary and contributions

5.1.1 Objective I

The aim of objective I is to collaborate with the different medical centres and gastroenterologists, collect the datasets, and make them publicly available to the research community. Once a dataset is collected, we sort the still images and videos from the whole dataset. Then, we check the duplication, anonymize the dataset (if needed), label it, and store it in a particular format.

We maintain the heterogeneity during the dataset creation and adapt specific protocol while selecting the dataset. The images that belong to a particular class (for us, colorectal polyp in most cases) are double-checked with the medical experts. Once the images are confirmed, we upload the images to the annotation toolbox and label the region of interest (ROI) (for our case area covered by polyp or polyps). All the annotations are exported, and the ground truth masks are generated. The generated masks will be useful for medical image segmentation tasks. Additionally, bounding boxes around the ROI are formed. The bounding boxes information are useful for the detection tasks.

Regarding collecting medical datasets, we have made sure that privacy and ethical concerns have been addressed before we released the dataset. So far, we have publicly

released Kvasir-SEG [89], HyperKvasir [26], KvasirCapsule [180], PolypGen [5], Kvasir-Instrument [88], and KvasirCapsule-SEG [92]. Additionally, we have also released datasets through the "Endotect challenge [71]" and "Medico automatic polyp segmentation challenge [91]". However, in these cases, new test images were released. Similarly, we released Kvasir-Sessile [85], which is a sub-set of the Kvasir-SEG dataset, that consists of flat or sessile polyps. Furthermore, we benchmark the algorithms on the datasets to initiate research and invite computer vision and medical multimedia researchers to develop new methods and test on the publicly available datasets.

5.1.2 Objective II

Objective II aims to explore, investigate, and design ML methods for GI findings classification and design of the automated polyp detection model. In this respect, we approach the multi-class GI tract classification problem based on global image features and ML classifiers. We obtain the best MCC of 0.8353 [193]. We further extended this work by evaluating the performance of our ML models to the specific classification problem, with and without retraining. We have evaluated the performance using different standard computer vision metrics, including ROC and precision recall curve (PRC). This study has emphasized the importance of cross-data test and performance metrics interpretation rather than choosing single performance metrics and one dataset [190].

In another study [85], we have performed a detailed analysis of automated ML classification methods with the endoscopic imaging dataset. This paper summarizes the methods presented at Medico and BioMedia competitions from 2017 to 2019. Our analysis revealed that participants showed improved performance over consecutive years both in accuracy and computational speed. Our clinical relevance ranking results showed that the team that achieved the highest accuracy had a lower rank than that team with decent accuracy. This was due to both speed and accuracy used for the computation of the clinical ranking. Additionally, we proposed LightLayers [90], a method to reduce the number of trainable parameters in the DNN and tested it across non-medical datasets and showed promising results.

For the automated polyp detection, we have trained algorithms such as EfficientDet, YOLOv4, Faster RCNN, RetinaNet, and YOLO3 with different backbones. We have compared these algorithms with our ColonSegNet. The comparison demonstrated the performance of the ColonSegNet on the Kvasir-SEG dataset and established a strong benchmark. Basically, it is a segmentation architecture trained end-to-end. In our case,

we have converted the predicted masks into bounding boxes and calculated the speed and other performance metrics.

5.1.3 Objective III

In objective III, we aim to design new medical image segmentation architectures that can segment medical images with high accuracy and high processing speed. Our research is mainly focused on colorectal polyp segmentation and surgical instrument segmentation. As the supporting cases, we have investigated and showed that our architectures could be extended to the general medical image segmentation tasks. For the automated polyp segmentation, we have designed some architectures considering higher DSC, whereas other architectures are designed considering real-time processing speed and competitive DSC. The architectures such as ResUNet++ [94], "ResUNet++ + CRF + TTA [85]", DoubleUNet [86], Ensemble MultiResUNet [197], and UNet-ResNet50 [3] focus on the scores such as higher DSC and higher mIoU. The architectures such as NanoNet [92], DDANet [195], ColonSegNet [93], and PNS-Net [95] are mostly focused on processing speed by design. Detailed explanations and results about all the architectures can be found in their respective papers. The codes of all of the segmentation architectures are available here: `https://github.com/DebeshJha`.

Towards addressing the generalizability issue, we have done an extensive cross-data test on publicly available datasets [85]. Moreover, we have organized "Polyp Detection & Segmentation: Addressing generalisability" to address the generalizability issue in polyp segmentation and detection[1]. For this challenge, we have also released multi-centre datasets toward addressing the need for generalizability in polyp segmentation and detection [5]. Additionally, we have organized the competitions, namely, "Medico automatic polyp segmentation challenge[2]" where we ask participants to develop efficient algorithms for automatic polyp segmentation focused on highest accuracy and speed. Moreover, we have also organized the Endotect 2020 challenge, where we ask participants to automatically classify GI findings and segment polyps. Through various challenges and competitions, our contributions are providing a platform, releasing new datasets, evaluating and reproducing different proposed methodologies on the same train and test dataset. Crowdsourcing based solution provides an opportunity to learn jointly and utilize the cross-domain knowledge to solve the same problem, which leads to the development of better algorithms.

[1] `https://endocv2021.grand-challenge.org/`
[2] `https://multimediaeval.github.io/editions/2020/tasks/medico/`

We have evaluated and compared popular semantic segmentation architectures for the surgical instrument segmentation [87]. The comparison showed that our Dual decoder attention network (DDANet) architecture outperform other SOTA methods used in the comparison both in terms of DSC and speed. Moreover, we participated in the ROBUST-MIS challenge[3], where we proposed a solution based on Refined Attention Segmentation Network (RASNet) [142]. Our solution was ranked 8th for the binary instrument segmentation task in the challenge. An overview of the challenge, dissection of each team's results, and the result summary on binary and multi-instance instrument segmentation tasks can be found in our paper [168].

5.1.4 Contribution to the main goal

The main goal of our research is to develop classification, detection, and segmentation algorithms for automated examination of GI tract findings. We aim to achieve this through the three objectives mentioned above. The significant problem in the field of GI endoscopy is the lack of publicly available datasets for research. Towards achieving this, we have collected, labeled, and annotated multiple datasets with the help from expert gastroenterologists, and publicly released new GI endoscopy datasets such as HyperKvasir [26], KvasirCapsule [180], PolypGen [5], Kvasir-SEG [89], Kvasir-Instrument [88], Medico automatic polyp segmentation challenge [91], EndoTect challenge [71], and KvasirCapsule-SEG [92] *(objective I)*. These datasets have been acquired in close collaboration with Norwegian hospitals, except PolypGen [5], which comes from the collaboration with hospitals from France, Italy, United Kingdom, Egypt, and Norway. The senior gastroenterologists identified the problems in detecting lesions in the hospitals and provided feedback on the datasets and results. Additionally, we have provided benchmarks on these datasets to encourage other researchers to use our datasets and develop novel and reproducible methods on the publicly available datasets. Therefore, we achieve our first objective (*objective I*) through several GI endoscopy dataset collection public releases.

Next, we designed automated methods for multi-class GI tract findings classification using GFs and ML techniques (*objective II*). We demonstrated that our best method achieved MCC of 0.8353 on medico 2018 challenge dataset [193]. We performed cross-dataset bias study on four GI endoscopy datasets and five ML techniques in the context of GI findings and abnormalities classification. Our experimental results suggested that a multi-center or a cross-dataset evaluation is important for a realistic understanding of the

[3] `https://robustmis2019.grand-challenge.org/`

performance of the ML models in the real-world setting [190]. Moreover, we performed a comprehensive study, where we evaluated, compared, dissected, ranked, and summarized 23 automated classification methods presented in the different GI endoscopy competitions [84]. We analyzed from ML methods using global image features to recent CNN based approaches using transfer learning and specialized data augmentation. Our study showed significant results improvement for GI tract finding classification, efficiency, and automatic reporting tasks over three consecutive years. We advocate organizing more competitions and analyzing the clinical applicability of the developed methods based on their merits such as higher accuracy, higher speed, robustness and transparency. Furthermore, we organized the 2020 Endotect challenge, where we proposed classification tasks, evaluation metrics, datasets and evaluated the participant's results. In GI endoscopy, crowdsourcing is a popular technique to solve complex problems. In our competition, Team Howard [67] achieved a MCC score of 0.9030 and processing speed of 129.74 FPS. In the context of automated polyp detection, we have developed a method, ColonSegNet [93], that can detect polyps at 180 FPS and produces AP of 0.8000 and mIoU of 0.8100. We achieve our second objective (*objective II*) through the presented methods and studies.

We have designed several architectures for automated polyp segmentation [94, 86, 92, 195, 93, 85, 95, 3, 197] (*objective III*). One of example contributions of our algorithm is the combination of "ResUNet++, CRF, and TTA [85]", where the proposed method achieved DSC of ≥ 0.8500 with three still images and achieved ≥ 0.8800 with two video datasets. The proposed architectures can identify and segment flat and sessile polyps with high accuracy, which is one of the significant contributions of semantic segmentation architecture. In the polyp segmentation tasks, usually processing speed was often neglected. We have built several architectures considering speed [92, 195, 95, 93]. One of our polyp segmentation architectures designed considering processing speed is ColonSegNet [93], that achieved real-time processing speed of 182.38 FPS, with decent DSC and mIoU scores. The results suggests that ColonSegNet could be useful to the clinician during the live examination. Similarly, we have developed NanoNet [92], which is a lightweight architecture with a smaller model size and low computational cost. NanoNet has only 36,561 parameters. The model could be integrated into low-end endoscope hardware devices. The architecture produces a real-time processing speed of 30 FPS with five datasets and an acceptable DSC. Similarly, for the surgical instrument segmentation in laparoscopy, we have researched and proposed two methods [168, 87]. DDANet achieved DSC of 0.8739 and mIoU of 0.8183 and real-time processing speed of 101.36 FPS with

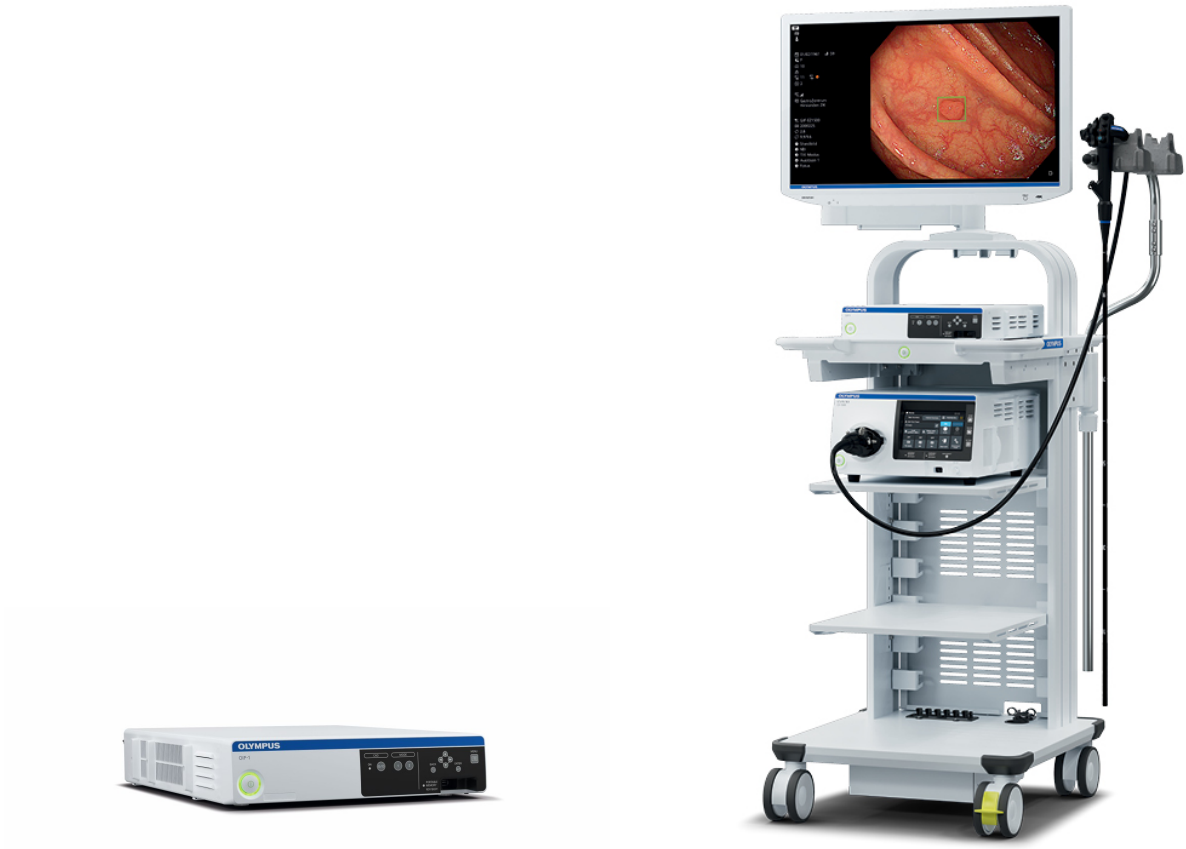
the ROBUST-MIS challenge dataset. We achieve our third objective (*objective III*) by designing several CNN based architectures for automated polyp and surgical instrument segmentation.

We achieve our main goal by connecting three different objectives (*objective I, objective II, and objective III*) to develop classification, detection and segmentation methods for CADx and CADe for the automated examination of anomalies in the GI tract. We have highlighted some of the results that show that our algorithm has the potential to identify, detect, and segment potential lesions with high accuracy and high speed. We have shown that the designed algorithm performs well with other medical imaging datasets as well. However, we conjecture that it will also perform well with the non-medical imaging datasets. Moreover, our lightweight models have the potential to be integrated with the endoscopic device and could also be useful to other mobile applications.

5.2 Possible limitations

The objective wise contribution shows that our method achieved promising results for each of the objectives. However, there are also limitations associated with our study. Our research is based on a retrospective study. Prospective DL studies and randomized trials are less subjected to bias. However, we have the limitation of resources to conduct prospective studies. Although we proposed several datasets during our study, our models are still trained on the limited datasets. Better solutions could be achieved with datasets having a large and diverse samples.

Despite CNN based approaches provide a better performance, there are also certain challenges associated with it. Interpretation of the CNN based approaches are difficult. It is difficult to decide how much data is required or how many layers are needed to achieve desirable performance. Catastrophic forgetting is another challenge associated with the CNN [128]. Additionally, CNN are bad at encoding representation [169]. In our work, we have resized images during training of the DL models for reducing the complexity of the network. However, it could also potentially lose relevant information, which can influence the network performance. Furthermore, we have optimized the code. However, further optimization of code might exist.



(a) ENDO-AID (b) EVIS X1 trolley

Figure 5.1: Olympus' endoscopy system (olympus-europa.com)

5.3 Commercial systems

Since the end of 2019, there are some AI based commercial products designed to support gastroenterologists in the clinic especially developed for detection and treatment of colorectal polyps. The CADe based application can highlight suspicious lesions such as adenomas, polyps, and malignant neoplasm during the colonoscopy examination and prevent from CRC. However, this is the initial stage, and their systems have not been tested on the cohort populations, and there are still chances of improving the technology. There are devices such as DISCOVERY™[4] from Pentax Medical, GI Genius™[5] intelligent endoscopy module from Medtronic, CAD EYE module[6] from FujiFilm, and ENDO-AID[7] from Olympus.

[4] https://www.pentaxmedical.com/pentax/en/95/2/DISCOVERY-new
[5] https://www.medtronic.com/us-en/c/digestive-gastrointestinal/gi-genius.html
[6] https://www.onis.com/en/cad-eye-artificial-intelligence-fujifilm-eluxeo
[7] https://mdc.olympus.eu/asset/084438885177/00b6f92967815b6dde91959494a5fc2e

5.4 Summary

In this chapter, we discussed our contributions based on each objective and showed how we fulfilled our objectives. We showed that how it helped us to achieve the main goal. We have presented some of the example results to demonstrate the effectiveness of our methods. We also discuss the possible limitation of the study and highlight the need for more public datasets for research. Additionally, we also present some of the currently available commercial systems. However, we also point out that the current systems are at the initial stage, and the current systems miss a large number of polyps in colonoscopy examination. Therefore, there is still a huge potential for algorithmic improvement for deployment across several hospitals for saving the life of the patients.

In summary, we believe that our contributions will have a societal impact in the field of upper endoscopy and colonoscopy. During our research, we have been able to expand our collaboration with several medical experts and hospitals from where we can collect, develop, and release more datasets that could be useful to the researchers working in the GI endoscopy and medical image analysis. We have developed algorithms and established collaboration with computer scientists working to solve the same challenge. We plan to compare our methods' performance with the team of gastroenterologists and observe the difference in the results. These results can help in better understanding the problem and help us in designing better frameworks.

Chapter 6

Conclusion and future work

We have proposed various methods and datasets for the development of new technology in the field of GI endoscopy. The scope of the research was to build novel automated classification, detection, and segmentation methods in the field of GI endoscopy that could assist gastroenterologists in the standard colonoscopy or endoscopy procedure. In this chapter, we reveal the main conclusions of our work. Additionally, we highlight possible limitations and areas of improvement. Moreover, we also point out some of the future research directions.

6.1 Conclusion

We have developed high performing architectures for CADx and CADe systems based on DL techniques for automated classification, detection, and segmentation of GI tract abnormalities and findings. The retrospective study showed that our models could identify the potential presence of abnormalities and lesions (for example, polyps in the colon, flat or depressed polyp, adenomas) with high performance. We have shown that our proposed architectures can reduce the miss rate and improve the detection rate even for flat or sessile polyps. We have shown that our architectures can automatically classify GI abnormalities and findings with high accuracy and segment colon polyps and surgical instruments in real-time with high processing speed.

Moreover, we have collected, labelled, annotated, and released several GI endoscopy and colonoscopy datasets, including datasets from multi-centre that can be freely downloaded under open source license for academic, research and industrial purpose (prior consent required). To the best of our knowledge, HyperKvasir [26] and Kvasir-Capsule [180] are the world largest and diverse publicly available GI tract and VCE datasets. Simi-

larly, PolypGen [5] is also only the publicly available multi-center dataset available for academic research and innovation. We invite multimedia and medical image analysis community researchers through the open-access datasets to provide their solutions through the challenge to tackle generalizability. Moreover, we have explored and analyzed the generalizability of the proposed methods using the cross-dataset test towards designing generalizable polyp segmentation models. We have highlighted the lack of generalizability issues of the best performing model on the completely new (independent) datasets through the comprehensive study. We have made the source code and train-test split of the datasets available for most proposed methods. It helps in method reproducibility and result comparison with the other recent SOTA methods.

The results have shown that our architectures can improve clinical outcomes and help endoscopists as our methods can identify multiple polyps simultaneously, including lesions such as flat polyps or sessile polyps that are overlooked by endoscopists in real-time. The architectures presented in the study can make the endoscopy efficient, easy, accessible, and reduce the miss-rate and overall load of endoscopists, nurses, and hospitals. We conjecture that our architectures can be helpful to detect missed lesions regardless of endoscopists experience and current attentiveness. Additionally, the designed models for CADe detection and segmentation have the potential to minimize the eye movement of the endoscopists, which would make the endoscopy procedure easier. As the endoscopy procedure is time-consuming, CNN models have large potential to provide convenient support during an examination.

6.2 Future work

Although we have presented promising results for automatic GI tract disease classification, colorectal polyp segmentation and detection, and comprehensive dataset collection, there is still potential for improvement in each area. We have developed algorithms that perform reasonably well on the available datasets. However, computational scientists can improve the metrics such as accuracy and speed. To show the full potential of the developed algorithms, we still need high-quality and diverse datasets to explore the strength and weaknesses of our methods and their usability in clinical settings. Currently, we have only labelled a few thousand images out of a hundred thousand's images in our released datasets. The released datasets have a high potential to build robust and generalizable algorithms for building CADe system to reach clinical goals. In the future, we plan to label it further with the collaboration of a team of expert gastroenterologists. Additionally, we

plan to collect more multi-centre and diverse datasets to solve the challenges related to the lack of heterogeneous datasets in the field of colonoscopy. The open-access datasets provide an opportunity for benchmarking and development of better generalizable DL models for automated colorectal polyp segmentation.

Our work is mainly focused on colorectal polyps, where we have achieved good results. However, our model should be investigated for other bowel conditions as well. Multi-centre datasets and randomized trials with information from thousands of patients are essential to evaluate better if our methods are clinically significant. Additionally, in our papers, we have mainly shown both best performing cases and failure cases. In the future, we plan to develop new methods to improve the results of the failure cases. Future research should focus more on improving the understanding and interpretability of the results of the CNN models. Additionally, statistical analysis of the methods is required to account for the differences between the baseline and the proposed methods. A potentially optimized combination of robust, generalizable, reproducible, and interpretable CNN models could be useful to build clinical relevant systems.

We have released a video polyp sequence dataset with both positive samples and negative samples. In the future, researchers could build automated methods on such datasets. Most of our segmentation models are specifically focused on the binary segmentation problem due to the lack of relevant datasets. Future research should be focused on the multi-class segmentation problem. Most of our research is based on developing encoder-decoder and transfer learning-based architectures. Future research should explore the use of different modalities, such as textual information in the form of patient information, and medical history along with imaging data. Similarly, transformer-based architectures should be explored in future. Although, we have summarized the results through the competition on automatic report generation tasks, future research should focus more on generating standardized endoscopy reports in GI endoscopy. This will reduce the overall administrative burden of the clinicians. It will also help to prepare a standard report to interpret and describe the finding consistently. Researchers could explore learning paradigms such as multi-task learning, meta-learning, domain adaptation, online learning, and continual learning in the future. Finally, we aim to test our methods on the hardware for clinical reliability.

Bibliography

[1] Sang Bong Ahn et al. "The miss rate for colorectal adenoma determined by quality-adjusted, back-to-back colonoscopies". In: *Gut and liver* 6.1 (2012), p. 64.

[2] Mojtaba Akbari et al. "Polyp segmentation in colonoscopy images using fully convolutional network". In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2018, pp. 69–72.

[3] Saruar Alam et al. "Automatic Polyp Segmentation using U-Net-ResNet50". In: *Proceedings of CEUR Multimedia Benchmark Workshop (MediaEval)* (2020).

[4] Sharib Ali et al. "An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy". In: *Scientific reports* 10.1 (2020), pp. 1–15.

[5] Sharib Ali et al. "PolypGen: A multi-center polyp detection and segmentation dataset for generalisability assessment". In: *arXiv preprint arXiv:2106.04463* (2021).

[6] Sharib Ali *et al.* "Endoscopy disease detection challenge 2020". In: *arXiv preprint arXiv:2003.03376* (2020). URL: `https://arxiv.org/abs/2003.03376`.

[7] Max Allan et al. "2017 robotic instrument segmentation challenge". In: *arXiv preprint arXiv:1902.06426* (2019).

[8] Max Allan et al. "2018 robotic scene segmentation challenge". In: *arXiv preprint arXiv:2001.11190* (2020).

[9] Victor de Almeida Thomaz, Cesar A Sierra-Franco, and Alberto B Raposo. "Training data enhancements for robust polyp segmentation in colonoscopy images". In: *Proceedings of IEEE International Symposium on Computer-Based Medical Systems (CBMS)*. 2019, pp. 192–197.

[10] Iakovos Amygdalos. "Detection and classification of gastrointestinal cancer and other pathologies through quantitative analysis of optical coherence tomography data and goniophotometry". PhD thesis. Imperial College London, 2014.

[11] Quentin Angermann et al. “Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis”. In: *Proceedings of Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures (CARE CLIP)*. Vol. 10550. 2017, pp. 29–41.

[12] Diego Ardila et al. “End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography”. In: *Nature medicine* 25.6 (2019), pp. 954–961.

[13] Harry R Aslanian et al. “Nurse observation during colonoscopy increases polyp detection: a randomized prospective study”. In: *American Journal of Gastroenterology* 108.2 (2013), pp. 166–172.

[14] Debapriya Banik, Debotosh Bhattacharjee, and Mita Nasipuri. “A multi-scale patch-based deep learning system for polyp segmentation”. In: *Advanced Computing and Systems for Security*. 2020, pp. 109–119.

[15] David W Bates et al. “The potential of artificial intelligence to improve patient safety: a scoping review”. In: *NPJ digital medicine* 4.1 (2021), pp. 1–8.

[16] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Meskó. “The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database”. In: *NPJ digital medicine* 3.1 (2020), pp. 1–8.

[17] Jorge Bernal and Histace Aymeric. *Gastrointestinal Image ANAlysis (GIANA) Angiodysplasia D&L challenge*. `https://endovissub2017-giana.grand-challenge.org/home/`. Accessed: 2017-11-20. 2017.

[18] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. “Towards automatic polyp detection with a polyp appearance model”. In: *Pattern Recognition* 45.9 (2012), pp. 3166–3182.

[19] Jorge Bernal et al. “Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 endoscopic vision challenge”. In: *IEEE transactions on medical imaging* 36.6 (2017), pp. 1231–1249.

[20] Jorge Bernal et al. “Polyp detection benchmark in colonoscopy videos using gtcreator: A novel fully configurable tool for easy and fast annotation of image databases”. In: *Proceedings of Computer Assisted Radiology and Surgery (CARS)*. 2018.

[21] Jorge Bernal et al. “WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians”. In: *Computerized Medical Imaging and Graphics* 43 (2015), pp. 99–111.

[22] Jorge Bernal del Nozal. “Polyp localization and segmentation in colonoscopy images by means of a model of appearance for polyps”. In: *ELCVIA: electronic letters on computer vision and image analysis* 13.2 (2013).

[23] Kirsten Bibbins-Domingo et al. “Screening for colorectal cancer: US Preventive Services Task Force recommendation statement”. In: *The Journal of the American Medical Association (JAMA)* 315.23 (2016), pp. 2564–2575.

[24] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. “Yolov4: Optimal speed and accuracy of object detection”. In: *arXiv preprint arXiv:2004.10934* (2020).

[25] Sebastian Bodenstedt et al. “Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery”. In: *arXiv preprint arXiv:1805.02475* (2018).

[26] Hanna Borgli et al. “HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy”. In: *Scientific Data* 7.1 (2020), pp. 1–14.

[27] Patrick Brandao et al. “Towards a computed-aided diagnosis system in colonoscopy: automatic polyp segmentation using convolution neural networks”. In: *Journal of Medical Robotics Research* 3.02 (2018), p. 1840002.

[28] Matthias Breier, Sebastian Gross, and Alexander Behrens. “Chan-Vese-segmentation of polyps in colonoscopic image data”. In: *Proceedings of the International Student Conference on Electrical Engineering POSTER*. Vol. 2011. 2011.

[29] Titus J Brinker et al. “Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task”. In: *European Journal of Cancer* 113 (2019), pp. 47–54.

[30] Anna M Buchner et al. “Trainee participation is associated with increased small adenoma detection”. In: *Gastrointestinal endoscopy* 73.6 (2011), pp. 1223–1231.

[31] Michael F Byrne et al. “Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model”. In: *Gut* 68.1 (2019), pp. 94–100.

[32] Suzanne K Chambers et al. “A five-year prospective study of quality of life after colorectal cancer”. In: *Quality of Life Research* 21.9 (2012), pp. 1551–1564.

[33] Yuan Chang et al. “Gastrointestinal Tract Diseases Detection with Deep Attention Neural Network”. In: *Proceedings of the ACM International Conference on MultiMedia (ACMMM)*. 2019, pp. 2568–2572.

[34] Liang-Chieh Chen et al. “Encoder-decoder with atrous separable convolution for semantic image segmentation”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 801–818.

[35] Liang-Chieh Chen et al. “Rethinking atrous convolution for semantic image segmentation”. In: *arXiv preprint arXiv:1706.05587* (2017).

[36] Peng-Jen Chen et al. “Accurate classification of diminutive colorectal polyps using computer-aided analysis”. In: *Gastroenterology* 154.3 (2018), pp. 568–575.

[37] Douglas A Corley et al. “Adenoma detection rate and risk of colorectal cancer and death”. In: *New england journal of medicine* 370.14 (2014), pp. 1298–1306.

[38] Jeffrey De Fauw et al. “Clinically applicable deep learning for diagnosis and referral in retinal disease”. In: *Nature medicine* 24.9 (2018), pp. 1342–1350.

[39] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)*. 2009, pp. 248–255.

[40] Peter J. Denning et al. “Computing as a discipline”. In: *Computer* 22.2 (1989), pp. 63–70.

[41] Willem Dijkstra et al. “Towards a Single Solution for Polyp Detection, Localization and Segmentation in Colonoscopy Images”. In: *VISIGRAPP (4: VISAPP)*. 2019, pp. 616–625.

[42] Gordana Dodig-Crnkovic. “Scientific methods in computer science”. In: *Proceedings of the Conference for the Promotion of Research in IT at New Universities and at University Colleges in Sweden, Skövde, Suecia*. 2002, pp. 126–130.

[43] *El Salvador Atlas of Gastrointestinal Video Endoscopy*. `http://www.gastrointestinalatlas.com/index.html`. Accessed: 2019-12-16.

[44] Andre Esteva et al. “Dermatologist-level classification of skin cancer with deep neural networks”. In: *nature* 542.7639 (2017), pp. 115–118.

[45] Deng-Ping Fan et al. “Pranet: Parallel reverse attention network for polyp segmentation”. In: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2020, pp. 263–273.

[46] Yuqi Fang et al. “Selective feature aggregation network with area-boundary constraints for polyp segmentation”. In: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2019, pp. 302–310.

[47] Deborah A Fisher et al. “Complications of colonoscopy”. In: *Gastrointestinal endoscopy* 74.4 (2011), pp. 745–752.

[48] US Preventive Services Task Force. “Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement”. In: *The Journal of the American Medical Association (JAMA)* 325.19 (2021), pp. 1965–1977.

[49] Adrian Galdran, Gustavo Carneiro, and Miguel A González Ballester. “A hierarchical multi-task approach to gastrointestinal image analysis”. In: *Proceedings of Pattern Recognition International Workshops and Challenges (ICPR)*. 2021, pp. 275–282.

[50] Tianyuan Gana et al. “Detection Of Polyps During Colonoscopy Procedure Using YOLOv5 Network”. In: *Proceedings of the 3rd International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV 2021)*. 2021.

[51] Melanie Ganz, Xiaoyun Yang, and Greg Slabaugh. “Automatic segmentation of polyps in colonoscopic narrow-band imaging data”. In: *IEEE Transactions on Biomedical Engineering* 59.8 (2012), pp. 2144–2151.

[52] Enrique Garcia-Ceja et al. “HTAD: A Home-Tasks Activities Dataset with Wrist-accelerometer and Audio Features”. In: *Proceedings of the International Conference on Multimedia Modeling (MMM)*. 2021, pp. 196–205.

[53] *Gastrointestinal Lesions in Regular Colonoscopy Dataset*. `http://www.depeca.uah.es/colonoscopy_dataset/`. Accessed: 2019-12-12.

[54] *GASTROLAB - the Gastrointestinal Site*. `http://www.gastrolab.net/index.htm`. Accessed: 2019-12-12.

[55] Raman Ghimirea, Sahadev Poudelb, and Sang-Woong Leec. “An Augmentation Strategy with Lightweight Network for Polyp Segmentation”. In: *Proceedings of the 3rd International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV 2021)*. 2021.

[56] Ross Girshick. “Fast r-cnn”. In: *Proceedings of the IEEE international conference on computer vision (ICCV)*. 2015, pp. 1440–1448.

[57] Ross Girshick et al. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 2014, pp. 580–587.

[58] European Colorectal Cancer Screening Guidelines Working Group et al. “European guidelines for quality assurance in colorectal cancer screening and diagnosis: overview and introduction to the full supplement publication”. In: *Endoscopy* 45.01 (2013), pp. 51–59.

[59] Varun Gulshan et al. “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs”. In: *The Journal of the American Medical Association (JAMA)* 316.22 (2016), pp. 2402–2410.

[60] Yun Bo Guo and Bogdan Matuszewski. “Giana polyp segmentation with fully convolutional dilation neural networks”. In: *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. 2019, pp. 632–641.

[61] Mark Hall et al. “The WEKA data mining software: an update”. In: *ACM SIGKDD explorations newsletter (SIGKDD Explor. Newsl.)* 11.1 (2009), pp. 10–18.

[62] Awni Y Hannun et al. “Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network”. In: *Nature medicine* 25.1 (2019), pp. 65–69.

[63] Gavin C Harewood and David A Lieberman. “Colonoscopy practice patterns since introduction of medicare coverage for average-risk screening”. In: *Clinical Gastroenterology and Hepatology* 2.1 (2004), pp. 72–77.

[64] Philipp Harzig, Moritz Einfalt, and Rainer Lienhart. “Automatic disease detection and report generation for gastrointestinal tract examination”. In: *Proceedings of the ACM International Conference on MultiMedia (ACMMM)*. 2019, pp. 2573–2577.

[65] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 2016, pp. 770–778.

[66] Kaiming He et al. “Mask r-cnn”. In: *Proceedings of the IEEE international conference on computer vision (ICCV)*. 2017, pp. 2961–2969.

[67] Qi He et al. “Hybrid Loss with Network Trimming for Disease Recognition in Gastrointestinal Endoscopy”. In: *Proceedings of the Pattern Recognition ICPR International Workshops and Challenges*. 2021.

[68] Steven Hicks et al. “ACM Multimedia BioMedia 2019 Grand Challenge Overview”. In: *Proceedings of the ACM International Conference on Multimedia (ACMMM)*. 2019, pp. 2563–2567.

[69] Steven A Hicks et al. “ACM Multimedia BioMedia 2020 Grand Challenge Overview”. In: *Proceedings of the ACM International Conference on Multimedia (ACMMM)*. 2020, pp. 4655–4658.

[70] Steven A Hicks et al. “Deep Learning Based Disease Detection Using Domain Specific Transfer Learning”. In: *Proceedings of CEUR Multimedia Benchmark Workshop (MediaEval)*. 2018.

[71] Steven Alexander Hicks et al. “The EndoTect 2020 Challenge: Evaluation and Comparison of Classification, Segmentation and Inference Time for Endoscopy.” In: *International Conference on Pattern Recognition Workshops and challenges*. 2020, pp. 263–274.

[72] Trung-Hieu Hoang, Hai-Dang Nguyen, and Thanh-An Nguyen. “An application of Residual Network and Faster - RCNN for Medico: Multimedia Task at MediaEval 2018”. In: *Proceedings of CEUR Multimedia Benchmark Workshop (MediaEval)*. 2018.

[73] Trung-Hieu Hoang et al. “Enhancing Endoscopic Image Classification with Symptom Localization and Data Augmentation”. In: *Proceedings of the ACM International Conference on MultiMedia (ACMMM)*. 2019, pp. 2578–2582.

[74] Øyvind Holme et al. “Effect of flexible sigmoidoscopy screening on colorectal cancer incidence and mortality: a randomized clinical trial”. In: *Journal of the American Medical Association (JAMA)* 312.6 (2014), pp. 606–615.

[75] Jie Hu, Li Shen, and Gang Sun. “Squeeze-and-excitation networks”. In: *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 7132–7141.

[76] Chien-Hsiang Huang, Hung-Yu Wu, and Youn-Long Lin. “Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps”. In: *arXiv preprint arXiv:2101.07172* (2021).

[77] Sae Hwang et al. “Polyp detection in colonoscopy video using elliptical shape feature”. In: *Proceedings of IEEE International conference on the Image Processing (ICIP)*. Vol. 2. 2007, pp. 465–468.

[78] Stephanie L Hyland et al. “Early prediction of circulatory failure in the intensive care unit using machine learning”. In: *Nature medicine* 26.3 (2020), pp. 364–373.

[79] Nabil Ibtehaz and M Sohel Rahman. “MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation”. In: *Neural Networks* 121 (2020), pp. 74–87.

[80] Gavriel Iddan et al. “Wireless capsule endoscopy”. In: *Nature* 405.6785 (2000), p. 417.

[81] NationalCancer Institute. *Cancer Stat Facts: Colorectal Cancer*. URL: `https://seer.cancer.gov/statfacts/html/colorect.html`.

[82] Fabian Isensee and Klaus H Maier-Hein. “OR-UNet: an Optimized Robust Residual U-Net for Instrument Segmentation in Endoscopic Images”. In: *arXiv preprint arXiv:2004.12668* (2020).

[83] L Jansen et al. “Quality of life among long-term ($\geq$ 5 years) colorectal cancer survivors–systematic review”. In: *European Journal of Cancer* 46.16 (2010), pp. 2879–2888.

[84] Debesh Jha et al. “A comprehensive analysis of classification methods in gastrointestinal endoscopy imaging”. In: *Medical image analysis* 70 (2021), p. 102007.

[85] Debesh Jha et al. “A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation”. In: *IEEE journal of biomedical and health informatics* 25.6 (2021), pp. 2029–2040.

[86] Debesh Jha et al. “Doubleu-net: A deep convolutional neural network for medical image segmentation”. In: *Proceedings of the IEEE International Symposium on Computer-Based Medical Systems (CBMS)*. 2020, pp. 558–564.

[87] Debesh Jha et al. “Exploring Deep Learning Methods for Real-Time Surgical Instrument Segmentation in Laparoscopy”. In: *Proceedings of the IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. 2021.

[88] Debesh Jha et al. “Kvasir-Instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy”. In: *Proceedings of International Conference on Multimedia Modeling (MMM)*. 2021, pp. 218–229.

[89] Debesh Jha et al. “Kvasir-seg: A segmented polyp dataset”. In: *Proceedings of International Conference on Multimedia Modeling (MMM)*. 2020, pp. 451–462.

[90] Debesh Jha et al. “LightLayers: Parameter Efficient Dense and Convolutional Layers for Image Classification”. In: *Proceedings of Joint conference of Parallel and Distributed Computing, Applications and Technologies and International Symposium on Parallel Architectures, Algorithms and Programming (PDCAT-PAAP)*. 2021, pp. 285–296.

[91] Debesh Jha et al. “Medico Multimedia Task at MediaEval 2020: Automatic Polyp Segmentation”. In: *Proceedings of CEUR Multimedia Benchmark Workshop (MediaEval)*. 2020.

[92] Debesh Jha et al. “NanoNet: Real-Time Polyp Segmentation in Video Capsule Endoscopy and Colonoscopy”. In: *Proceedings of International Symposium on Computer-Based Medical Systems (CBMS)*. 2021.

[93] Debesh Jha et al. “Real-time polyp detection, localization and segmentation in colonoscopy using deep learning”. In: *Ieee Access* 9 (2021), pp. 40496–40510.

[94] Debesh Jha et al. “Resunet++: An advanced architecture for medical image segmentation”. In: *Proceedings of the IEEE International Symposium on Multimedia (ISM)*. 2019, pp. 225–2255.

[95] Ge-Peng Ji et al. “Progressively Normalized Self-Attention Network for Video Polyp Segmentation”. In: *Proceedings of 24th international conference on medical image computing & computer assisted intervention (MICCAI 2021)*. 2021.

[96] C Daniel Johnson et al. “Accuracy of CT colonography for detection of large adenomas and cancers”. In: *New England Journal of Medicine* 359.12 (2008), pp. 1207–1217.

[97] Michal F Kaminski et al. “Quality indicators for colonoscopy and the risk of interval cancer”. In: *New England Journal of Medicine* 362.19 (2010), pp. 1795–1803.

[98] Jaeyong Kang and Jeonghwan Gwak. “Ensemble of instance segmentation models for polyp segmentation in colonoscopy images”. In: *IEEE Access* 7 (2019), pp. 26440–26447.

[99] Stavros A Karkanis et al. “Computer-aided tumor detection in endoscopic video using color wavelet features”. In: *IEEE transactions on information technology in biomedicine* 7.3 (2003), pp. 141–152.

[100] Sara Hosseinzadeh Kassani et al. “Automatic Polyp Segmentation Using Convolutional Neural Networks”. In: *arXiv preprint arXiv:2004.10792* (2020).

[101] Anastasia Katsoula et al. “Diagnostic accuracy of fecal immunochemical test in patients at increased risk for colorectal cancer: a meta-analysis”. In: *JAMA internal medicine* 177.8 (2017), pp. 1110–1118.

[102] Daniel S Kermany et al. “Identifying medical diagnoses and treatable diseases by image-based deep learning”. In: *Cell* 172.5 (2018), pp. 1122–1131.

[103] Rabindra Khadga et al. “Few-shot segmentation of medical images based on meta-learning with implicit gradients”. In: *arXiv preprint arXiv:2106.03223* (2021).

[104] Philipp Kickingereder et al. “Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study”. In: *The Lancet Oncology* 20.5 (2019), pp. 728–740.

[105] Nam Hee Kim et al. “Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies”. In: *Intestinal research* 15.3 (2017), p. 411.

[106] Anastasios Koulaouzidis et al. “KID Project: an internet-based digital video atlas of capsule endoscopy for research purposes”. In: *Endoscopy international open* 5.6 (2017), E477–E483.

[107] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems (NIPS)* 25 (2012), pp. 1097–1105.

[108] Chang Kyun Lee et al. “Participation by experienced endoscopy nurses increases the detection rate of colon polyps during a screening colonoscopy: a multicenter, prospective, randomized study”. In: *Gastrointestinal endoscopy* 74.5 (2011), pp. 1094–1102.

[109] Eung-Joo Lee et al. “Segmentation of surgical instruments in laparoscopic videos: training dataset generation and deep-learning-based framework”. In: *Proceedings of the Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling*. Vol. 10951. 2019, 109511T.

[110] Ji Young Lee et al. “Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets”. In: *Scientific reports* 10.1 (2020), pp. 1–9.

[111] AM Leufkens et al. “Factors influencing the miss rate of polyps in a back-to-back colonoscopy study”. In: *Endoscopy* 44.05 (2012), pp. 470–475.

[112] Baopu Li and Max Q-H Meng. “Tumor recognition in wireless capsule endoscopy images using textural features and SVM-based feature selection”. In: *IEEE Transactions on Information Technology in Biomedicine* 16.3 (2012), pp. 323–329.

[113] Jing Nan Li and Si Yi Yuan. “Fecal occult blood test in colorectal cancer screening”. In: *Journal of digestive diseases* 20.2 (2019), pp. 62–64.

[114] Qiaoliang Li et al. “Colorectal polyp segmentation using a fully convolutional neural network”. In: *Proceedings of the international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI)*. 2017, pp. 1–5.

[115] David A Lieberman. “Screening for colorectal cancer”. In: *New England Journal of Medicine* 361.12 (2009), pp. 1179–1187.

[116] Jennifer S Lin et al. “Screening for colorectal cancer: updated evidence report and systematic review for the US Preventive Services Task Force”. In: *The Journal of the American Medical Association (JAMA)* 315.23 (2016), pp. 2576–2594.

[117] Jennifer S Lin et al. “Screening for colorectal cancer: updated evidence report and systematic review for the US Preventive Services Task Force”. In: *The Journal of the American Medical Association (JAMA)* 325.19 (2021), pp. 1978–1997.

[118] Tsung-Yi Lin et al. “Focal loss for dense object detection”. In: *Proceedings of the IEEE international conference on computer vision (ICCV)*. 2017, pp. 2980–2988.

[119] Ming Liu, Jue Jiang, and Zenan Wang. “Colonic polyp detection in endoscopic videos with single shot detection based deep convolutional neural network”. In: *IEEE Access* 7 (2019), pp. 75058–75066.

[120] Wei Liu et al. “Ssd: Single shot multibox detector”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2016, pp. 21–37.

[121] Xiaoxuan Liu et al. “A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis”. In: *The lancet digital health* 1.6 (2019), e271–e297.

[122] Yun Liu et al. “Artificial intelligence–based breast cancer nodal metastasis detection: insights into the black box for pathologists”. In: *Archives of pathology & laboratory medicine* 143.7 (2019), pp. 859–868.

[123] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 2015, pp. 3431–3440.

[124] Zhipeng Luo et al. "Adaptive Ensemble: Solution to the Biomedia ACM MM GrandChallenge 2019". In: *Proceedings of the ACM International Conference on Multimedia (ACMMM)*. 2019, pp. 2583–2587.

[125] Mathias Lux and Savvas A Chatzichristofis. "Lire: lucene image retrieval: an extensible java cbir library". In: *Proceedings of the ACM international conference on Multimedia (ACMMM)*. 2008, pp. 1085–1088.

[126] Nadim Mahmud et al. "Computer vision and augmented reality in gastrointestinal endoscopy". In: *Gastroenterology report* 3.3 (2015), pp. 179–184.

[127] Alexander V Mamonov et al. "Automated polyp detection in colon capsule endoscopy". In: *IEEE transactions on medical imaging* 33.7 (2014), pp. 1488–1502.

[128] Gary Marcus. "Deep learning: A critical appraisal". In: *arXiv preprint arXiv:1801.00631* (2018).

[129] Wenhua Meng et al. "Biomedia ACM MM Grand Challenge 2019: Using Data Enhancement to Solve Sample Unbalance". In: *Proc. ACM Int. Conf. Multim.* 2019, pp. 2588–2592.

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

[131] Masashi Misawa et al. "Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video)". In: *Gastrointestinal Endoscopy* 93.4 (2021), pp. 960–967.

[132] Xi Mo et al. "An efficient approach for polyps detection in endoscopic videos based on faster R-CNN". In: *Proceedings of the international conference on pattern recognition (ICPR)*. 2018, pp. 3929–3934.

[133] Ahmed Mohammed et al. "Y-net: A deep convolutional neural network for polyp detection". In: *arXiv preprint arXiv:1806.01907* (2018).

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

[135] Yuichi Mori et al. “Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy: a prospective study”. In: *Annals of internal medicine* 169.6 (2018), pp. 357–366.

[136] Myura Nagendran et al. “Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies”. In: *Bmj* 368 (2020).

[137] Ju Gang Nam et al. “Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs”. In: *Radiology* 290.1 (2019), pp. 218–228.

[138] Kimmie Ng, Folasade P. May, and Deborah Schrag. “US Preventive Services Task Force Recommendations for Colorectal Cancer Screening”. In: *Journal of American Medical Association* 325.19 (2021), pp. 1943–1945.

[139] Ngoc-Quang Nguyen, Duc My Vo, and Sang-Woong Lee. “Contour-Aware Polyp Segmentation in Colonoscopy Images Using Detailed Upsamling Encoder-Decoder Networks”. In: *IEEE Access* 8 (2020), pp. 99495–99508.

[140] Quang Nguyen and Sang-Woong Lee. “Colorectal segmentation using multiple encoder-decoder network in colonoscopy images”. In: *Proceedings of the first international conference on artificial intelligence and knowledge engineering (AIKE)*. 2018, pp. 208–211.

[141] Zhen-Liang Ni et al. “Pyramid attention aggregation network for semantic segmentation of surgical instruments”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020, pp. 11782–11790.

[142] Zhen-Liang Ni et al. “RASNet: segmentation for tracking surgical instruments in surgical videos using refined attention segmentation network”. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2019, pp. 5735–5738.

[143] Olympus. *The ENDOCAPSULE 10 System. Olympus homepage*, `https://www.olympus-europa.com/medical/en/Products-and-Solutions/Products/Product/ENDOCAPSULE-10-System.html`. 2013.

[144] Daniil Pakhomov et al. “Deep residual learning for instrument segmentation in robotic surgery”. In: *Proceedings of the International Workshop on Machine Learning in Medical Imaging (MLMI)*. 2019, pp. 566–573.

[145] Sun Young Park et al. “A colon video analysis framework for polyp detection”. In: *IEEE Transactions on Biomedical Engineering* 59.5 (2012), pp. 1408–1418.

[146] Stefan Petscharnig, Klaus Schöffmann, and Mathias Lux. “An Inception-like CNN Architecture for GI Disease and Anatomical Landmark Classification”. In: *Proceedings of CEUR Multimedia Benchmark Workshop (MediaEval)*. 2017.

[147] Konstantin Pogorelov. “DeepEIR: A Holistic Medical Multimedia System for Gastrointestinal Tract Disease Detection and Localization”. In: *PhD thesis, University of Oslo* (2019).

[148] Konstantin Pogorelov et al. “Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection”. In: *Proceedings of the ACM on Multimedia Systems Conference (ACMMM)*. 2017, pp. 164–169.

[149] Konstantin Pogorelov et al. “Medico multimedia task at mediaeval 2018”. In: *Proceedings of CEUR Multimedia Benchmark Workshop (MediaEval)*. Vol. 2283. 2018, pp. 1–4.

[150] Konstantin Pogorelov et al. “Nerthus: A bowel preparation quality video dataset”. In: *Proceedings of the ACM on Multimedia Systems Conference (MMSys)*. 2017, pp. 170–174.

[151] JM Poomeshwaran et al. “Polyp Segmentation using Generative Adversarial Network”. In: *Proceedings of International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*. 2019, pp. 7201–7204.

[152] Carmen CY Poon et al. “AI-doscopist: a real-time deep-learning-based algorithm for localising polyps in colonoscopy videos with edge computing devices”. In: *NPJ Digital Medicine* 3.1 (2020), pp. 1–8.

[153] JM Poorneshwaran et al. “Polyp segmentation using generative adversarial network”. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2019, pp. 7201–7204.

[154] Hemin Ali Qadir et al. “Polyp detection and segmentation using mask R-CNN: Does a deeper feature extractor cnn always perform better?” In: *Proceedings of the International Symposium on Medical Information and Communication Technology (ISMICT)*. 2019, pp. 1–6.

[155] Hemin Ali Qadir Qadir. “Development of Image Processing Algorithms for the AutomaticScreening of Colon Cancer”. In: *PhD thesis, University of Oslo* (2020).

[156] Joseph Redmon and Ali Farhadi. "YOLO9000: better, faster, stronger". In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 2017, pp. 7263–7271.

[157] Joseph Redmon and Ali Farhadi. "Yolov3: An incremental improvement". In: *arXiv preprint arXiv:1804.02767* (2018).

[158] Jaroslaw Regula et al. "Colonoscopy in colorectal-cancer screening for detection of advanced neoplasia". In: *New England Journal of Medicine* 355.18 (2006), pp. 1863–1872.

[159] Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *arXiv preprint arXiv:1506.01497* (2015).

[160] Félix Renard et al. "Variability and reproducibility in deep learning for medical image segmentation". In: *Scientific Reports* 10.1 (2020), pp. 1–16.

[161] Alessandro Repici et al. "Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial". In: *Gastroenterology* 159.2 (2020), pp. 512–520.

[162] DOUGLAS K Rex et al. "Colonoscopic miss rates of adenomas determined by back-to-back colonoscopies". In: *Gastroenterology* 112.1 (1997), pp. 24–28.

[163] Michael Riegler. "Eir-a medical multimedia system for efficient computer aided diagnosis". In: *PhD thesis, University of Oslo* (2017).

[164] Michael Riegler et al. "Multimedia for medicine: the medico task at Mediaeval 2017". In: *Proceedings of CEUR Multimedia Benchmark Workshop (MediaEval)*. 2017.

[165] Michael Riegler et al. "From annotation to computer-aided diagnosis: Detailed evaluation of a medical multimedia system". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 13.3 (2017), p. 26.

[166] Michael Riegler et al. "Multimedia and medicine: Teammates for better disease detection and survival". In: *Proceedings of the ACM Multimedia (ACMMM)*. 2016, pp. 968–977.

[167] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2015, pp. 234–241.

[168] Tobias Roß et al. “Comparative validation of multi-instance instrument segmentation in endoscopy: results of the ROBUST-MIS 2019 challenge”. In: *Medical image analysis* 70 (2021), p. 101920.

[169] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. “Dynamic routing between capsules”. In: *arXiv preprint arXiv:1710.09829* (2017).

[170] Mark Sandler et al. “Mobilenetv2: Inverted residuals and linear bottlenecks”. In: *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)*. 2018, pp. 4510–4520.

[171] Aasma Shaukat et al. “Longer withdrawal time is associated with a reduced incidence of interval cancer after screening colonoscopy”. In: *Gastroenterology* 149.4 (2015), pp. 952–957.

[172] Younghak Shin and Ilangko Balasingham. “Comparison of hand-craft feature based SVM and CNN based deep learning framework for automatic polyp classification”. In: *Proceedings of IEEE the Engineering in Medicine and Biology Society (EMBC)*. 2017, pp. 3277–3280.

[173] Younghak Shin et al. “Automatic colon polyp detection using region based deep cnn and post learning approaches”. In: *IEEE Access* 6 (2018), pp. 40950–40962.

[174] Alexey A Shvets et al. “Automatic instrument segmentation in robot-assisted surgery using deep learning”. In: *Proceedings of the IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2018, pp. 624–628.

[175] Rebecca L Siegel et al. “Colorectal cancer statistics, 2017”. In: *CA: a cancer journal for clinicians* 67.3 (2017), pp. 177–193.

[176] Rebecca L Siegel et al. “Colorectal cancer statistics, 2020”. In: *CA: a cancer journal for clinicians* 70.3 (2020), pp. 145–164.

[177] Juan Silva et al. “Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer”. In: *International Journal of Computer Assisted Radiology and Surgery* 9.2 (2014), pp. 283–293.

[178] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[179] Anita Slomski. “Fecal Immunochemical Testing vs Sigmoidoscopy for Cancer Screening”. In: *Journal of the American Medical Association (JAMA)* 325.4 (2021), pp. 334–334.

[180] Pia H Smedsrud et al. “Kvasir-Capsule, a video capsule endoscopy dataset”. In: *Scientific Data* (2020).

[181] Abhishek Srivastava et al. “MSRF-Net: A Multi-Scale Residual Fusion Network for Biomedical Image Segmentation”. In: *IEEE Journal of Biomedical and Health Informatics* (2021).

[182] Xinzi Sun et al. “Colorectal polyp segmentation by u-net with dilation convolution”. In: *Proceedings of IEEE International Conference On Machine Learning And Applications (ICMLA)*. 2019, pp. 851–858.

[183] Hyuna Sung et al. “Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries”. In: *CA: a cancer journal for clinicians* (2021).

[184] Hyuna Sung et al. “Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries”. In: *CA: a cancer journal for clinicians* 71.3 (2021), pp. 209–249.

[185] Christian Szegedy et al. “Rethinking the inception architecture for computer vision”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 2016, pp. 2818–2826.

[186] Nima Tajbakhsh. “Ensuring High-Quality Colonoscopy by Reducing Polyp Miss-Rates.” PhD thesis. Arizona State University, 2015.

[187] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. “Automated polyp detection in colonoscopy videos using shape and context information”. In: *IEEE Transaction of Medical Imaging* 35.2 (2015), pp. 630–644.

[188] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. “Automatic polyp detection using global geometric constraints and local intensity variation patterns”. In: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2014, pp. 179–187.

[189] Mingxing Tan, Ruoming Pang, and Quoc V Le. “Efficientdet: Scalable and efficient object detection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 10781–10790.

[190] Vajira Thambawita et al. “An extensive study on cross-dataset bias and evaluation metrics interpretation for machine learning applied to gastrointestinal tract abnormality classification”. In: *ACM Transactions on Computing for Healthcare* 1.3 (2020), pp. 1–29.

[191] Vajira Thambawita et al. “DivergentNets: Medical Image Segmentation by Network Ensemble”. In: *Proceedings of the 3rd International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV 2021)*. 2021.

[192] Vajira Thambawita et al. “Pmdata: a sports logging dataset”. In: *Proceedings of the ACM Multimedia Systems Conference*. 2020, pp. 231–236.

[193] Vajira Thambawita et al. “The Medico-Task 2018: Disease Detection in the Gastrointestinal Tract using Global Features and Deep Learning”. In: *Proceedings of CEUR Multimedia Benchmark Workshop (MediaEval)*. 2018.

[194] *The Atlas of Gastrointestinal Endoscope*. `http://www.endoatlas.com/atlas_1.html`. Accessed: 2019-12-12.

[195] Nikhil Kumar Tomar et al. “DDANet: Dual Decoder Attention Network for Automatic Polyp Segmentation”. In: *Proceedings of the International Conference on Pattern Recognition (ICCP)*. 2021, pp. 307–314.

[196] Nikhil Kumar Tomar et al. “FANet: A Feedback Attention Network for Improved Biomedical Image Segmentation”. In: *arXiv preprint arXiv:2103.17235* (2021).

[197] Nikhil Kumar Tomar et al. “Improving Generalizability in Polyp Segmentation using Ensemble Convolutional Neural Network”. In: *Proceedings of the 3rd International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV 2021)*. 2021.

[198] Janet M. Torpy, Cassio Lynm, and Richard M. Glass. “Stomach Cancer”. In: *The Journal of the American Medical Association (JAMA)* 303.17 (2010), pp. 1771–1771.

[199] Cees Van Wijk et al. “Detection and segmentation of colonic polyps on implicit isosurfaces by second principal curvature flow”. In: *IEEE Transactions on Medical Imaging* 29.3 (2010), pp. 688–698.

[200] Ashish Vaswani et al. “Attention is all you need”. In: *Proceedings of the Neural Information Processing Systems (NIPS)*. 2017, pp. 5998–6008.

[201] David Vázquez et al. “A benchmark for endoluminal scene segmentation of colonoscopy images”. In: *Journal of healthcare engineering* 2017 (2017).

[202] Jasper LA Vleugels, Meta CJ Van Lanschot, and Evelien Dekker. “Colorectal cancer screening by colonoscopy: putting it into perspective”. In: *Digestive Endoscopy* 28.3 (2016), pp. 250–259.

[203] Brian Wahl et al. “Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings?” In: *BMJ global health* 3.4 (2018), e000798.

[204] Jingdong Wang et al. “Deep high-resolution representation learning for visual recognition”. In: *IEEE transactions on pattern analysis and machine intelligence* (2020).

[205] Pu Wang et al. “Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy”. In: *Nature biomedical engineering* 2.10 (2018), pp. 741–748.

[206] Pu Wang et al. “Lower adenoma miss rate of computer-aided detection-assisted colonoscopy vs routine white-light colonoscopy in a prospective tandem study”. In: *Gastroenterology* 159.4 (2020), pp. 1252–1261.

[207] Yi Wang et al. “Part-based multiderivative edge cross-sectional profiles for polyp detection in colonoscopy”. In: *IEEE Journal of Biomedical and Health Informatics* 18.4 (2013), pp. 1379–1389.

[208] Yi Wang et al. “Part-based multiderivative edge cross-sectional profiles for polyp detection in colonoscopy”. In: *IEEE Journal of Biomedical and Health Informatics* 18.4 (2014), pp. 1379–1389.

[209] Yi Wang et al. “Polyp-alert: Near real-time feedback during colonoscopy”. In: *International Journal of Computer methods and programs in biomedicine* 120.3 (2015), pp. 164–179.

[210] Joan L Warren et al. “Adverse events after outpatient colonoscopy in the Medicare population”. In: *Annals of internal medicine* 150.12 (2009), pp. 849–857.

[211] Milton C Weinstein et al. “Recommendations of the Panel on Cost-effectiveness in Health and Medicine”. In: *Journal of the American Medical Association (JAMA)* 276.15 (1996), pp. 1253–1258.

[212] *WEO Clinical Endoscopy Atlas*. `http://www.endoatlas.org/index.php`. Accessed: 2019-12-12.

[213] Itsara Wichakam et al. “Real-time polyps segmentation for colonoscopy video frames using compressed fully convolutional network”. In: *Proceedings of International Conference on Multimedia Modeling (MMM)*. 2018, pp. 393–404.

[214] Sidney J Winawer. “Screening sigmoidoscopy: can the road to colonoscopy be less traveled?” In: *Annals of internal medicine* 139.12 (2003), p. 1034.

[215] Lequan Yu et al. “Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos”. In: *IEEE journal of biomedical and health informatics* 21.1 (2017), pp. 65–75.

[216] Lingtao Yu et al. “A Holistically-Nested U-Net: Surgical Instrument Segmentation Based on Convolutional Neural Network”. In: *Journal of digital imaging* (2019), pp. 1–7.

[217] Yixuan Yuan, Dengwang Li, and Max Q-H Meng. “Automatic polyp detection via a novel unified bottom-up and top-down saliency approach”. In: *IEEE journal of biomedical and health informatics* 22.4 (2018), pp. 1250–1260.

[218] Lei Zhang, Sunil Dolwani, and Xujiong Ye. “Automated polyp segmentation in colonoscopy frames using fully convolutional neural network and textons”. In: *Proceedings of the Annual Conference on Medical Image Understanding and Analysis*. 2017, pp. 707–717.

[219] Luxia Zhang et al. “Big data and medical research in China”. In: *bmj* 360 (2018).

[220] Ruikai Zhang et al. “Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker”. In: *Pattern recognition* 83 (2018), pp. 209–219.

[221] Xu Zhang et al. “Real-time gastric polyp detection using convolutional neural networks”. In: *PloS one* 14.3 (2019), e0214133.

[222] Yundong Zhang, Huiye Liu, and Qiang Hu. “Transfuse: Fusing transformers and cnns for medical image segmentation”. In: *arXiv preprint arXiv:2102.08005* (2021).

[223] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. “Road extraction by deep residual u-net”. In: *IEEE Geoscience and Remote Sensing Letters* 15.5 (2018), pp. 749–753.

[224] Hengshuang Zhao et al. “Pyramid scene parsing network”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 2017, pp. 2881–2890.

[225] Mingda Zhou et al. “Polyp detection and radius measurement in small intestine using video capsule endoscopy”. In: *Proceedings of IEEE Biomedical Engineering and Informatics (BMEI)*. 2014, pp. 237–241.

[226] Zongwei Zhou et al. “Unet++: A nested u-net architecture for medical image segmentation”. In: *Proceedings of the Deep learning in medical image analysis and multimodal learning for clinical decision support*. 2018, pp. 3–11.

[227] Katharina Zimmermann-Fraedrich et al. “Right-sided location not associated with missed colorectal adenomas in an individual-level reanalysis of tandem colonoscopy studies”. In: *Gastroenterology* 157.3 (2019), pp. 660–671.

Appendix A

List of Papers

In this chapter, we list the published papers included in the Ph.D. research. We also describe the candidate's own contributions and show how each paper matches the given objectives defined in Section 1.x.

A.1 Paper I: ResUNet++: An Advance architecture for Medical image Segmentation

Authors: D. Jha, P.H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen

Abstract: Accurate computer-aided polyp detection and segmentation during colonoscopy examinations can help endoscopists resect abnormal tissue and thereby decrease chances of polyps growing into cancer. Towards developing a fully automated model for pixel-wise polyp segmentation, we propose ResUNet++, which is an improved ResUNet architecture for colonoscopic image segmentation. Our experimental evaluations show that the suggested architecture produces good segmentation results on publicly available datasets. Furthermore, ResUNet++ significantly outperforms U-Net and ResUNet, two key state-of-the-art deep learning architectures, by achieving high evaluation scores with a dice coefficient of 81.33%, and a mean Intersection over Union (mIoU) of 79.27% for the Kvasir-SEG dataset and a dice coefficient of 79.55%, and a mIoU of 79.62% with CVC-612 dataset.

Published: Proceedings of IEEE International Symposium on Multimedia (ISM), pp. 225-230, 2019.

Candidate contributions: D. Jha conceptualized this work and performed all the experiments in the paper. He prepared the manuscript and performed all the evaluation and analysis. Additionally, he made a subsequent revision to the manuscript with the input from all of the co-authors and presented it at the conference.

Thesis objectives: Objective III

ResUNet++: An Advanced Architecture for Medical Image Segmentation

Debesh Jha*‡, Pia H. Smedsrud*†§, Michael A. Riegler*§, Dag Johansen‡,
Thomas de Lange†§, Pål Halvorsen*¶, Håvard D. Johansen‡

*SimulaMet, Norway †Augere Medical AS, Norway
‡UiT The Arctic University of Norway, Norway §University of Oslo, Norway
¶Oslo Metropolitan University, Norway
Email: debesh@simula.no

***Abstract*—Accurate computer-aided polyp detection and segmentation during colonoscopy examinations can help endoscopists resect abnormal tissue and thereby decrease chances of polyps growing into cancer. Towards developing a fully automated model for pixel-wise polyp segmentation, we propose ResUNet++, which is an improved ResUNet architecture for colonoscopic image segmentation. Our experimental evaluations show that the suggested architecture produces good segmentation results on publicly available datasets. Furthermore, ResUNet++ significantly outperforms U-Net and ResUNet, two key state-of-the-art deep learning architectures, by achieving high evaluation scores with a dice coefficient of 81.33%, and a mean Intersection over Union (mIoU) of 79.27% for the Kvasir-SEG dataset and a dice coefficient of 79.55%, and a mIoU of 79.62% with CVC-612 dataset.**

***Index Terms*—Medical image analysis, semantic segmentation, colonoscopy, polyp segmentation, deep learning, health informatics.**

I. Introduction

Colorectal Cancer (CRC) is one of the leading causes of cancer related deaths worldwide. Polyps are predecessors to this type of cancers and therefore important to discover early by clinicians through colonoscopy examinations. To reduce the occurrence of CRC, it is routine to resect the neoplastic lesions (for example, adenomatous polyps) [1]. Unfortunately, many adenomatous polyps are missed during the endoscopic examinations [2]. A Computer-Aided Detection (CAD) system that, in real-time, can highlight the locations of polyps in the video stream from the endoscope, can act as a second observer, potentially drawing the endoscopist's attention to the polyps displayed on the monitor. This can reduce the chance that some polyps are overlooked [3]. For this purpose, an important improvement of pure anomaly detection approaches, which only identify whether or not there is something abnormal in an image, we also want our CAD system to have pixel-wise segmentation capability so that the specific regions of interest within each abnormal image can be identified.

A key challenge for designing a precise CAD system for polyps is the high costs of collecting and labeling proper medical datasets for training and testing. Polyps come in a wide variety of shapes, sizes, colors, and appearances as shown in Figure 1. For the four main classes of polyps: adenoma, serrated, hyperplastic, and mixed (rare), there are high inter-class similarity and intra-class variation. There can also be high background object similarity, for instance, where parts of a polyp is covered with stool or when they blend into the background mucosa. Although these factors make our task challenging, we conjecture that there is still a high potential for designing a system with a performance acceptable for clinical use.

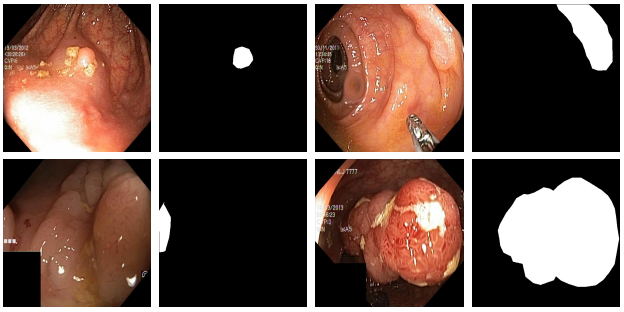


Fig. 1. Examples of polyp images and their corresponding masks from Kvasir-SEG dataset. The first and third column represents the original images, and the second column and fourth column represents their corresponding ground truth.

Motivated by the recent success of semantic segmentation-based approaches for medical image analysis [4]–[6], we explore how these methods can be used to improve the performance for automatic polyp segmentation and detection. A popular deep learning architecture in the field of semantic segmentation for biomedical application is U-Net [5], which have shown state-of-the-art performance at the 2015 ISBI cell tracking challenge [1]. The ResUNet [6] architecture, is a variant of U-Net architecture that has provided state-of-the-art results for the road image extraction. We therefore adapt this architecture as a basis for our work.

In this paper, we propose the ResUNet++ architecture for medical image segmentation. We have evaluated our model on two publicly available datasets. Our experimental results reveal that the improved model is efficient and achieved a performance boost compared to the popular U-Net [5] and ResUNet [6] architectures.

[1]http://brainiac2.mit.edu/isbi_challenge/.

In summary, the contributions of the paper are as follows:

1) We propose the novel ResUNet++ architecture, which is a semantic segmentation neural network that takes advantage of residual blocks, squeeze and excitation blocks, Atrous Spatial Pyramidal Pooling (ASPP), and attention blocks. ResUNet++ improved the segmentation results significantly for the colorectal polyps compared to other state-of-the-art methods. The proposed architecture works well with a smaller number of images.
2) We annotated the polyp class from the Kvasir dataset [7] with the help of an expert gastroenterologist to create the new Kvasir-SEG dataset [8]. We make this polyp segmentation dataset available to the research community to foster development of new methods and reproducible research.

II. Related Work

Automatic gastrointestinal (GI) tract disease detection and classification in colonoscopic videos has been an active area of research for the past two decades. Polyp detection has in particular been given attention. The performance of the machine learning software has come close to the level of expert endoscopists [9]–[12].

Apart from work on algorithm development, researchers have also investigated complete CAD systems, from data annotation, analysis, and evaluation to visualization for the medical experts [13]–[15]. Thambawita et al. [16] explored various methods, ranging from Machine Learning (ML) to deep Convolutional Neural Network (CNN), and suggested five novel models as a potential solution for classifying GI tract findings into sixteen classes. Guo et al. [17] presented two variants of fully convolutional neural networks, which secured the first position at the 2017 Gastrointestinal Image ANAlysis (GIANA) challenge and second position at the 2018 GIANA challenge.

Long et al. [18] proposed a state-of-the-art semantic segmentation approach for image segmentation known as a Fully Convolutional Network (FCN). FCN are trained end-to-end, pixels-to-pixels, and outputs segmentation result without any additional post-processing steps. Ronneberger et al. [5] modified and extended the FCN architecture to an U-Net architecture. There are various modification and extension based on U-Net architecture [4], [6], [17], [19]–[22] to achieve better segmentation results on both natural images and biomedical images.

Most of the published work in the field of polyp detection perform well on the specific datasets, and test scenarios often used small training and validation datasets [11], [23]. The model evaluated on the smaller dataset is neither generalizable nor robust. Moreover, some of the research work only focus on a specific type of polyps. Some of the current work also use non-publicly datasets, which makes it difficult to compare and reproduce results. Therefore, the goal of the ML models to reach a performance level similar to, or better than colonoscopists has not been achieved yet. There exists a potential for improvement in boosting the performance of the system.

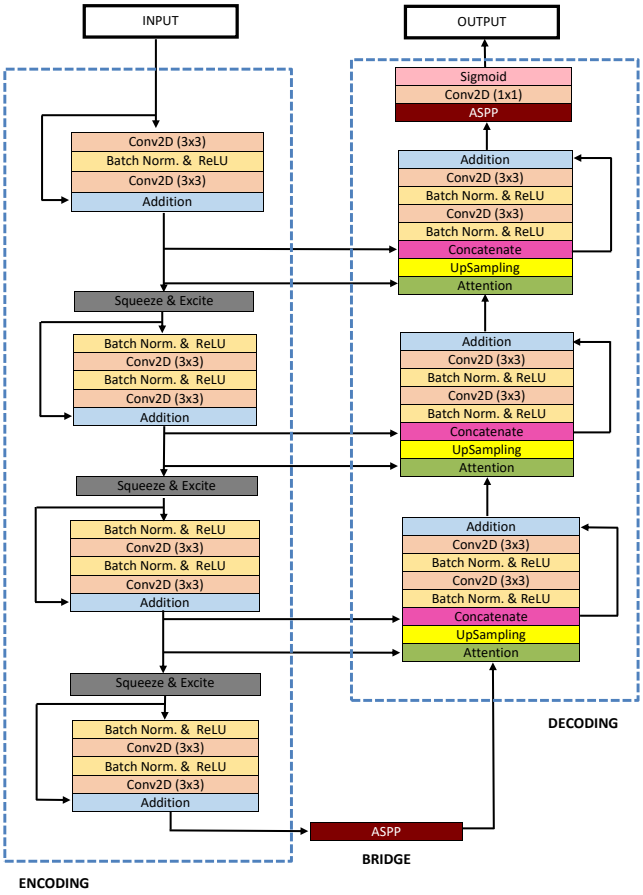


Fig. 2. Block diagram of the proposed ResUNet++ architecture.

III. ResUNet++

The ResUNet++ architecture is based on the Deep Residual U-Net (ResUNet) [6], which is an architecture that uses the strength of deep residual learning [24] and U-Net [5]. The proposed ResUNet++ architecture takes advantage of the residual blocks, the squeeze and excitation block, ASPP, and the attention block.

The residual block propagates information over layers, allowing to build a deeper neural network that could solve the degradation problem in each of the encoders. This improves the channel inter-dependencies, while at the same time reducing the computational cost. The proposed ResUNet++ architecture contains one stem block followed by three encoder blocks, ASPP, and three decoder blocks. The block diagram of the proposed ResUNet++ architecture is shown in Figure 2. In the block diagram, we can see that the residual unit is a combination of batch normalization, Rectified Linear Unit (ReLU) activation, and convolutional layers.

Each encoder block consists of two successive 3×3 convolutional block and an identity mapping. Each convolution block includes a batch normalization layer, a ReLU activation

layer, and a convolutional layer. The identity mapping connects the input and output of the encoder block. A strided convolution layer is applied to reduce the spatial dimension of the feature maps by half at the first convolutional layer of the encoder block. The output of encoder block is passed through the squeeze-and-excitation block. The ASPP acts as a bridge, enlarging the field-of-view of the filters to include a broader context. Correspondingly, the decoding path consists of residual units, too. Before each unit, the attention block increases the effectiveness of feature maps. This is followed by a nearest-neighbor up-sampling of feature maps from the lower level and the concatenation with feature maps from their corresponding encoding path.

The output of the decoder block is passed through ASPP, and finally, we apply a 1×1 convolution with sigmoid activation, that provides the segmentation map. The extension of the ResUNet++ is the squeeze-and-excitation blocks marked in light blue, the ASPP block marked in dark red, and attention block marked in light green. A brief explanation of each of the parts is given in the following subsections.

A. Residual Units

Deeper Neural Networks are comparatively challenging to train. Training a deep neural network with an increasing network depth can improve accuracy. However, it can hamper the training process and cause a degradation problem [6], [24]. He et al. [24] proposed a deep residual learning framework to facilitate the training process and address the problem of degradation. ResUNet [6] uses full pre-activation residual units. The deep residual unit makes the deep network easy to train and the skip connection within the networks helps to propagate information without degradation, improving the design of the neural network by decreasing the parameters along with comparable performance or boost in performance on semantic segmentation task [6], [24]. Because of these advantages, we use ResUNet as the backbone architecture.

B. Squeeze and Excitation Units

The squeeze-and-excitation network [25] boosts the representative power of the network by re-calibrating the features responses employing precise modeling inter-dependencies between the channels. The goal of the squeeze and excite block is to ensure that the network can increase its sensitivity to the relevant features and suppress the unnecessary features. This goal is achieved in two steps. The first step is squeeze (global information embedding), where each channel is squeezed by using global average pooling for generating channel-wise statistics. The second step is excitation (active calibration) that aims to capture the channel-wise dependencies fully [25]. In the proposed architecture, the squeeze and excitation block is stacked together with the residual block to increase effective generalization over different datasets and improve the performance of the network.

C. Atrous Spatial Pyramidal Pooling

The idea of ASPP comes from spatial pyramidal pooling [26], which is successful at re-sampling features at multiple scales. In ASPP, the contextual information is captured at various scales [27], [28] and many parallel atrous convolutions [29] with different rates in the input feature map are fused. Atrous convolution allows controlling the field-of-view for capturing multi-scale information precisely. In the proposed architecture, ASPP acts as a bridge between encoder and decoder in our architecture, as shown in Figure 2. The ASPP model has shown promising results on various segmentation tasks by providing multi-scale information. Therefore, we use ASPP to capture the useful multi-scale information for the semantic segmentation task.

D. Attention Units

The attention mechanism is mostly popular in Natural Language Processing (NLP) [30]. It gives attention to the subset of its input. Moreover, it has been employed in semantic segmentation tasks, like pixel-wise prediction [31]. The attention mechanism determines which parts of the network require more attention in the neural network. The attention mechanism also reduces the computational cost of encoding the information in each polyp image into a vector of fixed dimension. The main advantage of the attention mechanism is that they are simple, can be applied to any input size, enhance the quality of features that boosts the results.

In the previous two approaches, U-Net [5] and ResUNet [6], there exists a direct concatenation of the encoder feature maps with the decoder feature maps. Inspired by the success of attention mechanism, both in NLP and computer vision tasks, we implemented the attention block in the decoder part of our architecture to be able to focus on the essential areas of the feature maps.

IV. Experiments

To evaluate the ResUNet++ architecture, we train, validate, and test models using two publicly available datasets. We compare the performance of our ResUNet++ models with ones trained using U-Net and ResUNet.

A. Datasets

For the task of polyp image segmentation, each pixel in the training images must be labeled as belonging to either the polyp class or the non-polyp class. For the evaluation of ResUNet++, we use the Kvasir-SEG dataset [8], which consists of 1,000 polyp images and their corresponding ground truth masks annotated by expert endoscopists from Oslo University Hospital (Norway). Example images and their corresponding masks from the Kvasir-SEG dataset are shown in Figure 1. The second dataset we have used is the CVC-ClinicDB database [32], which is an open-access dataset of 612 images with a resolution of 384×288 from 31 colonoscopy sequences.

B. Implementation details

All architectures were implemented using the Keras framework [33] with TensorFlow [34] as backend. We performed our experiment on a single Volta 100 GPU on a powerful Nvidia DGX-2 AI system capable of 2-petaFLOPS tensor performance. The system is part of Simula Research Laboratories

heterogeneous cluster and has dual Intel(R) Xeon(R) Platinum 8168 CPU@2.70GHz, 1.5TB of DDR4-2667MHz DRAM, 32TB of NVMe scratch space, and 16 of NVIDIAs latest Volta 100 GPGPUs interconnected using Nvidia's NVlink fully non-blocking crossbars switch capable of 2.4 TB/s of bisectional bandwidth. The system was running Ubuntu 18.04.3LTS OS and had the latest Cuda 10.1.243 installed. We start the training with a batch size of 16, and the proposed architecture is optimized by Adam optimizer. The learning rate of the algorithm is set to $1e{-}4$. A lower learning rate is preferred, although a lower learning rate slowed down convergence, and a larger learning rate often causes convergence failures.

The size of the image within the same dataset varies. Both the dataset used in the study consists of different resolution images. For efficient GPU utilization and to reduce the training time, we crop the images by putting a crop margin of 320×320 to increase the training dataset. Then, the images are resized to 256×256 pixels before feeding the images to the model. We have used the data augmentation technique such as center crop, random crop, horizontal flip, vertical flip, scale augmentation, random rotation, cutout, and brightness augmentation, etc., to increase the number of training samples. The rotation angle is randomly chosen from 0 to $90°$. We have utilized 80% of the dataset for training, 10% for validation, and 10% for the testing. We trained all the models for 120 epochs with a lower learning rate so that a more generalized model can be built. The batch size, epoch, and learning rate were reset depending upon the need. There was an accuracy trade-off if we decrease the batch size; however, we preferred a larger batch size over accuracy because smaller batch size can lead to over-fitting. We also used the Stochastic Gradient Descent with Restart (SGDR) to improve the performance of the model.

V. Results

To show the effectiveness of ResUNet++, we conducted two sets of experiments on Kvasir-SEG and CVC-612 datasets. For the model comparison, we compared the results of the proposed ResUNet++ with the original U-Net and original ResUNet architecture, as both of them are the common preference for the semantic segmentation task. The original implementation of ResUNet, which uses Mean Square Error (MSE) as the loss function, did not produce satisfactory results with Kvasir-SEG and CVC-612 datasets. Therefore, we replaced the MSE loss function with dice coefficient loss and did hyperparameter optimization to improve the results and named the architecture as ResUNet-mod. With this modification, we achieved a performance boost in ResUNet-mod architecture for both the datasets.

A. Results on the Kvasir-SEG dataset

We have tried different sets of hyperparameters (i.e., learning rate, number of epochs, optimizer, batch size, and filter size) for the optimization of ResUNet++ architecture. Hyperparameter tuning is done manually by training the models with different sets of hyperparameters and evaluating their results. The results of ResUNet++, ResUNet-mod, ResUNet [6], and U-Net [5] are presented in Table I. Table I shows that the proposed model achieved the highest dice coefficient, mIoU, recall, and competitive precision for the Kvasir-SEG dataset. U-Net achieved the highest precision. However, the dice coefficient and mIoU scores are not competitive, which is an important metric for semantic segmentation task. The proposed architecture has outperformed the baseline architectures by a significant margin in terms of mIoU.

TABLE I
THE TABLE SHOWS THE EVALUATION RESULTS OF ALL THE MODELS ON KVASIR-SEG DATASET.

Method	Dice	mIoU	Recall	Precision
ResUNet++	**0.8133**	**0.7927**	**0.7064**	0.8774
ResUNet-mod	0.7909	0.4287	0.6909	0.8713
ResUNet	0.5144	0.4364	0.5041	0.7292
U-Net	0.7147	0.4334	0.6306	**0.9222**

B. Results on the CVC-612 dataset

We have performed additional experiments for in-depth performance analysis for automatic polyp segmentation. Therefore, we attempted for the generalization of the model to check the generalizability capability of the proposed architecture on a different dataset. Generalizability would be a further step toward building clinical acceptable model. Table II shows the results for all the architectures on CVC-612 datasets. The proposed model obtained highest dice coefficient, mIoU, and recall and competitive precision.

Figure 3 shows the qualitative results for all the models. From Table I, Table II, and Figure 3 we demonstrate the superiority of ResUNet++ over the baseline architectures. The quantitative and qualitative result shows that the ResUNet++ model trained on Kvasir-SEG and CVC-612 dataset performs well and outperforms all other models in terms of dice coefficient, mIoU, and recall. Therefore, the ResUNet++ architecture should be considered over these baselines architecture in the medical image segmentation task.

VI. Discussion

The proposed ResUNet++ architecture produces satisfactory results on both Kvasir-SEG and CVC-612 datasets. From Figure 3, it is evident that the segmentation map produced

TABLE II
THE TABLE SHOWS THE EVALUATION RESULTS OF ALL THE MODELS ON CVC-612 DATASET.

Method	Dice	mIoU	Recall	Precision
ResUNet++	**0.7955**	**0.7962**	**0.7022**	0.8785
ResUNet-mod	0.7788	0.4545	0.6683	**0.8877**
ResUNet	0.4510	0.4570	0.5775	0.5614
U-Net	0.6419	0.4711	0.6756	0.6868

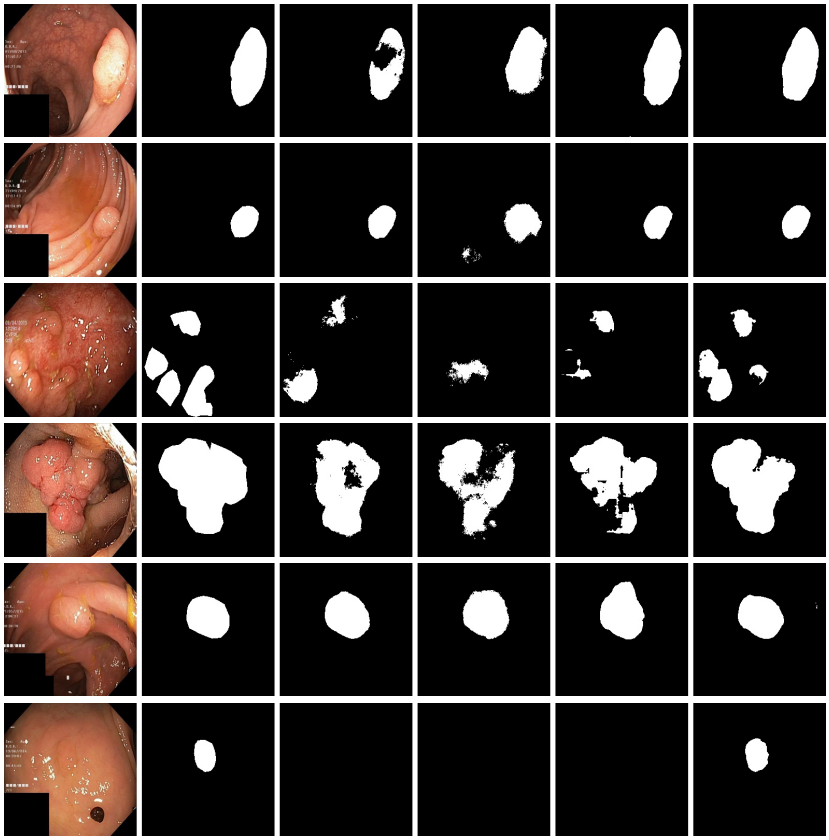


Fig. 3. Qualitative results comparison on the Kvasir-SEG dataset. From the left: image (1), (2) Ground truth, (3) U-Net, (4) ResUNet, (5) ResUNet-mod, and (6) ResUNet++. From the experimental results, we can say that ResUNet++ produces better segmentation masks than other competitors.

by ResUNet++ outperforms other architectures in capturing shape information, in the Kvasir-SEG dataset. It means the generated segmentation mask in ResUNet++ has more similar ground truth then the presented state-of-the-art models. However, ResUNet-mod and U-Net also produced competitive segmentation masks.

We trained the model using different available loss functions, for example, binary cross-entropy, the combination of binary cross-entropy and dice loss, and mean square loss. We observed that the model achieved a higher dice coefficient value with all the loss function. However, mIoU were significantly lower with all other except dice coefficient loss function. We selected the dice coefficient loss function based on our empirical evaluation. Moreover, we also observed that the number of filters, batch size, optimizer, and loss function can influence the result.

We conjecture that the performance of the model can be further improved by increasing the dataset size, applying more augmentation techniques, and by applying some post-processing steps. Despite increased numbers of parameters with the proposed architecture, we trained the model to achieve higher performance. We conclude that the application of ResUNet++ should not only limited to biomedical image segmentation but could also be expanded to the natural image segmentation and other pixel-wise classification tasks, which need further detailed validations. We have optimized the code as much as possible based on our knowledge and experience. However, there may exist further optimization, which may also influence the results of the architectures. We have run the code only on a Nvidia-DGX-2 machine, and the images were resized, which may have lead to the loss of some useful information. Additionally, ResUNet++ uses more parameters, which increases training time.

VII. Conclusion

In this paper, we presented ResUNet++, which is an architecture to address the need for more accurate segmentation of colorectal polyps found in colonoscopy examinations. The suggested architecture takes advantage of residual units, squeeze and excitation units, ASPP, and attention units. Comprehensive evaluation using different available datasets demonstrates that the proposed ResUNet++ architecture outperforms the state-of-the-art U-Net and ResUNet architectures in terms of producing semantically accurate predictions. Towards achieving the generalizability goal, the proposed architecture can be a strong baseline for further investigation in the direction of developing a clinically useful method. Post-processing techniques can potentially be applied to our model to achieve even better segmentation results.

ACKNOWLEDGEMENT

This work is funded in part by Research Council of Norway project number 263248. The computations in this paper were performed on equipment provided by the Experimental Infrastructure for Exploration of Exascale Computing (eX3), which is financially supported by the Research Council of Norway under contract 270053.

REFERENCES

[1] A. G. Zauber, S. J. Winawer, M. J. O'Brien, I. Lansdorp-Vogelaar, M. van Ballegooijen, B. F. Hankey, W. Shi, J. H. Bond, M. Schapiro, J. F. Panish *et al.*, "Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths," *New England Journal of Medicine*, vol. 366, no. 8, pp. 687–696, 2012.

[2] J. C. Van Rijn, J. B. Reitsma, J. Stoker, P. M. Bossuyt, S. J. Van Deventer, and E. Dekker, "Polyp miss rate determined by tandem colonoscopy: a systematic review," *The American journal of gastroenterology*, vol. 101, no. 2, p. 343, 2006.

[3] Y. Mori and S.-e. Kudo, "Detecting colorectal polyps via machine learning," *Nature biomedical engineering*, vol. 2, no. 10, p. 713, 2018.

[4] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proceeding of International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 565–571.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[6] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[7] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt, M. Riegler, and P. Halvorsen, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proc. of MMSYS*, june 2017, pp. 164–169.

[8] D. Jha, P. H. Smedsrud, M. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. Johansen, "Kvasir-seg: A segmented polyp dataset," in *International Conference on Multimedia Modeling*. Springer, 2020. [Online]. Available: https://datasets.simula.no/kvasir-seg/

[9] Y. Wang, W. Tavanapong, J. Wong, J. Oh, and P. C. De Groen, "Part-based multiderivative edge cross-sectional profiles for polyp detection in colonoscopy," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1379–1389, 2014.

[10] Y. Mori, S.-e. Kudo, T. M. Berzin, M. Misawa, and K. Takeda, "Computer-aided diagnosis for colonoscopy," *Endoscopy*, vol. 49, no. 8, pp. 813–819, 2017.

[11] P. Brandao, O. Zisimopoulos, E. Mazomenos, G. Ciuti, J. Bernal, M. Visentini-Scarzanella, A. Menciassi, P. Dario, A. Koulaouzidis, A. Arezzo *et al.*, "Towards a computed-aided diagnosis system in colonoscopy: automatic polyp segmentation using convolution neural networks," *Journal of Medical Robotics Research*, vol. 3, no. 2, p. 1840002, 2018.

[12] P. Wang, X. Xiao, J. R. G. Brown, T. M. Berzin, M. Tu, F. Xiong, X. Hu, P. Liu, Y. Song, D. Zhang *et al.*, "Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy," *Nature biomedical engineering*, vol. 2, no. 10, pp. 741–748, 2018.

[13] Y. Wang, W. Tavanapong, J. Wong, J. H. Oh, and P. C. De Groen, "Polyp-alert: Near real-time feedback during colonoscopy," *International Journal of Computer methods and programs in biomedicine*, vol. 120, no. 3, pp. 164–179, 2015.

[14] M. Riegler, K. Pogorelov, S. L. Eskeland, P. T. Schmidt, Z. Albisser, D. Johansen, C. Griwodz, P. Halvorsen, and T. D. Lange, "From annotation to computer-aided diagnosis: Detailed evaluation of a medical multimedia system," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, no. 3, p. 26, 2017.

[15] S. A. Hicks, S. Eskeland, M. Lux, T. de Lange, K. R. Randel, M. Jeppsson, K. Pogorelov, P. Halvorsen, and M. Riegler, "Mimir: an automatic reporting and reasoning system for deep learning based analysis in the medical domain," in *Proceedings of the ACM Multimedia Systems Conference*. ACM, 2018, pp. 369–374.

[16] V. Thambawita, D. Jha, M. Riegler, P. Halvorsen, H. L. Hammer, H. D. Johansen, and D. Johansen, "The medico-task 2018: Disease detection in the gastrointestinal tract using global features and deep learning," in *Working Notes Proceedings of the MediaEval Workshop*. CEUR Workshop Proceedings, 2018.

[17] Y. B. Guo and B. Matuszewski, "Giana polyp segmentation with fully convolutional dilation neural networks," in *Proceedings of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SCITEPRESS-Science and Technology Publications, 2019, pp. 632–641.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)*, 2015, pp. 3431–3440.

[19] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *Proceeding of International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424–432.

[20] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep Learning and Data Labeling for Medical Applications*. Springer, 2016, pp. 179–187.

[21] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data," *arXiv preprint arXiv:1904.00592*, 2019.

[22] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 3–11.

[23] Y. Wang, W. Tavanapong, J. Wong, J. Oh, and P. C. De Groen, "Part-based multiderivative edge cross-sectional profiles for polyp detection in colonoscopy," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1379–1389, 2013.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 770–778.

[25] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)*, 2018, pp. 7132–7141.

[26] K. He., X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[27] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[28] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[29] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[31] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," *arXiv preprint arXiv:1805.10180*, 2018.

[32] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.

[33] F. Chollet *et al.*, "Keras," 2015.

[34] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proceeding of {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI})*, 2016, pp. 265–283.

A.2 Paper II: Kvasir-SEG: A segmented polyp dataset

Authors: D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen

Abstract: Pixel-wise image segmentation is a highly demanding task in medical-image analysis. In practice, it is difficult to find annotated medical images with corresponding segmentation masks. In this paper, we present Kvasir-SEG: an open-access dataset of gastrointestinal polyp images and corresponding segmentation masks, manually annotated by a medical doctor and then verified by an experienced gastroenterologist. Moreover, we also generated the bounding boxes of the polyp regions with the help of segmentation masks. We demonstrate the use of our dataset with a traditional segmentation approach and a modern deep-learning based Convolutional Neural Network (CNN) approach. The dataset will be of value for researchers to reproduce results and compare methods. By adding segmentation masks to the Kvasir dataset, which only provide frame-wise annotations, we enable multimedia and computer vision researchers to contribute in the field of polyp segmentation and automatic analysis of colonoscopy images.

Published: Proceedings of International Conference on Multimedia Modeling (MMM), pp. 451–462, 2020.

Candidate contributions: D. Jha conceptualized this work, annotated and prepared the dataset with the help of a medical doctor and an expert gastroenterologist. He conducted all the experiments and wrote the manuscript with input from all of the co-authors. He also hosted the dataset on the web page and made the dataset publicly available to the research community for academic and industrial research. Additionally, he revised the manuscript and presented it at the conference.

Thesis objectives: Objective I

Kvasir-SEG: A Segmented Polyp Dataset

Debesh Jha[1,2( )], Pia H. Smedsrud[1,3,4], Michael A. Riegler[1,7], Pål Halvorsen[1,6], Thomas de Lange[4,5], Dag Johansen[2], and Håvard D. Johansen[2]

[1] SimulaMet, Oslo, Norway
debesh@simula.no
[2] UIT The Arctic University of Norway, Tromsø, Norway
[3] Augere Medical AS, Oslo, Norway
[4] University of Oslo, Oslo, Norway
[5] Oslo University Hospital, Oslo, Norway
[6] Oslo Metropolitan University, Oslo, Norway
[7] Kristiania University College, Oslo, Norway

Abstract. Pixel-wise image segmentation is a highly demanding task in medical-image analysis. In practice, it is difficult to find annotated medical images with corresponding segmentation masks. In this paper, we present Kvasir-SEG: an open-access dataset of gastrointestinal polyp images and corresponding segmentation masks, manually annotated by a medical doctor and then verified by an experienced gastroenterologist. Moreover, we also generated the bounding boxes of the polyp regions with the help of segmentation masks. We demonstrate the use of our dataset with a traditional segmentation approach and a modern deep-learning based Convolutional Neural Network (CNN) approach. The dataset will be of value for researchers to reproduce results and compare methods. By adding segmentation masks to the Kvasir dataset, which only provide frame-wise annotations, we enable multimedia and computer vision researchers to contribute in the field of polyp segmentation and automatic analysis of colonoscopy images.

Keywords: Medical images · Polyp segmentation · Semantic segmentation · Kvasir-SEG dataset · Fuzzy C-mean clustering · ResUNet

1 Introduction

Colorectal cancer is the second most common cancer type among women and third most common among men [25]. Polyps are precursors to colorectal cancer and therefore important to detect and remove at an early stage. Polyps are found in nearly half of the individuals at age 50 that undergo a colonoscopy screening, and their frequency increase with age [21]. Polyps are abnormal tissue growth from the mucous membrane, which is lining the inside of the GI tract, and can sometimes be cancerous. Colonoscopy is the gold standard for detection and

Y. M. Ro et al. (Eds.): MMM 2020, LNCS 11962, pp. 451–462, 2020.
https://doi.org/10.1007/978-3-030-37734-2_37

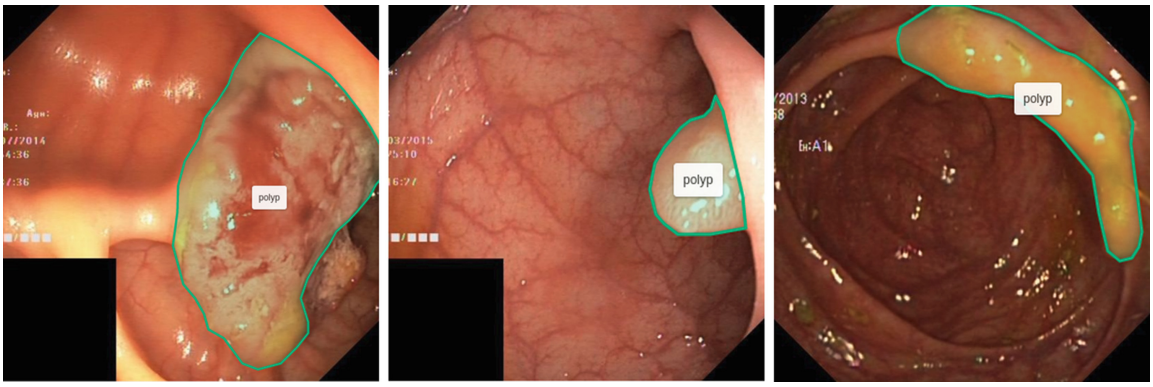


Fig. 1. Example frames from the Kvasir dataset where we additionally have marked the polyp tissue with green outlines. (Color figure online)

assessment of these polyps with subsequent biopsy and removal of the polyps. Early disease detection has a huge impact on survival from colorectal cancer [11]. In addition, several studies show that polyps are often overlooked during colonoscopies, with polyp miss rates of 14 to 30% depending on type and size of the polyps [26]. Increasing the detection of polyps has been shown to decrease risk of colorectal cancer [12]. Thus, automatic detection of more polyps at an early stage can play a crucial role in prevention and survival from colorectal cancer. This is the main motivation behind the development of a polyp segmentation dataset.

Image segmentation is the technique of dividing images into meaningful Regions of Interests (ROIs) that are simple to analyze and interpret. Further research in medical image segmentation can assist processes such as monitoring pathology, improving the diagnostic ability by increasing accuracy, precision, and reducing manual intervention [15]. In particular, for Computer Assisted Interventions (CAIs), pixel-wise semantic segmentation methods have a huge potential to become part of fast, accurate and cost-effective systems.

The goal of image segmentation is to assign a label to each pixel of the image so the pixels with the same label share specific characteristics, e.g., the pixels covered by the outline in the Fig. 1 show a polyp. Manual segmentation by physicians is still the gold standard for most of the medical imaging modalities, for example, Magnetic Resonance Imaging for evaluating hippocampal atrophy in Alzheimer's Disease [5] and tumor segmentation of glioma [27]. However, manual image segmentation is tedious, time-consuming, and subject to physician's bias and inter-observer variation. Therefore, there is a need for an automated and efficient image segmentation technique. Methods for automated and efficient image segmentation are difficult to develop as state-of-the art machine learning methods often require large number of annotated and labelled quality data, which is difficult to obtain in this field. Annotating medical data such as polyp images manually requires a lot of time and effort. It also requires medical experts, gastroenterologists in our case, which can be expensive and inaccessible. Also, there are problems related to the collection of medical images, with concern to privacy and security for patients and hospitals. Riegler et al. [19] have raised

several open questions about the medical world that need to be addressed, where they emphasized the need for test datasets, including annotations and ground truth that meet current medical standards. Although there are a few available datasets, open-access datasets for comparable evaluations are missing in this field. We therefore provide the Kvasir-SEG dataset and propose a baseline model for evaluation.

The main contribution of this paper are as follows:

1. We extend the Kvasir dataset [16] with polyp images along with their corresponding segmentation masks and bounding boxes. The ROIs are the pixels depicting polyp tissue. These are represented by a white foreground in the segmentation masks. The ROIs are generated from manual annotations verified by an experienced gastroenterologist. The bounding boxes are the set of coordinates that encloses the polyp regions. The Kvasir-SEG dataset is made publicly available and open access.
2. In this article, we include a first attempt to use the Kvasir-SEG dataset for pixel-wise semantic segmentation based analysis. For the experiment, we have used Fuzzy C-mean clustering (FCM) [6] and Deep Residual U-Net (ResUNet) [28] architecture. We achieved promising results with our proposed methods when evaluated on the same dataset. We evaluated the proposed method using Dice Coefficients and mean Intersection over Union (IoU). These metrics were selected for fair comparison, and we encourage the use of these and similar metrics in future work on the dataset. The promising results demonstrated in this paper serves as baseline and motivation for further research and evaluation done on the same dataset.
3. Multiple datasets are prerequisites for comparing computer vision based algorithms, and this dataset is useful both as a training dataset or as a validation dataset. This dataset can assist the development of state-of-the-art solutions on images captured by colonoscopes. Further research in this field has the potential to help reduce the polyp miss rate and thus improve examination quality.

This paper is organized as follows: Sect. 2 discusses related datasets. We discuss the Kvasir-SEG dataset in Sect. 3. In Sect. 4, we define the suggested metrics for the segmentation of polyps. Section 5 describes the baseline experiments, results and discussion. We conclude our work and give future directions in Sect. 6.

2 Related Work

There are only few available polyp datasets that consist of ground truth and corresponding segmentation mask. These are CVC-ColonDB [24], ASU-Mayo Clinic Colonoscopy Video © Database [3], ETIS-Larib Polyp DB [23], and CVC-Clinic DB [2].

CVC-ColonDB [24] is the second largest database available and consists of annotated video sequences from colonoscopy videos. From 15 short colonoscopy

sequences, 1200 image frames are extracted. Out of these images, only 300 frames are annotated. These annotated frames were specifically chosen to maximize the the visual differences between them. Use of the CVC-ColonDB requires registration.

The ASU-Mayo Clinic Colonoscopy Video © Database [3] is the first and largest available dataset captured using standard colonoscopes. The training dataset consists of 18,781 frames extracted from 20 short videos. Of these, there are 10 videos of polyps (positive) and 10 videos without polyps (negative). Ground truth and its corresponding segmentation masks are provided with the more than 3,500 frames showing polyps. For testing, 18 videos without ground truth are included. The images in the dataset are very similar to each other, which raise the problem of overfitting [16]. The ASU-Mayo Clinic Colonoscopy database is copyrighted, and is only available through direct contact with the administrators at Arizona State University.

The ETIS-Larib Polyp DB [23] consists of 196 frames of polyps extracted from colonoscopy videos and their corresponding masks. This database is available through registration.

The CVC-Clinic DB [2] consists of 612 image frames extracted from 29 different colonoscopy sequences and their corresponding ground truth as segmentation masks. The use of this database is public and open access.

The CVC-Clinic DB, ETIS-Larib and ASU-Mayo Clinic Colonoscopy Video DB were used at Medical Image Computing and Computer Assisted Intervention (MICCAI) 2015 Automatic Polyp Detection in Colonoscopy Videos Sub-Challenge. More details about the dataset and competition can be found in the paper by Bernal et al. [4].

The literature review shows that there are few available datasets. However, an open-access dataset for comparable evaluation is missing in this field. Therefore, it was a logical next step to extend the Kvasir dataset with segmentation masks. The presented data and baseline work can be a important source for addressing the problem of standard datasets for evaluation, and help develop robust and efficient systems.

3 The Kvasir-SEG Dataset

The Kvasir-SEG dataset is based on the previous Kvasir [16] dataset, which is the first multi-class dataset for gastrointestinal (GI) tract disease detection and classification.

3.1 The Original Kvasir Dataset

The original Kvasir dataset [16] comprises 8,000 GI tract images from 8 classes where each class consists of 1000 images. We replaced the 13 images from the polyp class with new images to improve the quality of the dataset. These images were collected and verified by experienced gastroenterologists from Vestre Viken Health Trust in Norway. The classes include anatomical landmarks, pathological

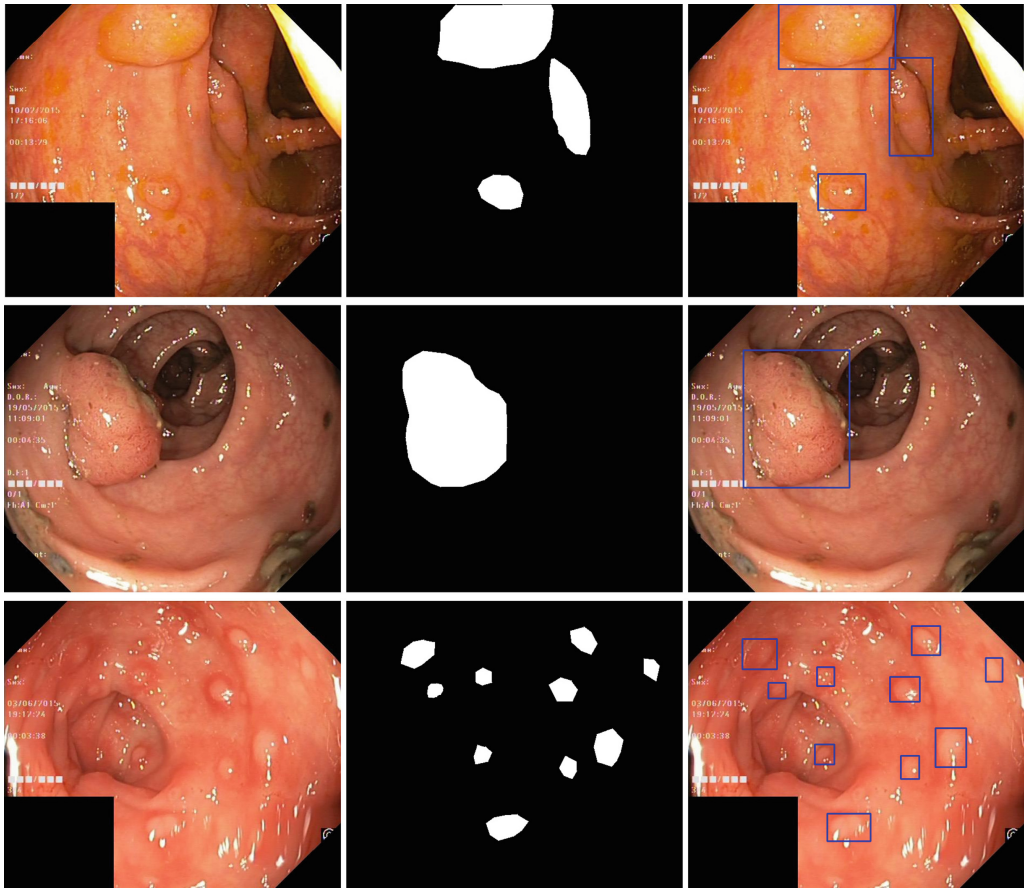


Fig. 2. Examples of polyp images and their corresponding masks from Kvasir-SEG. The third image is generated from the original image using the bounding box information from the JSON file.

findings and endoscopic procedures. A more detailed explanation about each image classes, the data collection procedure and the dataset details can be found in [16].

The Kvasir dataset was used for the Multimedia for Medicine Challenge (the Medico Task) in 2017 [20] and 2018 [17] at the MediaEval Benchmarking Initiative for Multimedia Evaluation[1] to develop and compare methods to reach clinical level performance on multiclass classification of endoscopic findings in the large bowel. However, the dataset was limited to frame classification only, due to only a frame-wise annotations. Thus, Pozdeev et al. [18] trained their model on the CVC-ClinicDB, and tried to predict the segmentation masks for the Kvasir dataset, but could not report the experimental scores because of missing ground truth.

3.2 The Kvasir-SEG Dataset Details

To address the high incidence of colorectal cancer, we selected the polyp class of the Kvasir dataset for the initial investigation. The Kvasir-SEG dataset contains

[1] http://www.multimediaeval.org.

annotated polyp images and their corresponding masks. As shown in Fig. 2, the pixels depicting polyp tissue, the ROI, are represented by the foreground (white mask), while the background (in black) does not contain positive pixels. Some of the original images contain the image of the endoscope position marking probe from the ScopeGuide (Olympus).

The Kvasir-SEG dataset is made up of two folders: one for images and one for masks. Each folder contains 1000 images. The bounding boxes for the corresponding images are stored in a JSON file. Therefore, the kvasir-SEG dataset has image folder, masks folder and JSON file. The image and its corresponding mask have the same filename. The image files are encoded using JPEG compression, and online browsing is facilitated. The open-access dataset can be easily downloaded for research purposes at: https://datasets.simula.no/kvasir-seg/.

3.3 Mask Extraction

We uploaded the entire Kvasir polyp class to Labelbox [22] and created all the segmentations using this application. The Labelbox is a tool used for labelling the ROI in image frames, i.e., the polyp regions for our case. A team consisting of an engineer and a medical doctor manually outlined the margins of all polyps in all 1000 images. The annotations were then reviewed by an experienced gastroenterologist.

Figure 1 shows example frames from the kvasir dataset where we have additionally marked the polyp tissue with green outline. After annotation, we exported the files to generate masks for each annotation. The exported JSON file contained all the information about the image and the coordinate points for generating the mask. To create a mask, we used ROI coordinates to draw contours on an empty black image and fill the contours with white color. The generated masks are a 1-bit color depth images with white foreground and black background. Figure 2 shows example images, their corresponding segmentation masks and bounding boxes from the Kvasir-SEG dataset.

4 Suggested Metrics

Different metrics for evaluating and comparing the performance of the architectures exist. For medical image segmentation tasks, the perhaps most commonly used metrics are Dice coefficient and IoU. These are used in particular for several medical related Kaggle competitions [10]. In this medical image segmentation approach, each pixel of the image either belongs to a polyp or non-polyp region. We calculate the Dice coefficient and mean IoU based on this principle.

Dice coefficient: Dice coefficient is a standard metric for comparing the pixel-wise results between predicted segmentation and ground truth. It is defined as:

$$\text{Dice coefficient}(A, B) = \frac{2 \times |A \cap B|}{|A| + |B|} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{1}$$

where A signifies the predicted set of pixels and B is the ground truth of the object to be found in the image. Here, TP represents true positive, FP represents false positive, and FN represents the false negative.

Intersection over Union: The Intersection over Union (IoU) is another standard metric to evaluate a segmentation method. The IoU calculates the similarity between predicted (A) and its corresponding ground truth (B) as shown in the equation below:

$$IoU(A, B) = \frac{A \cap B}{A \cup B} = \frac{TP(t)}{TP(t) + FP(t) + FN(t)} \tag{2}$$

In Eq. 2, t is the threshold. At each threshold value t, a precision value is calculated based on the above equation and parameters, which is done by calculating the predicted object to all the ground truth objects. There are other parameters such as recall, specificity, precision, and accuracy which are mostly used for frame-wise image classification tasks. The detailed explanation about these parameters can be found in the Kvasir dataset paper [16].

5 Evaluation

The Kvasir-SEG dataset is intended for research and development of new and improved methods for segmentation, localization, and classification of polyps. To show that the dataset is useful for these purposes, we conducted several experiments, which we will described next.

5.1 Baseline Models

As our baseline, we have conducted initial investigations using two different methods. The first method is based on the efficient FCM [6] unsupervised clustering algorithm. The second method is based on the deep-learning ResUNet [28] architecture, utilizing the advantage of the residual block.

When using basic CNN architectures to predict outcomes in computer-vision tasks, millions of labelled training data are often needed to counteract overfitting and ensuring the model's ability to generalize when tested on new data [9]. Because large datasets of medical images are hard to come by, using CNNs for medical-image segmentation systems remains challenging. Image augmentation techniques [7] and encoder-decoder architecture such as ResUNet [28] are popular methods to use CNNs with smaller training sets.

5.2 Implementation Details

Before applying the FCM algorithm, several pre-processing steps were applied to the dataset. First, we converted the image to grayscale and applied median blur to reduce noise. Then, we applied the Median-based Otsu method [14], which gave us the ROI. Next, we converted image pixels between 0 and 1 and subtracted

Table 1. Quantitative performance of ResUNet model on Kvasir-SEG dataset.

ResUNet	Loss	Dice coefficient	Mean IoU
Train	0.059389	0.940609	0.920957
Validation	0.196520	0.803479	0.792339
Test	0.212236	0.787763	0.777771

the image with its blurred version. We then used a threshold value and created an image with edges in it. Afterwards, we performed the dilation operation, which increases the foreground (white) region of the image. We subtracted edges from the image and clipped the image pixels value between 0 and 1. After that, we reshaped the image into 1D, which is the input to the FCM. Finally, the output of the FCM was reshaped into a 2D binary mask.

For our experiment with the ResUNet model, we used image augmentation techniques like flipping, random crop, scaling, rotation, brightness, cutout, and random erasing to increase the size of our training dataset. After all preprocessing was completed, we resized our images to 320×320 pixels. We used 80% of the dataset for training, and 10% for validation. The remaining of 10% was used for testing. We used five convolutional blocks both in the encoder and the decoder of the ResUNet model. The batch size was set to 8, and we trained the model for 150 epochs. The proposed model converged at 91 epochs. We used a Nadam optimizer with the learning rate of 0.0001, $\beta 1$ of 0.9 and $\beta 2$ of 0.999. We chose Dice coefficient as the loss function and Relu as non-linearity. We used a threshold value t of 0.5 to convert the predicted masks pixels to foreground or background.

For our deep-learning implementations, we used the Keras framework [8] and Tensorflow [1] as a backend. We performed our experiment on a single Volta 100 GPU on a powerful Nvidia DGX-2 [6] AI system capable of 2 PFLOPS tensor performance. The system is part of Simula Research Laboratories heterogeneous cluster and has dual Intel(R) Xeon(R) Platinum 8168 CPU@2.70 GHz, 1.5 TB of DDR4-2667 MHz DRAM, 32 TB of NVMe scratch space, and 16 of NVIDIAs latest Volta 100 GPGPUs [7] interconnected using Nvidia's NVlink fully non-blocking crossbars switch capable of 2.4 TB/s of bisectional bandwidth. The system was running Ubuntu 18.04.3 LTS OS and had the latest Cuda 10.1.243 installed.

5.3 Results and Discussions

The FCM clustering algorithm achieved a Dice coefficient of 0.239002 and a mean IoU of 0.314187. The ResUNet model achieved a Dice coefficient of 0.787763 and mean IoU of 0.777771 (see Table 1) using the test dataset. We have included the training, validation and testing scores for the ResUNet model in Table 1. Examples of qualitative result comparisons for the FCM algorithm and ResUNet model on the Kvasir-SEG dataset is shown in the Fig. 3.

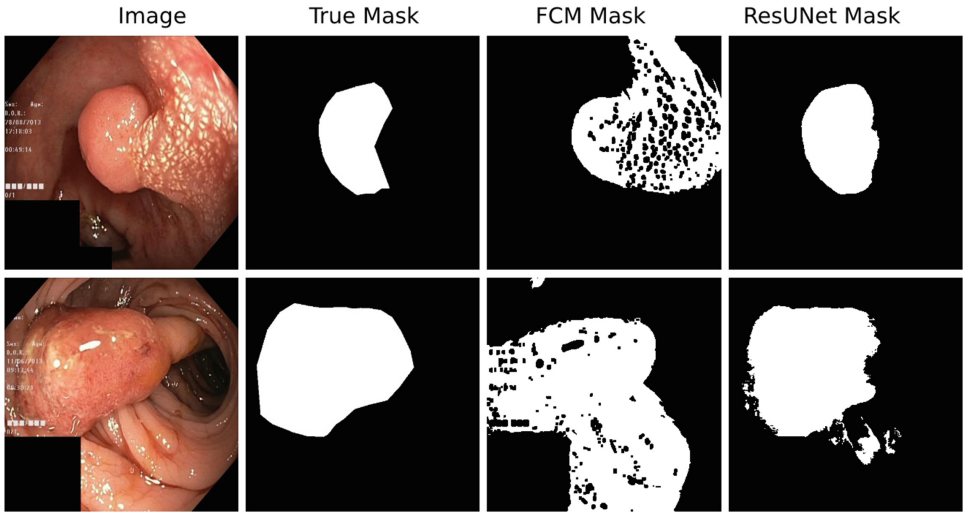


Fig. 3. Qualitative comparison provided by both methods: Coloumn one shows the original image, column two shows the ground truth of the corresponding image. Column three shows the result of FCM clustering and column four shows the results of ResUNet. (Color figure online)

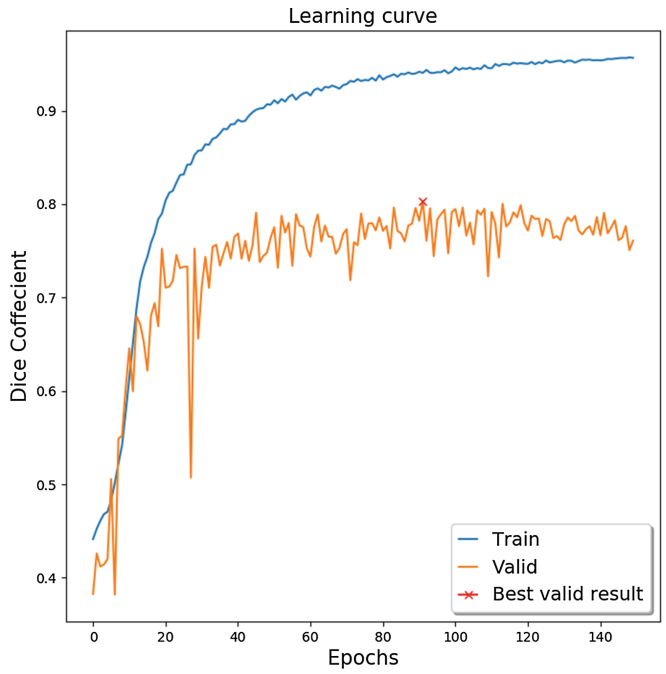


Fig. 4. The learning curve of the proposed ResUNet model on Kvasir-SEG dataset showing Dice coefficient versus number of epochs. (Color figure online)

Considering the quantitative and qualitative results (see Table 1 and Fig. 3), the study shows a superior performance of the ResUNet model over the FCM algorithm in segmenting the polyp pixels. It should be noted that the FCM algorithm uses no data augmentation because it does not have any learning mechanism or learning parameters, whereas the ResUNet utilizes the advantage of the data augmentation techniques. Another important reason why FCM

clustering approach did not perform well as it uses color as a significant feature for discriminating normal tissue and polyp. However, in practice, it is difficult to distinguish between polyps and other conditions inside the GI tract on the basis of color features because of their similar appearances. We achieved promising results with the ResUNet model.

There are no directly comparable papers with regards to our results. Nevertheless, compared to the work of Kang et al. [13] which obtained the IoU of 0.6607 and the work of Pozdeev et al. [18] that showed dice ranging from 0.6200 to 0.8600, we can say our results are either comparable or better. We think that the performance of our ResUNet model can be improved by providing it with more diverse polyp images. The plot of the learning curve of the ResUNet model is shown in Fig. 4. The red mark in the learning curve "x" denote the best model. This best model was used for testing the previously unseen test dataset. The presented results are good; however, we believe that more research is required to achieve performance applicable to the clinic.

6 Conclusion

In this paper, we present Kvasir-SEG: a new polyp segmentation dataset developed to aid multimedia researchers in carrying out extensive and reproducible research. We also present a FCM clustering algorithm and a ResUNet-based approach for automatic polyp segmentation. Our results show that the ResUNet model is outperforming the FCM clustering.

The Kvasir-SEG dataset is released as open-source to the multimedia and medical research communities, hoping it can help evaluate and compare existing and future computer vision methods. This could boost the performance of computer vision methods, an important step towards building clinically acceptable CAI methods for improved patient care.

Acknowledgements. This work is funded in part by the Research Council of Norway projects number 263248 (Privaton). We performed all computations in this paper on equipment provided by the Experimental Infrastructure for Exploration of Exascale Computing (eX^3), which is financially supported by the Research Council of Norway under contract 270053.

References

1. Abadi, M., et al.: Tensorflow: a system for large-scale machine learning. In: Proceeding of the ACM Symposium on Operating Systems Design and Implementation (SOSP), pp. 265–283 (2016)
2. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. Comput. Med. Imag. Graph. **43**, 99–111 (2015)
3. Bernal, J., Sánchez, J., Vilarino, F.: Towards automatic polyp detection with a polyp appearance model. Pattern Recogn. **45**(9), 3166–3182 (2012)

4. Bernal, J., et al.: Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 endoscopic vision challenge. IEEE Trans. Med. Imag. **36**(6), 1231–1249 (2017)
5. Boccardi, M., et al.: Survey of protocols for the manual segmentation of the hippocampus: preparatory steps towards a joint EADC-ADNI harmonized protocol. J. Alzheimer's Dis. **26**(s3), 61–75 (2011)
6. Cai, W., Chen, S., Zhang, D.: Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. Pattern Recogn. **40**(3), 825–838 (2007)
7. Chollet, F.: Building powerful image classification models using very little data. Keras Blog (2016)
8. Chollet, F.: Keras: The Python Deep Learning Library. Astrophysics Source Code Library (2018)
9. Dravid, A.: Employing deep networks for image processing on small research datasets. Microsc. Today **27**(1), 18–23 (2019)
10. Goldbloom, A., Hamner, B., et al.: Kaggle: your home for data science. Competition, Kaggle Inc. (2019). https://www.kaggle.com. Accessed 12 July 2019
11. Haggar, F.A., Boushey, R.P.: Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. Clin. Colon Rectal Surg. **22**(04), 191–197 (2009)
12. Kaminski, M.F., et al.: Increased rate of adenoma detection associates with reduced risk of colorectal cancer and death. Gastroenterology **153**(1), 98–105 (2017)
13. Kang, J., Gwak, J.: Ensemble of instance segmentation models for polyp segmentation in colonoscopy images. IEEE Access **7**, 26440–26447 (2019)
14. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979)
15. Pham, D.L., Xu, C., Prince, J.L.: Current methods in medical image segmentation. Ann. Rev. Biomed. Eng. **2**(1), 315–337 (2000)
16. Pogorelov, K., et al.: Kvasir: a multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of Multimedia Systems Conference (MMSYS), pp. 164–169. ACM (2017)
17. Pogorelov, K., et al.: Medico multimedia task at MediaEval 2018. In: CEUR Workshop Proceedings - Multimedia Benchmark Workshop (MediaEval) (2018)
18. Pozdeev, A.A., Obukhova, N.A., Motyko, A.A.: Automatic analysis of endoscopic images for polyps detection and segmentation. In: IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), pp. 1216–1220. IEEE (2019)
19. Riegler, M., et al.: Multimedia and medicine: teammates for better disease detection and survival. In: Proceedings of ACM Multimedia (ACM MM), pp. 968–977. ACM (2016)
20. Riegler, M., et al.: Multimedia for medicine: the medico task at Mediaeval 2017. In: CEUR Workshop Proceedings - Multimedia Benchmark Workshop (MediaEval) (2017)
21. Rundle, A.G., Lebwohl, B., Vogel, R., Levine, S., Neugut, A.I.: Colonoscopic screening in average-risk individuals ages 40 to 49 vs 50 to 59 years. Gastroenterology **134**(5), 1311–1315 (2008)
22. Sharma, M., Rasmuson, D., Rieger, B., Kjelkerud, D., et al.: Labelbox: the best way to create and manage training data. Software, LabelBox Inc. (2019). https://www.labelbox.com/. Accessed 21 May 2019
23. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. Int. J. Comput. Assist. Radiol. Surg. **9**(2), 283–293 (2014)

24. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans. Med. Imag. **35**(2), 630–644 (2015)
25. Torre, L.A., Bray, F., Siegel, R.L., Ferlay, J., Lortet-Tieulent, J., Jemal, A.: Global cancer statistics, 2012. CA: Cancer J. Clin. **65**(2), 87–108 (2015)
26. Van Rijn, J.C., Reitsma, J.B., Stoker, J., Bossuyt, P.M., Van Deventer, S.J., Dekker, E.: Polyp miss rate determined by tandem colonoscopy: a systematic review. Am. J. Gastroenterol. **101**(2), 343 (2006)
27. Visser, M., et al.: Inter-rater agreement in glioma segmentations on longitudinal MRI. NeuroImage: Clin. **22**, 101727 (2019)
28. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual U-net. IEEE Geosci. Rem. Sens. Lett. **15**(5), 749–753 (2018)

A.3 Paper III: A Comprehensive Study on Colorectal Polyp Segmentation with ResUNet++, Conditional Random Field and Test-Time Augmentation

Authors: D. Jha, P. H. Smedsrud, D. Johansen, T. de Lange, H. Johansen, P. Halvorsen, and M.Riegler

Abstract: Colonoscopy is considered the gold standard for detection of colorectal cancer and its precursors. Existing examination methods are, however, hampered by high overall miss-rate, and many abnormalities are left undetected. Computer-Aided Diagnosis systems based on advanced machine learning algorithms are touted as a game-changer that can identify regions in the colon overlooked by the physicians during endoscopic examinations, and help detect and characterize lesions. In previous work, we have proposed the ResUNet++ architecture and demonstrated that it produces more efficient results compared with its counterparts U-Net and ResUNet. In this paper, we demonstrate that further improvements to the overall prediction performance of the ResUNet++ architecture can be achieved by using CRF and TTA. We have performed extensive evaluations and validated the improvements using six publicly available datasets: Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, ETIS-Larib Polyp DB, ASU-Mayo Clinic Colonoscopy Video Database, and CVC-VideoClinicDB. Moreover, we compare our proposed architecture and resulting model with other State-of-the-art methods. To explore the generalization capability of ResUNet++ on different publicly available polyp datasets, so that it could be used in a real-world setting, we performed an extensive cross-dataset evaluation. The experimental results show that applying CRF and TTA improves the performance on various polyp segmentation datasets both on the same dataset and cross-dataset. To check the model's performance on difficult to detect polyps, we selected, with the help of an expert gastroenterologist, 196 sessile or flat polyps that are less than ten millimeters in size. This additional data has been made available as a subset of Kvasir-SEG. Our approaches showed good results for flat or sessile and smaller polyps, which are known to be one of the major reasons for high polyp miss-rates. This is one of the significant strengths of our work and indicates that our methods should be investigated further for use in clinical practice.

Published: IEEE journal of biomedical and health informatics, 2021.

Candidate contributions: D. Jha conceptualized this work and implemented and configured ResUNet++ with conditional random field and test-time augmentation with two videos and four still images polyp datasets. He further classified sessile polyp with the help of expert gastroenterologists from the Kvasir-SEG dataset, conducted further experiments, and released the dataset publicly. He also performed additional experiments on the cross-dataset, promoting the generalizability study in the field of polyp segmentation. He designed and created all the figures. Finally, he prepared the manuscript and revised it subsequently with the critical input from all of the co-authors.

Thesis objectives: Objective III, Objective I

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3049304, IEEE Journal of Biomedical and Health Informatics

A Comprehensive Study on Colorectal Polyp Segmentation with ResUNet++, Conditional Random Field and Test-Time Augmentation

Debesh Jha, Pia H. Smedsrud, Dag Johansen, Thomas de Lange, Håvard D. Johansen, Pål Halvorsen, and Michael A. Riegler

***Abstract*—Colonoscopy is considered the gold standard for detection of colorectal cancer and its precursors. Existing examination methods are, however, hampered by high overall miss-rate, and many abnormalities are left undetected. Computer-Aided Diagnosis systems based on advanced machine learning algorithms are touted as a game-changer that can identify regions in the colon overlooked by the physicians during endoscopic examinations, and help detect and characterize lesions. In previous work, we have proposed the ResUNet++ architecture and demonstrated that it produces more efficient results compared with its counterparts U-Net and ResUNet. In this paper, we demonstrate that further improvements to the overall prediction performance of the ResUNet++ architecture can be achieved by using Conditional Random Field (CRF) and Test-Time Augmentation (TTA). We have performed extensive evaluations and validated the improvements using six publicly available datasets: Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, ETIS-Larib Polyp DB, ASU-Mayo Clinic Colonoscopy Video Database, and CVC-VideoClinicDB. Moreover, we compare our proposed architecture and resulting model with other State-of-the-art methods. To explore the generalization capability of ResUNet++ on different publicly available polyp datasets, so that it could be used in a real-world setting, we performed an extensive cross-dataset evaluation. The experimental results show that applying CRF and TTA improves the performance on various polyp segmentation datasets both on the same dataset and cross-dataset. To check the model's performance on difficult to detect polyps, we selected, with the help of an expert gastroenterologist, 196 sessile or flat polyps that are less than ten millimeters in size. This additional data has been made available as a subset of Kvasir-SEG. Our approaches showed good results for flat or sessile and smaller polyps, which are known to be one of the major reasons for high polyp miss-rates. This is one of the significant strengths of our work and indicates that our methods should be investigated further for use in clinical practice.**

***Index Terms*—Colonoscopy, polyp segmentation, ResUNet++, conditional random field, test-time augmentation, generalization**

I. Introduction

Cancer is a primary health problem of contemporary society, with colorectal cancer (CRC) being the third most prevailing type in terms of cancer incidence and second in terms of mortality globally [2]. Colorectal polyps are the precursors for the CRC. Early detection of polyps through high-quality colonoscopy and regular screening are cornerstones for the prevention of colorectal cancer [3], since adenomas can be found and resected before transforming to cancer and subsequently reducing CRC morbidity and mortality.

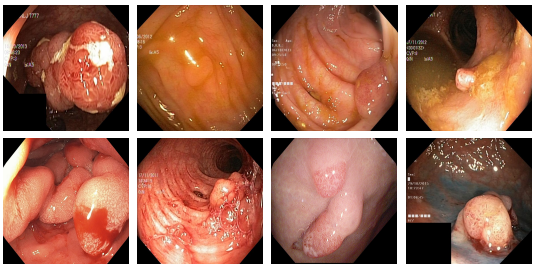


Fig. 1. Example images showing the variations in shape, size, color, and appearance of polyps from the Kvasir-SEG [4].

Regardless of the achievement of colonoscopy in lowering cancer burden, the estimated adenoma miss-rate is around 6-27% [5]. In a recent pooled analysis of 8 randomized tandem colonoscopy studies, polyps smaller than 10 mm, sessile, and flat polyps [6] are shown to most often be missed [7]. Another reason why polyps are missed may be that the polyp either was not in the visual field or was not recognized despite being in the visual field due to fast withdrawal of the colonoscope [8]. The adenoma miss-rate could be reduced by improving the quality of bowel preparation, applying optimal observation techniques, and ensuring a colonoscopy withdrawal time of at least six minutes [8]. Moreover, adenoma detection rate can also be improved by using advanced techniques or devices, for example, auxiliary imaging devices, colonoscopes with increased field of view, add-on-devices, and colonoscopes with integrated inflatable, reusable balloon [3].

The structure and characteristics of a colorectal polyp changes over time at different development stages. Polyps have different shapes, sizes, colors, and appearances, which makes them challenging to analyze (see Figure 1). Moreover, there are challenges such as the presence of image artifacts like blurriness, surgical instruments, intestinal contents, flares, and low-quality images that can cause errors during segmentation.

Polyp segmentation is of crucial relevance in clinical applications to focus on the particular area of the potential lesion,

A preliminary version of this paper was presented in [1].

Manuscript received xxxx-xx-xx; revised xxxx-xx-xx; accepted xxxx-xx-xx; Date of Publication xxxx-xx-xx.

The authors are with the SimulaMet, Norway, Augere Medical AS, Norway, UiT The Arctic University of Norway, University of Oslo, Norway, Oslo Metropolitan University, Norway, Sahlgrenska University Hospital, Mölndal, Sweden, and Bærum Hospital, Vestre Viken, Norway (Corresponding author: Debesh Jha (e-mail: debesh@simula.no))

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3049304, IEEE Journal of Biomedical and Health Informatics

extract detailed information, and possibly remove the polyp if necessary. A Computer-Aided Diagnosis (CADx) system for polyp segmentation can assist in monitoring and increasing the diagnostic ability by increasing the accuracy, precision, and reducing manual intervention. Moreover, it could lead to less segmentation errors than when conducted subjectively. Such systems could reduce doctor's workload and improve clinical workflow. Lumen segmentation helps clinicians navigate through the colon during screening, and it can be useful to establish a quality metric for the explored colon wall [9]. Thus, an automated CADx system could be used as a supporting tool to reduce the miss-rate of the overlooked polyps.

A CADx system could be used in a clinical setting if it addresses two common challenges: (i) Robustness (i.e., the ability of the model to consistently perform well on both easy and challenging images), and (ii) Generalization (i.e., a model trained on specific intervention in a specific hospital should generalize across different hospitals) [10]. Addressing these challenges is key to designing a powerful semantic segmentation system for medical images. Generalization capability checks the usefulness of the model across different available datasets coming from different hospitals and must finally be confirmed in multi-center randomized trials. A good generalizable model could be a significant step toward developing an acceptable clinical system. A cross-dataset evaluation is crucial to check the model on the unseen polyps from other sources and test the generalizability of it.

Toward developing a robust CADx system, we have previously proposed ResUNet++ [1]: an initial encoder-decoder based deep-learning architecture for segmentation of medical images, which we trained, validated, and tested on the publicly available Kvasir-SEG [4] and CVC-ClinicDB [11] datasets. In this paper, we describe how the ResUNet++ architecture can be extended by applying Conditional Random Field (CRF) and Test-Time Augmentation (TTA) to further improve its prediction performance on segmented polyps. We have tested our approaches on six publicly available datasets, including both image datasets and video datasets. We have intentionally incorporated video datasets from colonoscopies to support the clinical significance. Usually, still-frames have at least one polyp sample. Videos have a situation where frames consist of both polyp and non-polyp. Therefore, we have tested the model on these video datasets and provided a new benchmark for the segmentation task. We have used extensive data augmentation to increase the training sample and used a comprehensive hyperparameter search to find optimal hyperparameters for the dataset. We have provided a more in-depth evaluation by including more evaluation metrics, and added justification for the ResUNet++, CRF, and TTA.

Additionally, we have performed extensive experiments on the cross-data evaluation, in-depth analysis of best performing and worst performing cases, and comparison of the proposed method with other recent works. Moreover, we have pointed out the necessity of solving tasks related to the miss-detection of flat and sessile polyps, and showed that our combining approach could detect the overlooked polyps with high efficiency, which could be of significant importance in the clinical settings. For this, we also released a dataset consisting sessile or flat polyps publicly. Furthermore, we have emphasized the use of cross-dataset evaluation by training and testing the model with images coming from various sources to achieve the generalizability goal.

In summary, the main contributions are as follows:

1) We have extended the ResUNet++ deep-learning architecture [1] for automatic polyp segmentation with CRF and TTA to achieve better performance. The quantitative and qualitative results shows that applying CRF and TTA is effective.
2) We validate the extended architecture on a large range of datasets, i.e., Kvasir-SEG [4], CVC-ClinicDB [11], CVC-ColonDB [12], EITS-Larib [13], ASU-Mayo Clinic Colonoscopy Video Database [14] and CVC-VideoClinicDB [15], [16], and we compare our proposed approaches with the recent State-of-the-art (SOTA) algorithm and set new a baseline. Moreover, we have compared our work with other recent works, which is often lacking in comparable studies.
3) We selected 196 flat or sessile polyps that are usually missed during colonoscopy examination [7] from the Kvasir-SEG with the help of an expert gastroenterologist. We have conducted experiments on this separate dataset to show how well our model performs on challenging polyps. Moreover, we release these polyp images and segmentation masks as a part of the Kvasir-SEG dataset so that researchers can build novel architectures and improve the results.
4) Our model has better detection of smaller and flat or sessile polyps, which are frequently missed during colonoscopy [7], which is a major strength compared to existing works.
5) In medical clinical practice, generalizable models are essential to target patient population. Our work is focused on generalizability, previously not much explored in the community. To promote generalizable Deep Learning (DL) models, we have trained our models on Kvasir-SEG and CVC-ClinicDB and tested and compared the results over five publicly available diverse unseen polyp dataset. Moreover, we have mixed two diverse datasets and conducted further experiments on other unseen datasets to show the behaviour of the model on the images captured using different devices.

II. Related Work

Over the past decades, researchers have made several efforts at developing CADx prototypes for automated polyp segmentation. Most of the prior polyp segmentation approaches were based on analyzing either the polyp's edge or its texture. More recent approaches used Convolutional Neural Network (CNN) and pre-trained networks. Bernal et al. [11] introduced a novel method for polyp localization that used WM-DOVA energy maps for accurately highlighting the polyps, irrespective of its type and size. Pozdeev et al. [17] presented a fully automated

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3049304, IEEE Journal of Biomedical and Health Informatics

polyp segmentation framework using pixel-wise prediction based upon the Fully Convolutional Network (FCN). Bernal et al. [18] hosted the automatic polyp detection in colonoscopy videos sub-challenge, and later on, they presented a comparative validation of different methods for automatic polyp detection and concluded that the SOTA CNN based methods provide the most promising results.

Akbari et al. [19] used the FCN-8S network and Otsu's thresholding method for automated colon polyp segmentation. Wang et al. [20] used the SegNet [21] architecture to detect polyps. They obtained high sensitivity, specificity, and receiver operating characteristic (ROC) curve value. Their algorithm could achieve a speed of 25 frames per second with some latency during real-time video analysis. Guo et al. [22] used a Fully Convolutional Neural Network (FCNN) model for the Gastrointestinal Image ANAlysis (GIANA) polyp segmentation challenge. The proposed method won first place in the 2017 GIANA challenge for both standard definition (SD) and high definition image and won second place in the SD image segmentation task in the 2018 GIANA challenge. Yamada et al. [23] developed a CADx support system that can be used for the real-time detection of polyps reducing the number of missed abnormalities during colonoscopy.

Poorneshwaran et al. [24] used a Generative Adversarial Network (GAN) for polyp image segmentation. Kang et al. [25] used Mask R-CNN, which relies on ResNet50 and ResNet101, as a backbone structure for automatic polyp detection and segmentation. Ali et al. [26] presented various detection and segmentation methods that could classify, segment, and localize artifacts. Additionally, there are several recent really interesting studies on polyp segmentation [27]–[30]. They are useful steps toward building an automated polyp segmentation system. There are also some works which have hypothesized that coupling the existing architecture by applying careful post-processing technique could improve the model performance [1], [31].

From the presented related work, we observe that automatic CADx systems in the area of polyp segmentation are becoming mature. Researchers are conducting a variety of studies with different designs ranging from a retrospective study, prospective study, to post hoc examination of the prospectively obtained dataset. Some of the models achieve very high performance with smaller training and test datasets [1], [20], [32]. The algorithms used for building the models are the ones that use handcrafted-, CNN- or pre-trained-features from ImageNet [33], where DL based algorithms are outperforming and gradually replacing the traditional handcrafted or machine learning (ML) approaches. Additionally, the performance of the models improves by the use of advance DL algorithms, especially designed for polyp segmentation task or any other similar biomedical image segmentation task. Moreover, there is interest for testing the proposed architectures with more than one dataset [1], [20].

The main drawbacks in the field are the minimal effort applied towards testing the generalizability of the CADx system possible to achieve with the cross-dataset test. Additionally, there is almost no effort involved in designing an universal model that could accurately segment polyp coming from different sources, critical for the development of CADx for automated polyp segmentation. Besides, most of the current works have proposed algorithms that are tested on single, often small, imbalanced, and explicitly handpicked datasets. This renders conclusions regarding the performance of the algorithms almost useless (compared to other areas in ML like, for example, natural image classification or action recognition where the common practice is to test on more than one dataset and make source code and datasets publicly available). Additionally, the used datasets are often not public available (restricted and difficult to access), and the total number of images and videos used in the study are not sufficient to believe that the system is robust and generalizable for use in clinical trials. For instance, the model can produce output segmentation map with high sensitivity and precision on a particular dataset and completely fails on other modality images. Moreover, existing work often use small training and test datasets. These current limitations make it harder to develop a robust and generalizable systems.

Therefore, we aim to develop a CADx based support system that could achieve high performances irrespective of the datasets. To achieve the goal, we have done extensive experiments on various colonoscopy images and video datasets. Additionally, we have mixed the dataset from multiple centers and tested it on other diverse unseen datasets to achieve the goal of building a generalizable and robust CADx system that produces no segmentation errors. Moreover, we set a new benchmark for the publicly available datasets that can be improved in the future.

III. The ResUNet++ Architecture

ResUNet++ is a semantic segmentation deep neural network designed for medical image segmentation. The backbone for ResUNet++ architecture is ResUNet [34]: an encoder-decoder network and based on U-Net [35]. The proposed architecture takes the benefit of residual block, squeeze and excite block [36], atrous spatial pyramid pooling (ASPP) [37], and attention block [38]. What distinguishes ResUNet++ from ResUNet is the use of squeeze-and-excitation blocks (marked in dark gray) at the encoder, the ASPP block, (marked in the dark red) at bridge and decoder, and the attention block (marked in light green) at the decoder (see Figure 2).

In the ResUNet++ model, we introduce the sequence of squeeze and excitation block to the encoder part of the network. Additionally, we replace the bridge of ResUNet with ASPP. In the decoder stage, we introduce a sequence of attention block, nearest-neighbor up-sampling, and concatenate it with the relevant feature map from the residual block of the encoder through skip connection. This process is followed by the residual unit with identity mapping, as shown in Figure 2.

We also introduce a series of additional skip connections from the residual unit of the encoder section to the attention block of the decoder section. We assign the number of filters $[32, 64, 128, 256, 512]$, along with the levels in the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3049304, IEEE Journal of Biomedical and Health Informatics

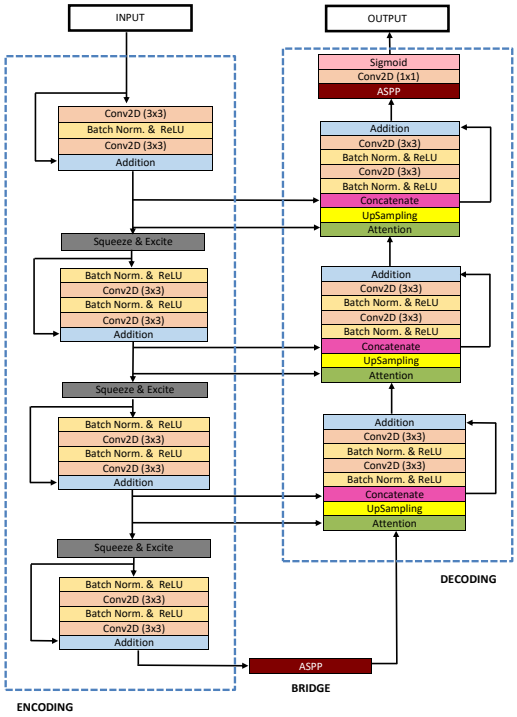


Fig. 2. ResUNet++ architecture [1]

encoder section, which are the values in our ResUNet++ architecture. These filter combinations achieved the best results in our ResUNet++ experiment. In the decoder section, the number of the filter is reversed, and the sequence becomes $[512, 256, 128, 64, 32]$. As the semantic gap between the feature map of the encoder and decoder blocks are supposed to decrease, the number of filters in the convolution layers of the decoder block are also decreased to achieve better semantic coverage. Through this, we ensure that the overall quality of the feature maps is more alike to the ground truth mask. This is especially important as the loss in semantic space is likely to decrease, and therefore it will become more feasible to find a meaningful representation in semantic space.

The overall ResUNet++ architecture consists of one stem block with three encoder blocks, an ASPP between the encoder and the decoder, and three decoder blocks. All the encoder and decoder blocks use the standard residual learning approach. Skip connections are introduced between encoder and decoder for the propagation of information. The output of the last decoder block is passed through the ASPP, followed by a 1×1 convolution and a sigmoid activation function. All convolutional layers except for the output layer are batch normalized [39] and are activated by a Rectified Linear Unit (ReLU) activation function [40]. Finally, we get the output as binary segmentation maps. A brief explanation of each block is provided in the following sub-sections.

A. Residual Blocks

Training a deep neural network by expanding network depth can potentially improve overall performance. Nevertheless, simply stacking the CNN layer could also hamper the training process and cause exploding/vanishing gradient when back-propagation occurs [41]. Residual connections facilitate the training process by directly routing the input information to the output and preserves the nobility of the gradient flow. The residual function simplifies the objective of optimization without any additional parameters and boosts the performance, which is the inspiration behind the deeper residual-based network [42]. Equation (1) below shows the working principle.

$$y_n = F(x_n, W_n) + x_n \tag{1}$$

Here, x_n is the input and $F(\cdot)$ is the residual function. The residual units consist of numerous combinations of Batch Normalization (BN), ReLU, and convolution layers. A detailed description of the combinations used and their impact can be found in the work of He et al. [43]. We have employed the concept of a pre-activation residual unit in the ResUNet++ architecture from ResUNet.

B. Squeeze and Excitation block

The squeeze and excitation (SE) block is the building block for the CNN that re-calibrates channel-wise feature response by explicitly modeling interdependencies between the channels [36]. The SE block learns the channel weights through global spatial information that increases the sensitivity of the effective feature maps, whereas it suppresses the irrelevant feature maps [1]. The feature maps produced by the convolution have only access to the local information, meaning they have no access to the global information left by the local receptive field. To address this limitation, we perform a squeeze operation on the feature maps using the global average pooling to generate a global representation. We then use the global representation and perform sigmoid activation that helps us to learn a non-linear interaction between the channels, and capture the channel-wise dependencies. Here, the sigmoid activation output acts as a simple gating mechanism that ensures us to adaptively recalibrate the feature maps produced by the convolution. The adaptive recalibration or excitation operation explicitly models the interdependencies between the feature channels. The SE net has the capability of generalizing exceptionally well across various datasets [36]. In the ResUNet++ architecture, we have stacked the SE block together with the residual block for improving the performance of the network, increasing the effective generalization across different medical datasets.

C. Atrous Spatial Pyramidal Pooling

Since the introduction of Atrous convolution by Chen et al. [44] to control the field-of-view to capture contextual information at multi-scale precisely, it has shown promising results for semantic image segmentation. Later, Chen et al. [45] proposed ASPP, which is a parallel atrous convolution block to capture multiple-scale information simultaneously. ASPP captures the contextual information at different scales, and multiple parallel atrous convolutions with varying rates in the input feature map are fused [45]. In ResUNet++, we use ASPP

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3049304, IEEE Journal of Biomedical and Health Informatics

as a bridge between the encoder and the decoder sections, and after the final decoder block. We adopt ASPP in ResUNet++ to capture the useful multi-scale information between the encoder and the decoder.

D. Attention Units

Chen et al. [46] proposed an attention model that can segment natural images by multi-scale input processing. Attention model is an improvement over average and max-pooling baseline and allows to visualize the features importance at different scales and positions [46]. With the success of attention mechanisms, various medical image segmentation methods have integrated an attention mechanism into their architecture [1], [47]–[49]. The attention block gives importance to the subset of the network to highlight the most relevant information. We believe that the attention mechanism in our architecture will boost the effectiveness of the feature maps of the network by capturing the relevant semantic class and filtering out irrelevant information. Motivated by the recent achievement of attention mechanism in the field of medical image segmentation and computer vision in general, we have integrated an attention block at the decoder part of the ResUNet++ model.

E. Conditional Random Field

Conditional Random Field (CRF) is a popular statistical modeling method used when the class labels for different inputs are not independent (e.g., image segmentation tasks). CRF can model useful geometric characteristics like shape, region connectivity, and contextual information [50]. Therefore, the use of CRF can further improve the models capability to capture contextual information of the polyps and thus improve overall results. We have used CRF as a further step to produce more refined output to the test dataset for improving the segmentation results. we have used a dense CRF for our experiment.

F. Test Time Augmentation

Test-Time Augmentation (TTA) is a technique of performing reasonable modifications to the test dataset to improve the overall prediction performance. In TTA, augmentation is applied to each test image, and multiple augmented images are created. After that, we make predictions on these augmented images, and the average prediction of each augmented image is taken as the final output prediction. Inspired by the improvement of recent SOTA [22], we have used TTA in our work. In this paper, we utilize both horizontal and vertical flip for TTA.

IV. Experiments

A. Datasets

We have used six different datasets of segmented polyps with ground truths in our experiments as shown in Table I, i.e., Kvasir-SEG [4], CVC-ClinicDB [11], CVC-ColonDB [12], ETIS Larib Polyp DB [13], CVC-VideoClinicDB [15], [16] and ASU-Mayo Clinic dataset [14]. They vary e.g., regarding number of images, image resolution, availability, devices used for capturing and the accuracy of the segmentation masks. One example is given from the Kvasir-SEG in Figure 3. The Kvasir-SEG dataset includes 196 polyps smaller than 10 mm classified as Paris class 1 sessile or Paris class IIa. We have released this dataset seperately as subset of Kvasir-SEG. Note that for CVC-VideoClinicDB, we have only used the data from the CVC-VideoClinicDBtrainvalid folder since only these data have ground truth masks. Moreover, the ASU-Mayo Clinic dataset, which was made available at the "Automatic Polyp Detection in Colonoscopy Videos" sub-challenge at Endovis 2015 had ten normal videos (negative shots) and ten videos with polyps. However, the test subset is not available because of issues related to licensing. In our experiment, while training, validating, testing with 80:10:10 split on the ASU-Mayo, we used all 20 videos for experimentation. However, for the cross-dataset test (i.e., Tables X and XI), we only tested on ten positive polyp videos.

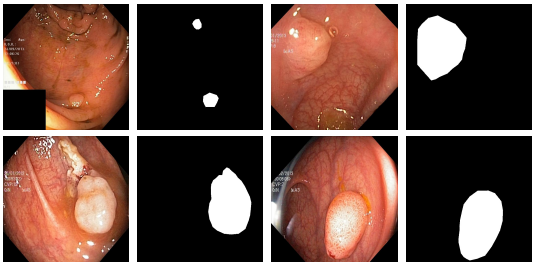


Fig. 3. Example polyp and corresponding ground truth from the Kvasir-SEG

TABLE I
THE BIOMEDICAL SEGMENTATION DATASETS USED IN OUR EXPERIMENTS

Dataset	Images	Input size	Availability
Kvasir-SEG [4]	1000	Variable	Public
CVC-ClinicDB [11]	612	384×288	Public
CVC-ColonDB [12]	380	574×500	Public
ETIS Larib Polyp DB [13]	196	1225×966	Public
CVC-VideoClinicDB [15], [16]$^{\dagger\diamond}$	11,954	384×288	Public
ASU-Mayo Clinic dataset [14]†	18,781	688×550	Copyrighted
Kvasir-Sessile$^{\bullet}$	196	Variable	Public

† Ground truth for test data not available $^{\diamond}$Ground truth oval or circle shaped
$^{\bullet}$ Part of Kvasir-SEG [4], only sessile polyps

B. Evaluation Method

To evaluate polyp segmentation methods, where individual pixels should be identified and marked, we use metrics used in earlier research [4], [18], [20], [22], [26], [51] and in competitions like GIANA[1], comparing the correctly and wrongly identified pixels of findings. The Dice coefficient (DSC) and the Intersection over Union (IoU) are the most commonly used metrics. We use the DSC to compare the similarity between the produced segmentation results and the original ground truth. Similarly, the IoU is used to compare the overlap between the output mask and original ground truth mask of the polyp. The mean Intersection over Union (mIoU) calculates IoU of each semantic class of the image and compute the mean over all the classes. There is a correlation between DSC and mIoU. However, we calculate both the metrics to

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3049304, IEEE Journal of Biomedical and Health Informatics

provide a comprehensive results analysis that could lead to better understanding of the results.

Moreover, other often-used metrics for the binary classification are recall (true positive rate) and precision (positive predictive value). For the polyp segmentation, precision is the ratio of the number of correctly segmented pixels versus the total number of all the pixels. Similarly, recall is the ratio of correctly segmented pixel versus the total number of pixels present in the ground truth. In the polyp image segmentation, precision and recall are used to indicate over-segmentation and under-segmentation. For formal definitions and formulas, see the definitions in for example [4], [51]. Finally, the receiver operating characteristic (ROC) curve analysis is also an important metric to characterize the performance of the binary classification system. In our study, we therefore calculate DSC, mIoU, recall, precision, and ROC when evaluating the segmentation models.

C. Data Augmentation

Data augmentation is a crucial step in increasing the number of polyp samples. This solves the data insufficiency problem, improves the performance of the model, and help to reduce over-fitting. We have used a large number of different data augmentation techniques to increase the training sample. We divide all the polyp datasets into training, validation, and testing sets using the ratio of 80:10:10 based on the random distribution except for the mixed datasets. After splitting the dataset, we apply data augmentation techniques such as center crop, random rotation, transpose, elastic transform, grid distortion, optical distortion, vertical flip, horizontal flip, grayscale, random brightness, random contrast, hue saturation value, RBG shift, course dropout, and different types of blur. For cropping the images, we have used a crop size of 256×256 pixels. For the experiment, we have resized the complete training, validation, and testing dataset to 256×256 pixels to reduce the computational complexity. We have only augmented the training dataset. The validation data is not augmented, and the test datasets were augmented while evaluation using TTA.

D. Implementation and Hardware Details

We have implemented all the models using the Keras framework [52] with Tensorflow [53] as a backend. Source code of our implementation and information about our experimental setup are made publicly available on Github[2]. Our experiments were performed using a Volta 100 Tensor Core GPU on a Nvidia DGX-2 AI system capable of 2-petaFLOPS tensor performance. We used a Ubuntu 18.04.3LTS operating system with Cuda 10.1.243 version installed. We have performed different experiments with different sets of hyperparameters manually on the same dataset in order to select the optimal set of hyperparameters for the ResUNet++. Our model performed well with the batch size of 16, Nadam as an optimizer, binary cross-entropy as the loss function, and learning rate of $1e{-}5$. The dice loss function was also competitive. These hyperparameters were chosen based on the empirical evaluation. All the models were trained for 300 epochs. We have used early stopping to prevent the model from over-fitting. To further improve the results, we have used stochastic gradient descent with warm restarts (SGDR). All the hyperparameters were same except the learning rate, which was adjusted based on the requirement. We have also included the Tensorboard for the analysis and visualization of the results.

[2]https://github.com/DebeshJha/ResUNet-with-CRF-and-TTA

TABLE II
RESULTS COMPARISON ON KVASIR-SEG

Method	DSC	mIoU	Recall	Precision
UNet [35]	0.7147	0.4334	0.6306	**0.9222**
ResUNet [34]	0.5144	0.4364	0.5041	0.7292
ResUNet-mod [34]	0.7909	0.4287	0.6909	0.8713
ResUNet++ [1]	0.8119	0.8068	0.8578	0.7742
ResUNet++ + CRF	0.8129	0.8080	0.8574	0.7775
ResUNet++ TTA	0.8496	0.8318	**0.8760**	0.8203
ResUNet++ +TTA + CRF	**0.8508**	**0.8329**	0.8756	0.8228

TABLE III
RESULTS COMPARISON ON CVC-CLINICDB

Method	DSC	mIoU	Recall	Precision
MultiResUNet$^{\diamond}$ [31]	-	0.8497	-	-
cGAN† [24]	0.8848	0.8127	-	-
SegNet [20]	-	-	0.8824	-
FCN$^{\bullet}$ [54]	-	-	0.7732	**0.8999**
CNN [55]	(0.62-0.87)	-	-	-
MSPB$^{\psi}$ CNN [56]	0.8130	-	0.7860	0.8090
UNet [35]	0.6419	0.4711	0.6756	0.6868
ResUNet [34]	0.4510	0.4570	0.5775	0.5614
PraNet [57]	0.8980	0.8400	-	-
ResUNet-mod [34]	0.7788	0.4545	0.6683	0.8877
ResUNet++ [1]	0.9199	0.8892	0.9391	0.8445
ResUNet++ + CRF	**0.9203**	**0.8898**	**0.9393**	0.8459
ResUNet++ + TTA	0.9020	0.8826	0.9065	0.8539
ResUNet++ + TTA + CRF	0.9017	0.8828	0.9060	0.8549

† Conditional generative adversarial network $^{\diamond}$Data augmentation
$^{\bullet}$Fully convolutional network $^{\psi}$ multi-scale patch-based

V. RESULTS

In our previous work, we have showed that ResUNet++ outperforms the SOTA UNet [35] and ResUNet [34] models trained on Kvasir-SEG and CVC-ClinicDB dataset [1]. In this work, we aim to improve the results of ResUNet++ by utilizing further hyperparameter optimization, CRF and TTA. In this section, we present and compare the results of ResUNet++ with CRF, TTA, and both approaches combined on the same dataset, mixed dataset, and cross-dataset. Although a direct comparison of approaches from the literature is difficult due to different testing mechanisms used by various authors, we nonetheless compare the results with the recent work for the evaluation.

A. Results comparison on Kvasir-SEG dataset

Table II and Figure 4 show the quantitative and qualitative results comparison. Figure 7 shows the ROC curve for all the models. As seen in the quantitative results (Table II), qualitative results (Figure 4), and ROC curve (Figure 7), our proposed methods outperform ResUNet++ on the Kvasir-SEG dataset. The improvement in results demonstrates the advantage of the use of the TTA, CRF and their combinations.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3049304, IEEE Journal of Biomedical and Health Informatics

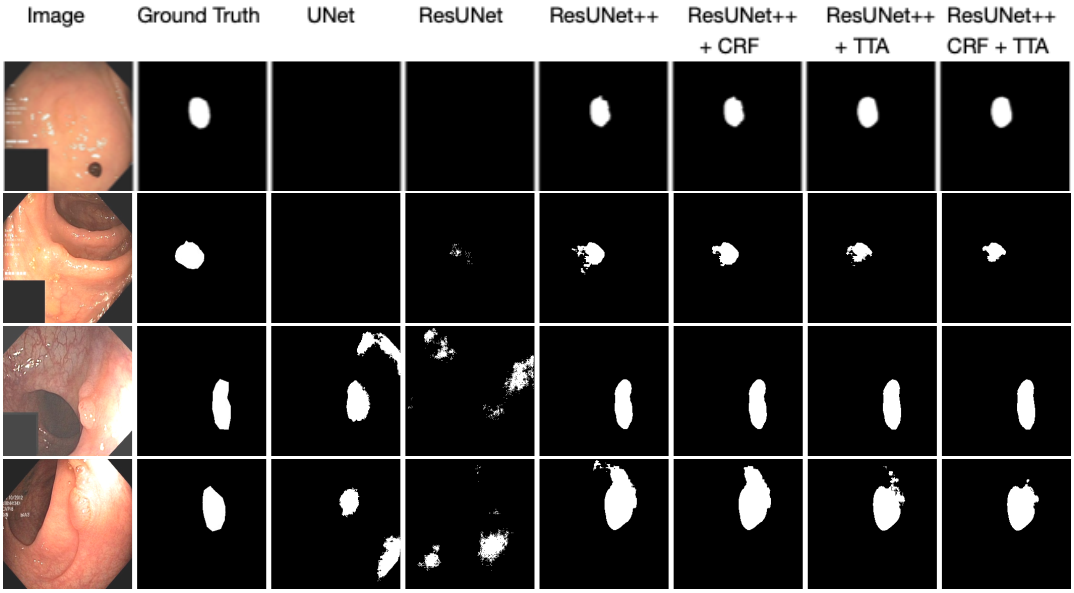


Fig. 4. Qualitative results comparison of the proposed models with UNet, ResUNet, and ResUNet++. The figure shows the example of polyps that are usually missed-out during colonoscopy examination. We see that there is a high similarity between ground truth and predicted mask for the proposed models.

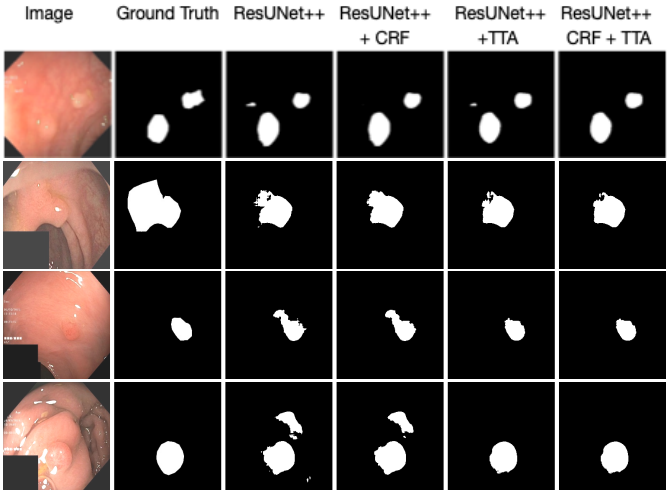


Fig. 5. Result of model trained on CVC-ClinicDB and tested on Kvasir-SEG

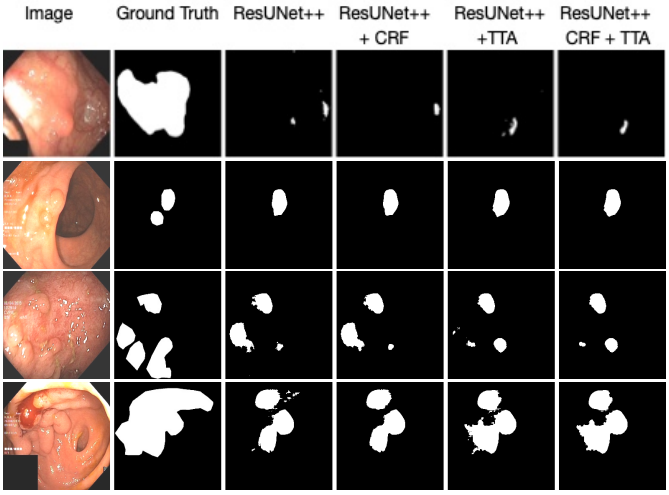


Fig. 6. Example images where the proposed models fails on Kvasir-SEG

B. Results comparison on CVC-ClinicDB

CVC-ClinicDB is a commonly used dataset for polyp segmentation. Therefore, it becomes important that we bring different work from the literature together and compare the proposed algorithms with the existing works. We compare our algorithms with the SOTA algorithms. Table III demonstrates that the combination of ResUNet++ and CRF achieves DSC of 0.9293 and mIoU of 0.8898, which is 2.23% improvement on PraNet [57] in DSC and 4.98% improvement in mIoU, respectively, and the proposed methods shows the SOTA result on CVC-ClinicDB.

The ROC curve measures the performance for the classification problem provided a set threshold. We have set the probability threshold of 0.5. The combination of ResUNet++ and TTA has the maximum Area Under Curve - Receiver Operating Characteristic (AUC-ROC) of 0.9814, as shown in Figure 8. Therefore, the results in Table III and Figure 8 show that applying TTA gives an improvement on CVC-ClinicDB.

C. Results comparison on CVC-ColonDB dataset

Our results using the CVC-ColonDB dataset are presented in Table IV. The table shows that proposed method of combining ResUNet++ and TTA achieved the highest DSC of 0.8474, which is 3.74% higher than SOTA [19], and mIoU of 0.8466 which is 20.66% higher than [57]. The recall and precision of all three proposed methods are quite acceptable. When compared with ResUNet++, there is an improvement of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3049304, IEEE Journal of Biomedical and Health Informatics

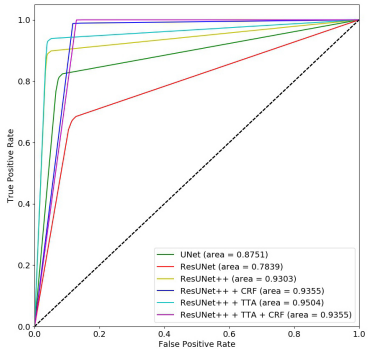


Fig. 7. ROC curve of proposed models on the Kvasir-SEG

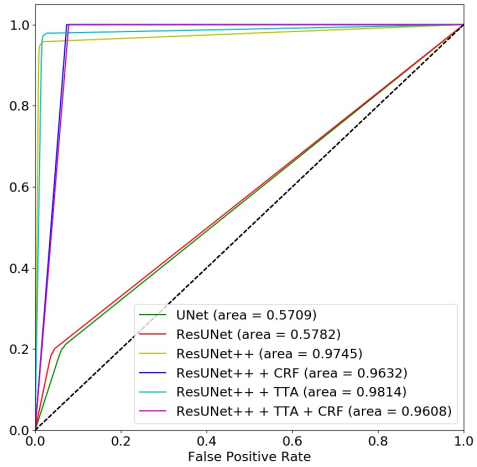


Fig. 8. ROC curve for all the models trained and tested on CVC-ClinicDB

TABLE IV
RESULTS COMPARISON ON CVC-COLONDB

Method	DSC	mIoU	Recall	Precision
FCN-8S + Otsu [19]	0.8100	-	0.7480	-
FCN-8s + Texton [58]	0.7014	-	0.7566	-
SA-DOVA Descriptor [12]	0.5533	-	0.6191	-
PraNet [57]	0.7090	0.6400	-	-
ResUNet++ [1]	0.8469	0.8456	**0.8511**	0.8003
ResUNet++ + CRF	0.8458	0.8456	0.8497	0.7767
ResUNet++ + TTA	**0.8474**	**0.8466**	0.8434	0.8118
ResUNet++ + TTA + CRF	0.8452	0.8459	0.8411	**0.8125**

TABLE V
RESULTS ON ETIS-LARIB POLYP DB

Method	DSC	mIoU	Recall	Precision
PraNet [57]	0.6280	0.5670	-	-
ResUNet++ [1]	**0.6364**	**0.7534**	**0.6346**	0.6467
ResUNet++ + CRF	0.6228	0.7520	0.6242	0.5648
ResUNet++ + TTA	0.6136	0.7458	0.5996	**0.6565**
ResUNet++ + TTA + CRF	0.6018	0.7426	0.5914	0.5755

1.22% in precision. There are negligible differences in recall, with ResUNet++ slightly outperforming the others.

D. Results comparison on ETIS-Larib Polyp DB

Table V shows the results of the proposed models on the ETIS-Larib Polyp DB. In this case, we do not compare the results with UNet and ResUNet, but compare the models directly with ResUNet++ as it already showed superior performance on Kvasir-SEG and CVC-ClinicDB [1]. Here, there are only marginal differences in the results of ResUNet++, "ResUNet++ + CRF", "ResUNet++ + TTA", and "ResUNet++ + CRF + TTA". However, ResUNet++ achieves maximum DSC of 0.6364, which is 0.84% improvement over SOTA [57] and mIoU of 0.7534 which is 18.64% improvement over [57]. The recall of ResUNet++ is 0.6346, which is slightly higher than the proposed methods. However, the precision of combining ResUNet++ and TTA is higher as compared to ResUNet++.

From the results, we can say that the performance of architecture is data specific. Our proposed methods outperformed SOTA over five independent datasets, however, ResUNet++ shows better results than the combinational approaches on ETIS-Larib dataset. Still, the precision of combining ResUNet++ and TTA is slightly higher than ResUNet++. It is to be noted that ETIS-Larib contains only 196 images, out of which only 156 images are used for training. Even with the small training dataset, the models are performing satisfactory as compared to the SOTA [57] with significant margin in mIoU, which can be considered as the strength of the algorithm.

TABLE VI
RESULTS ON KVASIR-SESSILE

Method	DSC	mIoU	Recall	Precision
ResUNet++ [1]	0.4600	0.64086	0.4382	0.5838
ResUNet++ + CRF	0.4522	0.6394	0.4326	0.5708
ResUNet++ + TTA	**0.5042**	**0.6606**	**0.4851**	**0.6796**
ResUNet++ + TTA + CRF	0.4901	0.6565	0.4766	0.6277

TABLE VII
RESULTS COMPARISON ON CVC-VIDEOCLINICDB

Method	DSC	mIoU	Recall	Precision
ResUNet++ [1]	0.8798	0.8730	**0.7749**	0.6702
ResUNet++ + CRF	**0.8811**	**0.8739**	0.7743	**0.6706**
ResUNet++ + TTA	0.8125	0.8467	0.6896	0.6421
ResUNet++ + TTA + CRF	0.8130	0.8477	0.6875	0.6276

E. Results on Kvasir-Sessile

As this is the first work on Kvasir-Sessile, we have compared the proposed methods with ResUNet++. Table VI shows that combining ResUNet++ and TTA gives the DSC of 0.5042, mIoU of 0.6606 which can be considered a decent score on a smaller size dataset. The dataset contains small, diverse images, which are difficult to generalize with very few training samples.

F. Results comparison on CVC-VideoClinicDB

Table VII shows the results of the proposed models on the CVC-VideoClinicDB. From the results, we can see that all models perform well on the dataset despite the fact that masks are not pixel perfect. One of the reasons for high performance is the presence of $11,954$ polyps and normal video frames that was used in training and testing. The combination of ResUNet++ and CRF obtained a DSC of 0.8811, mIoU of 0.8739, recall of 0.7743, and precision of 0.6706 which is quite acceptable for the segmentation task with this type of dataset. In CVC-VideoClinicDB, the ground-truth is marked

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3049304, IEEE Journal of Biomedical and Health Informatics

TABLE VIII
RESULTS COMPARISON ON ASUMAYO CLINIC

Method	DSC	mIoU	Recall	Precision
ResUNet++ [1]	0.8743	0.8569	**0.6534**	0.4896
ResUNet++ + CRF	**0.8850**	**0.8635**	0.6504	0.4858
ResUNet++ + TTA	0.8553	0.8535	0.6162	**0.4912**
ResUNet++ + TTA + CRF	0.8550	0.8551	0.6107	0.4743

TABLE IX
RESULTS COMPARISON USING (KVASIR-SEG + CVC-CLINICDB) AS THE TRAINING SET

Test set	Method	DSC	mIoU	Recall	Precision
CVC-ColonDB	ResUNet++ [1]	0.4974	0.6800	0.4787	**0.6019**
	ResUNet++ + CRF	0.4920	0.6788	0.4744	0.5636
	ResUNet++ + TTA	**0.5084**	**0.6859**	**0.4795**	0.5973
	ResUNet++ + TTA + CRF	0.5061	0.6852	0.4775	0.5770
CVC-Video-ClinicDB	ResUNet++ [1]	0.3460	0.6348	**0.2272**	**0.3383**
	ResUNet++ + CRF	0.3552	0.6412	0.2228	0.3065
	ResUNet++ + TTA	0.3573	0.6440	0.2104	0.3338
	ResUNet++ + TTA + CRF	**0.3603**	**0.6468**	0.2068	0.3038

with a oval or circle shape. However, it is understandable that pixel-precise annotations of this dataset will need great manual effort from expert endoscopists and engineers.

G. Results comparison on AUS-Mayo ClinicDB

Table VIII shows the results of the proposed models on the ASU-Mayo ClinicDB. ASU-Mayo contains 18,781 frames, both polyp and non-polyp images. The combination of ResUNet++ and CRF obtained a DSC of 0.8850 and mIoU of 0.8635. As in the real clinical settings, the models trained on this type of dataset are more meaningful (as it contains both polyp and non-polyp frames). The capability to achieve good performance for these more challenging datasets is one of the strengths of the proposed method. This is supported by the fact that this dataset also contains a sufficient amount of images to enable sufficient training.

H. Results comparison on mixed dataset

To check the performance of the proposed approaches on the images captured using different devices, we have mixed the Kvasir-SEG and CVC-ClinicDB and used them for training. The model were tested on CVC-ColonDB and CVC-VideoClinicDB. Table IX shows the result of the mixed dataset on both datasets. The combination of ResUNet++ and TTA obtains a DSC of 0.5084 and mIoU of 0.6859 with CVC-ColonDB. The combination of ResUNet++, CRF, and TTA obtained a DSC of 0.3603 and mIoU of 0.6468 with CVC-VideoClinicDB.

From the table, we can see that the combination of ResUNet++, CRF, and TTA performs better or very competitive in both still images and video frames. Here, it is also evident that the model trained on the smaller dataset (Kvasir-SEG and CVC-ClinicDB) which do not include non-polyp images is not performing well on larger and diverse datasets (CVC-VideoClinicDB) that contain both polyp and non-polyp frames. Additionally, for the CVC-VideoClinicDB datasets, the provided ground truth is not perfect (oval/circle) shaped. As the model trained on Kvasir-SEG and CVC-ClinicDB have perfect annotations, the model is good at predicting a perfect shaped mask. When we make predictions on the CVC-VideoClinicDB with imperfect masks, even if the predictions are good, the scores may not be high because of the difference in the provided ground truth and the predicted masks.

TABLE X
CROSS-DATASET RESULTS USING **KVASIR-SEG** AS THE TRAINING SET

Test set	Method	DSC	mIoU	Recall	Precision
CVC-ClinicDB	ResUNet++ [1]	0.6468	0.7311	0.6984	0.6510
	ResUNet++ + CRF	0.6458	0.7321	0.6955	0.6425
	ResUNet++ + TTA	**0.6737**	**0.7507**	**0.7108**	**0.6833**
	ResUNet++ + TTA + CRF	0.6712	0.7506	0.7078	0.6680
ETIS-Larib Polyp DB	ResUNet++ [1]	**0.4017**	0.6415	**0.4412**	0.3925
	ResUNet++ + CRF	0.4012	0.6427	0.4379	0.3755
	ResUNet++ + TTA	0.4014	**0.6468**	0.4294	**0.4014**
	ResUNet++ + TTA + CRF	0.3997	0.6466	0.4267	0.3710
CVC-ColonDB	ResUNet++ [1]	0.5135	0.6742	0.5398	0.5461
	ResUNet++ + CRF	0.5122	0.6748	0.5367	0.5285
	ResUNet++ + TTA	**0.5593**	**0.7030**	**0.5626**	**0.5944**
	ResUNet++ + TTA + CRF	0.5563	0.7024	0.5595	0.5811
CVC-Video-ClinicDB	ResUNet++ [1]	0.3175	0.6082	**0.2915**	0.3299
	ResUNet++ + CRF	0.3334	0.6185	0.2862	0.3141
	ResUNet++ + TTA	0.3505	0.6337	0.2601	**0.3488**
	ResUNet++ + TTA + CRF	**0.3601**	**0.6402**	0.2555	0.3252
ASU-Mayo	ResUNet++ [1]	0.3482	0.6346	**0.2196**	**0.2021**
	ResUNet++ + CRF	0.3747	0.6516	0.2136	0.1797
	ResUNet++ + TTA	0.3823	0.6583	0.1962	0.2165
	ResUNet++ + TTA + CRF	**0.3950**	**0.6681**	0.1890	0.1781

I. Cross-dataset result evaluation on Kvasir-SEG

For the cross-dataset evaluation, we trained the models on the Kvasir-SEG dataset and tested it on the other five independent datasets. Table X shows the results of cross-data generalizability of ResUNet++ alone, and with the CRF and TTA techniques. The results of the models trained on Kvasir-SEG produces an average best mIoU of 0.6817 and an average best DSC of 0.4779 for both image and video datasets. From the above table, we can see that the proposed combinational approaches are performing competitive. For the image datasets, the combination of ResUNet++ and TTA is performing better, and for the video datasets, the combination of ResUNet++, CRF, and TTA is performing best. It is to be noted that we are training a model with 1000 Kvasir-SEG pixel segmented polyps and testing on (for example, 11,954 frames) oval-shaped polyp ground truth. Here, even if the predictions are correct, the evaluation scores will not be good because of the oval/circle shaped ground truth. Moreover, the datasets such as ASU-Mayo and CVC-VideoClinicDB are heavily imbalanced, but the model trained on Kvasir-SEG contains at least one polyp. This may also have caused the poor performance.

J. Cross-dataset evaluation on CVC-ClinicDB

To further test generaliziblity, we trained the models on CVC-CliniDB and tested it across five independent, diverse image and video datasets. Tables XI shows the results of cross-data generalizability. Like the previous test on Kvasir-SEG, the results follow the same pattern with the combination of ResUNet++ and TTA outperforming others on the image datasets and the combination of ResUNet++, CRF, and TTA outperforming its competitors on video datasets. ResUNet++

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3049304, IEEE Journal of Biomedical and Health Informatics

TABLE XI
CROSS-DATASET RESULTS ON **CVC-CLINICDB** AS THE TRAINING SET

Test set	Method	DSC	mIoU	Recall	Precision
Kvasir-SEG	ResUNet++ [1]	0.6876	0.7374	0.7027	0.7354
	ResUNet++ + CRF	0.6877	0.7389	0.7004	0.7371
	ResUNet++ + TTA	**0.7218**	0.7616	**0.7225**	**0.7855**
	ResUNet++ + TTA + CRF	0.7208	**0.7621**	0.7204	0.7831
CVC-ColonDB	ResUNet++ [1]	0.5489	0.6942	0.5577	0.5816
	ResUNet++ + CRF	0.5470	0.6949	0.5546	0.5727
	ResUNet++ + TTA	**0.5686**	0.7080	**0.5702**	**0.5935**
	ResUNet++ + + TTA + CRF	0.5667	**0.7081**	0.5687	0.5773
ETIS-Larib Polyp DB	FCN-VGG [59]	0.7023	0.5420	-	-
	ResUNet++ [1]	0.4012	0.6398	**0.4232**	0.4013
	ResUNet++ + CRF	0.3990	0.6403	0.4191	0.3974
	ResUNet++ + TTA	0.4027	**0.6522**	0.3969	**0.4235**
	ResUNet++ + TTA + CRF	0.3973	0.6514	0.3906	0.4078
CVC-Video-ClinicDB	ResUNet++ [1]	0.3666	0.6422	**0.2568**	0.3632
	ResUNet++ + CRF	0.3788	0.6500	0.2530	0.3399
	ResUNet++ + TTA	0.3941	0.6582	0.2516	**0.3829**
	ResUNet++ + TTA + CRF	**0.3988**	**0.6616**	0.2481	0.3542
ASU-Mayo	ResUNet++ [1]	0.2797	0.6113	**0.1627**	0.1443
	ResUNet++ + CRF	0.3167	0.6323	0.1591	0.1348
	ResUNet++ + TTA	0.3085	0.6331	0.1265	**0.1571**
	ResUNet++ + TTA + CRF	**0.3233**	**0.6426**	0.1225	0.1270

and TTA still remain competitive. Moreover, the values of DSC and mIoU of the best model are similar for both the CVC-VideoClinicDB and the ASU-Mayo Clinic dataset. We have compared the results with the existing work that used CVC-CliniDB for training and ETIS-Larib for testing. Our model achieves highest mIoU of 0.6522.

K. Result summary

In summary, from all obtained results (i.e., qualitative, quantitative, and ROC curve), the following main observations can be drawn: (i) the proposed ResUNet++ is capable of segmenting the smaller, larger and regular polyps; (ii) the combination of ResUNet++ with CRF achieves the best performance in terms of DSC, mIoU, recall and precision when trained and tested on the same dataset (see Table III, Table VII, and Table VIII) whereas it remains competitive when tested on other datasets; (iii) the combination of ResUNet++ and TTA and the combination of ResUNet++, CRF and TTA performs similar for the mixed datasets; (iv) the combination of ResUNet++ and TTA outperforms others on still images; (v) the combination of ResUNet++, CRF and TTA shows improvement on all the video datasets compared to ResUNet++; (vi) all the models perform better when the images have higher contrast; (vii) ResUNet++ is particularly good at segmenting smaller and flat or sessile polyps, which is a prerequisite for developing an ideal CADx polyp detection system [1]; (viii) ResUNet++ fails especially on the images that contains overexposed regions termed as saturation or contrast (see Figure 6); (ix) ResUNet and ResUNet-mod particularly showed oversegmented or under-segmented results, (see Figure 4).

VI. DISCUSSION

A. General Performance

The tables and figures suggest that applying CRF and TTA improved the performance of ResUNet++ on the same datasets, mixed datasets and cross-datasets. Specifically, the combination of ResUNet++ and TTA, and the combination of ResUNet++, CRF and TTA are more generalizable for all the datasets, where TTA with ResUNet++ performs best on the still images, and the combinations of ResUNet++, CRF, and TTA are outperforming others on video datasets. For all of the proposed models, the value of AUC is greater than 0.93. This indicates that our models are good at distinguishing between the polyp and non-polyps. It also suggests that the model produces sufficient sensitivity.

TABLE XII
TOTAL NUMBER OF TRAINABLE PARAMETERS

Model	Trainable parameters
U-Net	5,400,289
ResUNet	8,221,121
ResUNet-mod	2,058,465
ResUNet++	16,228,001

The total number of trainable parameters increases by increasing the number of blocks in the networks (see Table XII). However, in ResUNet++, there is significant performance gain that compensates for the training time, and our model requires fewer parameters if we compare with the models that use pretrained encoders.

B. Cross Dataset Performance

The cross-data test is an excellent technique to determine the generalizing capability of a model. The presented work is an initiative towards improving the generalizability of segmentation methods. Our contribution towards generalizability is to train on one dataset and test on several other public datasets that may come from different centers and use different scope manufacturers. Thus, we believe that to tackle this issue, out-of-sample multicenter data must be used to test the built methods. The work is a step forward in raising an issue regarding method interpretability and we also raise questions about generalizability and domain adaptation of supervised methods in general.

From the results analyses, we can see that different proposed algorithms perform well with different types of datasets. For instance, CRF outperformed others on tables III, VII, and VIII. TTA showed improvement on tables IV, IX, X and XI. CRF performs better than TTA while trained and tested on video datasets (see tables VII and VIII). CRF also outperformed TTA on most of the images dataset. However, TTA still remains competitive. On the mixed dataset and the cross-dataset test, TTA performs better than CRF on all the datasets. On the mixed datasets and on the cross-dataset test on videos, the combination of ResUNet++, CRF, and TTA remains the best choice (see tables IX, X, and XI). There is a performance improvement over ResUNet++ while combining CRF, TTA, and the combination of CRF and TTA.

However, there is no significant performance improvement of any methods on the others. From the results, we can see that the results are typically data-dependent. However, as the proposed methods perform well on video frames, it may work better in the clinic, as the output from a colonoscope is a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3049304, IEEE Journal of Biomedical and Health Informatics

video stream. Thus, it becomes critical to show the results with all three approaches on each dataset. Therefore, we provide extensive experiments showing both success (Figure 4, Figure 5) and failure cases (Figure 6) and present the overall analysis.

C. Challenges

There are several challenges associated with segmenting polyps, such as bowel-quality preparation during colonoscopy, angle of the cameras, superfluous information, and varying morphology, which can affect the overall performance of a DL model. For some of the images, there even exists variation in the decision between endoscopists. While ResUNet++ with CRF and TTA also struggle with producing satisfactory segmentation maps for these images, it performs considerably better than our previous model and also outperforms another SOTA algorithm.

The quality of a colonoscopy examination is largely determined by the experience and skill of the endoscopist [23]. Our proposed model can help in two ways: (i) it can be used to segment a detected polyp, providing an extra pair of eyes to the endoscopist; and (ii) it performs well on both flat and small polyps, which are often missed during endoscopic examinations. The qualitative analysis (see Figure 4) and the quantitative analyses from the above tables and figures support this argument. This is a major strength of our work and makes it a candidate for clinical testing.

D. Possible Limitations

Possible limitations of this work are that it is a retrospective study. Prospective clinical evaluation is essential because data analyzed with the retrospective study is the different prospective study (for example, the case of missing data that should be considered on the basis of best-case and worse case scenarios) [60]. Also, all data in these experiments are curated, while a prospective clinical trial would mean testing on full colonscopy videos. During model training, we have resized all the images to 256×256 to reduce the complexity, which costs in loss of information, and can affect the overall performance. We have worked on optimizing the code, but further optimization may exist, that can potentially improve the performance of the model.

VII. Conclusion

In this paper, we have presented the ResUNet++ architecture for semantic polyp segmentation. We took inspiration from the residual block, ASPP, and attention block to design the novel ResUNet++ architecture. Furthermore, we applied CRF and TTA to improve the results even more. We have trained and validated the combination of ResUNet++ with CRF and TTA using six publicly available datasets, and analyzed and compared the results with the SOTA algorithm on specific datasets. Moreover, we analyzed the cross-data generalizability of the proposed model towards developing generalizable semantic segmentation models for automatic polyp segmentation. A comprehensive evaluation of the proposed model trained and tested on six different datasets showed good performance of the (ResUNet++ and CRF) on image datasets and (ResUNet++ and TTA), (ResUNet++, CRF, and TTA) model for the mixed datasets and cross-datasets. Further, a detailed study on cross-dataset generalizability of the models trained on Kvasir-SEG and CVC-ClinicDB and tested on five independent datasets, confirmed the robustness of the proposed ResUNet++ + TTA method for cross-dataset evaluation.

The strength of our method is that we successfully detected smaller and flat polyps, which are usually missed during colonoscopy examination [20], [61]. Our model can also detect the polyps that would be difficult for the endoscopists to identify without careful investigations. Therefore, we believe that the ResUNet++ architecture, along with the additional CRF and TTA steps, could be one of the potential areas to investigate, especially for the overlooked polyps. We also point out that the lack of generalization issues of the models, which is evidenced by the unsatisfactory result for cross-dataset evaluation in most of the cases. In the future, our CADx system should also be investigated on other bowel conditions. Moreover, a prospective trial should also be conducted with image and video datasets.

Acknowledgement

This work is funded in part by Research Council of Norway project number 263248. Experiments are performed on the Experimental Infrastructure for Exploration of Exascale Computing (eX3), supported by the Research Council of Norway under contract 270053.

References

[1] D. Jha *et al.*, "Resunet++: An advanced architecture for medical image segmentation," in *Proc. of IEEE ISM.*, 2019, pp. 225–230.

[2] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[3] T. Matsuda, A. Ono, M. Sekiguchi, T. Fujii, and Y. Saito, "Advances in image enhancement in colonoscopy for detection of adenomas," *Nat. Revi. Gastroenter. & Hepato.*, vol. 14, no. 5, pp. 305–314, 2017.

[4] D. Jha *et al.*, "Kvasir-seg: A segmented polyp dataset," in *Proc. of MMM*, 2020, pp. 451–462.

[5] S. B. Ahn, D. S. Han, J. H. Bae, T. J. Byun, J. P. Kim, and C. S. Eun, "The miss rate for colorectal adenoma determined by quality-adjusted, back-to-back colonoscopies," *Gut and liver*, vol. 6, no. 1, pp. 64–70, 2012.

[6] D. o. Heresbach, "Miss rate for colorectal neoplastic polyps: a prospective multicenter study of back-to-back video colonoscopies," *Endoscopy*, vol. 40, no. 04, pp. 284–290, 2008.

[7] Zimmermann-Fraedrich *et al.*, "Right-sided location not associated with missed colorectal adenomas in an individual-level reanalysis of tandem colonoscopy studies," *Gastroenterology*, vol. 157, no. 3, pp. 660–671, 2019.

[8] A. Shaukat *et al.*, "Longer withdrawal time is associated with a reduced incidence of interval cancer after screening colonoscopy," *Gastroenterology*, vol. 149, no. 4, pp. 952–957, 2015.

[9] D. Vázquez *et al.*, "A benchmark for endoluminal scene segmentation of colonoscopy images," *Journal of healthcare engineering*, vol. 2017, 2017.

[10] T. Roß *et al.*, "Robust medical instrument segmentation challenge 2019," *arXiv preprint arXiv:2003.10299v1*, 2020.

[11] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computeri. Med. Imag. and Graph.*, vol. 43, pp. 99–111, 2015.

[12] J. Bernal, J. Sánchez, and F. Vilarino, "Towards automatic polyp detection with a polyp appearance model," *Patt. Recognit.*, vol. 45, no. 9, pp. 3166–3182, 2012.

[13] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer," *Int. Jour. of Comput. Assis. Radiol. and Surg.*, vol. 9, no. 2, pp. 283–293, 2014.

[14] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Trans. Med. Imag.*, vol. 35, no. 2, pp. 630–644, 2015.

[15] Q. Angermann *et al.*, "Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis," in *Comput. Assis. and Robot. Endos. and Clin. Image-Based Proced.*, 2017, pp. 29–41.

[16] J. Bernal *et al.*, "Polyp detection benchmark in colonoscopy videos using gtcreator: A novel fully configurable tool for easy and fast annotation of image databases," in *Proceedings of CARS conference*, 2018.

[17] A. A. Pozdeev, N. A. Obukhova, and A. A. Motyko, "Automatic analysis of endoscopic images for polyps detection and segmentation," in *Proc. of EIConRus*, 2019, pp. 1216–1220.

[18] J. Bernal *et al.*, "Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge," *IEEE Trans. Med. Imag.*, vol. 36, no. 6, pp. 1231–1249, 2017.

[19] M. Akbari *et al.*, "Polyp segmentation in colonoscopy images using fully convolutional network," in *Proc. of EMBC*, 2018, pp. 69–72.

[20] P. Wang *et al.*, "Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy," *Nat. biomed. engineer.*, vol. 2, no. 10, pp. 741–748, 2018.

[21] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE trans. on patt. analys. and mach. intellige.*, vol. 39, no. 12, pp. 2481–2495, 2017.

[22] Y. B. Guo and B. Matuszewski, "Giana polyp segmentation with fully convolutional dilation neural networks," in *Proc. of VISIGRAPP*, 2019, pp. 632–641.

[23] M. Yamada *et al.*, "Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy," *Scienti. repo.*, vol. 9, no. 1, pp. 1–9, 2019.

[24] J. Poomeshwaran, K. S. Santhosh, K. Ram, J. Joseph, and M. Sivaprakasam, "Polyp segmentation using generative adversarial network," in *Proc. of EMBC*, 2019, pp. 7201–7204.

[25] J. Kang and J. Gwak, "Ensemble of instance segmentation models for polyp segmentation in colonoscopy images," *IEEE Access*, vol. 7, pp. 26 440–26 447, 2019.

[26] S. Ali *et al.*, "Endoscopy artifact detection (ead 2019) challenge dataset," *arXiv preprint arXiv:1905.03209*, 2019.

[27] N.-Q. Nguyen and S.-W. Lee, "Robust boundary segmentation in medical images using a consecutive deep encoder-decoder network," *IEEE Access*, vol. 7, pp. 33 795–33 808, 2019.

[28] V. de Almeida Thomaz, C. A. Sierra-Franco, and A. B. Raposo, "Training data enhancements for robust polyp segmentation in colonoscopy images," in *Proc. of CBMS*, 2019, pp. 192–197.

[29] X. Sun, P. Zhang, D. Wang, Y. Cao, and B. Liu, "Colorectal polyp segmentation by u-net with dilation convolution," *arXiv preprint arXiv:1912.11947*, 2019.

[30] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "Doubleu-net: A deep convolutional neural network for medical image segmentation," in *Proc. of IEEE CBMS*, 2020.

[31] N. Ibtehaz and M. S. Rahman, "Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, pp. 74–87, 2020.

[32] P. Brandao *et al.*, "Towards a computed-aided diagnosis system in colonoscopy: automatic polyp segmentation using convolution neural networks," *Jour. of Medi. Robot. Resear.*, vol. 3, no. 02, p. 1840002, 2018.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. of CVPR*, 2009, pp. 248–255.

[34] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geosci. and Remo. Sens. Lett.*, vol. 15, no. 5, pp. 749–753, 2018.

[35] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. of MICCAI*, 2015, pp. 234–241.

[36] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. of CVPR*, 2018, pp. 7132–7141.

[37] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of NIPS*, 2017, pp. 5998–6008.

[39] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[40] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of CVPR*, 2016, pp. 770–778.

[42] L. Wang, R. Chen, S. Wang, N. Zeng, X. Huang, and C. Liu, "Nested dilation network (ndn) for multi-task medical image segmentation," *IEEE Access*, vol. 7, pp. 44 676–44 685, 2019.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. of ECCV*, 2016, pp. 630–645.

[44] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[45] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE trans. on pattern anal. and mach. intelli.*, vol. 40, no. 4, pp. 834–848, 2017.

[46] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. of CVPR*, 2016, pp. 3640–3649.

[47] Y. Wang *et al.*, "Deep attentional features for prostate segmentation in ultrasound," in *Proc. of MICCAI*, 2018, pp. 523–530.

[48] D. Nie, Y. Gao, L. Wang, and D. Shen, "Asdnet: Attention based semi-supervised deep networks for medical image segmentation," in *Proc. of MICCAI*, 2018, pp. 370–378.

[49] A. Sinha and J. Dolz, "Multi-scale guided attention for medical image segmentation," *arXiv preprint arXiv:1906.02849*, 2019.

[50] F. I. Alam, J. Zhou, A. W.-C. Liew, X. Jia, J. Chanussot, and Y. Gao, "Conditional random field and deep feature learning for hyperspectral image classification," *IEEE Trans. on Geosci. and Remo. Sens.*, vol. 57, no. 3, pp. 1612–1628, 2018.

[51] K. Pogorelov *et al.*, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proc. of MMSYS*, 2017, pp. 164–169.

[52] F. Chollet *et al.*, "Keras," 2015.

[53] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proc. of OSDI*, 2016, pp. 265–283.

[54] Q. Li *et al.*, "Colorectal polyp segmentation using a fully convolutional neural network," in *Proc. of CISP-BMEI*, 2017, pp. 1–5.

[55] Q. Nguyen and S.-W. Lee, "Colorectal segmentation using multiple encoder-decoder network in colonoscopy images," in *Proc. of IKE*, 2018, pp. 208–211.

[56] D. Banik, D. Bhattacharjee, and M. Nasipuri, "A multi-scale patch-based deep learning system for polyp segmentation," in *Advan. Comput. and Syst. for Secur.*, 2020, pp. 109–119.

[57] D.-P. Fan *et al.*, "Pranet: Parallel reverse attention network for polyp segmentation," in *Proc. of MICCAI*, 2020, pp. 263–273.

[58] L. Zhang, S. Dolwani, and X. Ye, "Automated polyp segmentation in colonoscopy frames using fully convolutional neural network and textons," in *Proc. ov MIUA*, 2017, pp. 707–717.

[59] P. Brandao *et al.*, "Fully convolutional neural networks for polyp segmentation in colonoscopy," in *Medical Imaging 2017: Computer-Aided Diagnosis*, vol. 10134, 2017, pp. 101 340F1 – 101 340F1.

[60] Y. Mori and S.-e. Kudo, "Detecting colorectal polyps via machine learning," *Nat. biomed. engineer.*, vol. 2, no. 10, pp. 713–714, 2018.

[61] A. Leufkens, M. Van Oijen, F. Vleggaar, and P. Siersema, "Factors influencing the miss rate of polyps in a back-to-back colonoscopy study," *Endoscopy*, vol. 44, no. 05, pp. 470–475, 2012.

A.4 Paper IV: DoubleU-Net: A Deep Convolutional Neural Network for Medical Image Segmentation

Authors: D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen

Abstract: Semantic image segmentation is the process of labeling each pixel of an image with its corresponding class. An encoder-decoder based approach, like U-Net and its variants, is a popular strategy for solving medical image segmentation tasks. To improve the performance of U-Net on various segmentation tasks, we propose a novel architecture called DoubleU-Net, which is a combination of two U-Net architectures stacked on top of each other. The first U-Net uses a pre-trained VGG-19 as the encoder, which has already learned features from ImageNet and can be transferred to another task easily. To capture more semantic information efficiently, we added another U-Net at the bottom. We also adopt Atrous Spatial Pyramid Pooling (ASPP) to capture contextual information within the network. We have evaluated DoubleU-Net using four medical segmentation datasets, covering various imaging modalities such as colonoscopy, dermoscopy, and microscopy. Experiments on the MICCAI 2015 segmentation challenge, the CVC-ClinicDB, the 2018 Data Science Bowl challenge, and the Lesion boundary segmentation datasets demonstrate that the DoubleU-Net outperforms U-Net and the baseline models. Moreover, DoubleU-Net produces more accurate segmentation masks, especially in the case of the CVC-ClinicDB and MICCAI 2015 segmentation challenge datasets, which have challenging images such as smaller and flat polyps. These results show the improvement over the existing U-Net model. The encouraging results, produced on various medical image segmentation datasets, show that DoubleU-Net can be used as a strong baseline for both medical image segmentation and cross-dataset evaluation testing to measure the generalizability of Deep Learning (DL) models.

Published: Proceedings of IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), pp. 558-564, 2020.

Candidate contributions: D. Jha conceptualized this work and performed all the experiments and analyses in the paper. He wrote the manuscript, which was revised by all of the co-authors. Additionally, he presented this work at the conference.

A.4. Paper IV: DoubleU-Net: A Deep Convolutional Neural Network for Medical Image Segmentation

This paper was nominated for the best paper award at Computer-Based Medical Systems (CBMS 2020).

Thesis objectives: Objective III

A.5 Paper V: Real-Time Polyp Detection, Localization and Segmentation in Colonoscopy Using Deep Learning

Authors: D. Jha, S. Ali, N. K. Tomar, H. D. Johansen, D. Johansen, J. Rittscher, and P. Halvorsen

Abstract: Computer-aided detection, localisation, and segmentation methods can help improve colonoscopy procedures. Even though many methods have been built to tackle automatic detection and segmentation of polyps, benchmarking of state-of-the-art methods still remains an open problem. This is due to the increasing number of researched computer vision methods that can be applied to polyp datasets. Benchmarking of novel methods can provide a direction to the development of automated polyp detection and segmentation tasks. Furthermore, it ensures that the produced results in the community are reproducible and provide a fair comparison of developed methods. In this paper, we benchmark several recent state-of-the-art methods using Kvasir-SEG, an open-access dataset of colonoscopy images for polyp detection, localisation, and segmentation evaluating both method accuracy and speed. Whilst, most methods in literature have competitive performance over accuracy, we show that the proposed ColonSegNet achieved a better trade-off between an average precision of 0.8000 and mean IoU of 0.8100, and the fastest speed of 180 frames per second for the detection and localisation task. Likewise, the proposed ColonSegNet achieved a competitive dice coefficient of 0.8206 and the best average speed of 182.38 frames per second for the segmentation task. Our comprehensive comparison with various state-of-the-art methods reveals the importance of benchmarking the deep learning methods for automated real-time polyp identification and delineations that can potentially transform current clinical practices and minimise miss-detection rates.

Published: IEEE Access, vol. 9, pp. 40496–40510, 2021

Candidate contributions: D. Jha conceptualized and designed this work together with S. Ali (collaborator from the University of Oxford). D. Jha performed experiments and analysis presented in the paper. N. K. Tomar assisted in designing architecture. He wrote most of the manuscript and led the subsequent revisions of the manuscript to incorporate reviewers remarks and suggestions.

Thesis objectives: Objective II, Objective III

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 2, 2021, accepted February 15, 2021, date of publication March 4, 2021, date of current version March 17, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3063716

Real-Time Polyp Detection, Localization and Segmentation in Colonoscopy Using Deep Learning

DEBESH JHA[1,3], SHARIB ALI[2,4], NIKHIL KUMAR TOMAR[1], HÅVARD D. JOHANSEN[3], DAG JOHANSEN[3], JENS RITTSCHER[2,4], MICHAEL A. RIEGLER[1], AND PÅL HALVORSEN[1,5]

[1]SimulaMet, 0167 Oslo, Norway
[2]Department of Engineering Science, Big Data Institute, University of Oxford, Oxford OX4 2PGv, U.K.
[3]Department of Computer Science, UiT—The Arctic University of Norway, 9037 Tromsø, Norway
[4]Oxford NIHR Biomedical Research Centre, Oxford OX4 2PGv, U.K.
[5]Department of Computer Science, Oslo Metropolitan University, 0167 Oslo, Norway

Corresponding authors: Debesh Jha (debesh@simula.no) and Sharib Ali (sharib.ali@eng.ox.ac.uk)

The work of Debesh Jha is funded by the Research Council of Norway project number 263248 (Privaton). The computations in this paper were performed on equipment provided by the Experimental Infrastructure for Exploration of Exascale Computing (eX3), which is financially supported by the Research Council of Norway under contract 270053. Parts of computational resources were also used from the research supported by the National Institute for Health Research (NIHR) Oxford BRC with additional support from the Wellcome Trust Core Award Grant Number 203141/2/16/Z. The work of Sharib Ali is supported by the NIHR Oxford Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

ABSTRACT Computer-aided detection, localization, and segmentation methods can help improve colonoscopy procedures. Even though many methods have been built to tackle automatic detection and segmentation of polyps, benchmarking of state-of-the-art methods still remains an open problem. This is due to the increasing number of researched computer vision methods that can be applied to polyp datasets. Benchmarking of novel methods can provide a direction to the development of automated polyp detection and segmentation tasks. Furthermore, it ensures that the produced results in the community are reproducible and provide a fair comparison of developed methods. In this paper, we benchmark several recent state-of-the-art methods using Kvasir-SEG, an open-access dataset of colonoscopy images for polyp detection, localization, and segmentation evaluating both method accuracy and speed. Whilst, most methods in literature have competitive performance over accuracy, we show that the proposed ColonSegNet achieved a better trade-off between an average precision of 0.8000 and mean IoU of 0.8100, and the fastest speed of 180 frames per second for the detection and localization task. Likewise, the proposed ColonSegNet achieved a competitive dice coefficient of 0.8206 and the best average speed of 182.38 frames per second for the segmentation task. Our comprehensive comparison with various state-of-the-art methods reveals the importance of benchmarking the deep learning methods for automated real-time polyp identification and delineations that can potentially transform current clinical practices and minimise miss-detection rates.

INDEX TERMS Medical image segmentation, ColonSegNet, colonoscopy, polyps, deep learning, detection, localization, benchmarking, Kvasir-SEG.

I. INTRODUCTION

Colorectal Cancer (CRC) has the third highest mortality rate among all cancers. The overall five-year survival rate of colon cancer is around 68%, and stomach cancer is only around 44% [1]. Searching for and removing precancerous anomalies is one of the best working methods to avoid CRC based mortality. Among these abnormalities, polyps in the colon are important to detect because it can develop into the CRC at late stage. Thus, an early detection of CRC is crucial for survival.

After modification in the lifestyle, the prevention from the CRC is the screening of the colon regularly. Different research studies suggest that population-wide screening advances the prognosis and can even reduce the incidence of CRC [2]. Colonoscopy is an invasive medical procedure where an endoscopist examines and operates on the colon using a flexible endoscope. It is considered to be the best

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano.

diagnostic tool for colon examination for early detection and removal of polyps. Therefore, colonoscopic screening is the most preferred technique among gastroenterologists.

Polyps are abnormal growths of tissue protruding from the mucous membrane. They can occur anywhere in the gastrointestinal (GI) tract but are mostly found in the colorectal area and are often considered a predecessor of CRC [3], [4]. Polyps may be pedunculated (having a well-defined stalk) or sessile (without a defined stalk). The colorectal polyps can be categorised into two classes: non-neoplastic and neoplastic. Non-neoplastic polyps are further sub-categorised into hyperplastic, inflammatory, and hamartomatous polyps. These types of polyps are non-cancerous and not harmful. Neoplastic is further sub-categorised into adenomas and serrated polyps. These polyps can develop into the risk of cancer. Based on their size, colorectal polyps can be categorised into three classes, namely, diminutive ($\leq$5mm), small (6 to 9 mm), and advanced (large) ($\geq$10mm) [5]. Usually, larger polyps can be detected and resected.

There exists a significant risk with small and diminutive colorectal polyps [6]. A polypectomy is a technique for the removal of small and diminutive polyps. There are five different polypectomy techniques for resection of diminutive polyps, namely, cold forceps polypectomy, hot forceps polypectomy, cold snare polypectomy, hot snare polypectomy, and endoscopic mucosal resection [5]. Among these techniques, cold snare polypectomy is considered best polypectomy technique for resectioning small colorectal polyps [7].

Colonoscopy is an invasive procedure that requires high-quality bowel preparation as well as air insufflation during examination [8]. It is both an expensive and time-demanding procedure. Nevertheless, on average, 20% of polyps are missed during examinations. The risk of getting cancer therefore relates to the individual endoscopists' ability to detect polyps [9]. Recent studies have shown that new endoscopic devices and diagnostic tools have improved the adenoma detection rate and polyp detection rate [10], [11]. However, the problem of over-looked polyps remains the same.

The colonoscopy videos recorded at the clinical centers store a significant amount of colonoscopy data. However, the collected data are not used efficiently as they are labour intense for the endoscopists [12]. Thus, a second review of videos are often not done. This might lead to missed detection at an early stage largely. Automated data curation and annotation of video data is a prerequisite for building reliable Computer Aided Diagnosis (CADx) systems that can help to assess clinical endoscopy more thoroughly [13]. A fraction of the collected colonoscopy data can be curated to develop computer-aided systems for automated detection and delineation of polyps either during the clinical procedure or after the reporting. At the same time, to build a robust system, it is vital to incorporate data variability related to patients, endoscopic procedure, and endoscope manufacturers. Even though recent developments in computer vision and system designs have enabled us to built accurate and efficient systems, these largely depend on the data availability as most recent methods are data voracious. The lack of availability of public datasets [14] is a critical bottleneck to accelerate algorithm development in this realm.

In general, curating medical datasets are challenging and it requires domain knowledge expertise. Reaching a consensus to achieve ground truth labels from different experts on the same dataset is again another obstacle. Typically, in colonoscopy, smaller polyps or flat/sessile polyps that are usually missed out during a procedure can be difficult to observe even during manual labeling. Other challenges include the patient variability and presence of different sizes, shapes, textures, colors, and orientations of these polyps [3]. Therefore, during polyp data curation and developing of automated systems for the colonoscopy, it is vital that all various challenges often come along routine colonoscopy has to be taken into consideration.

Automatic polyp detection and segmentation systems based on Deep Learning (DL) have a high overall performance in both colonoscopy images and colonoscopy videos [15], [16]. Ideally, the automatic CADx systems for polyps detection, localization, and segmentation should have: 1) consistent performance and improved robustness to patient variability, i.e., the system should be able to produce reliable outputs, 2) high overall performance surpassing the set bar for algorithms, 3) real-time performance required for clinical applicability, and 4) easy-to-use system that can provide with clinically interpretable outputs. Scaling this to a population sized cohort is also a very resource-demanding and incurs enormous costs. As a first step, we therefore target the detection, localization, and segmentation of colorectal polyps known as precursors of CRC. The reason for starting with this scenario is that most colon cancers arise from benign adenomatous polyps (around 20%) containing dysplastic cells. Detection and removal of polyps prevent the development of cancer, and the risk of getting CRC in the following 60 months after a colonoscopy depends largely on the endoscopist ability to detect polyps [9].

Detection and localization of polyps are usually critical during routine surveillance and to measure the polyp load of the patient at the end of the surveillance while pixel-wise segmentation becomes vital to automate the polyp boundary delineation during the surgical procedures or radio-frequency ablations. In this paper, we evaluate DL methods for both detection (and localization referring to bounding box detection) and segmentation (pixel-wise classification or semantic segmentation) SOTA methods on Kvasir-SEG dataset [17] to provide a comprehensive benchmark for the colonoscopy images. The main aim of the paper is to establish a new strong benchmark with existing successful computer vision approaches. Our contributions can be summarised as follows:

- We propose ColonSegNet, an encoder-decoder architecture for segmentation of colonoscopic images. The architecture is very efficient in terms of processing speed

(i.e., produces segmentation of colonoscopic polyp in real-time) and competitive in terms of performance.
- A comprehensive comparison of the state-of-the-art computer vision baseline methods on the Kvasir-SEG dataset is presented. The best approaches show real-time performance for polyp detection, localization, and segmentation.
- We have established strong benchmark for detection and localization on the Kvasir-SEG dataset. Additionally, we have extended segmentation baseline as compared to [3], [17], [18]. These benchmarks can be useful to develop reliable and clinically applicable methods.
- Detection, localization, and semantic segmentation performances are evaluated on standard computer vision metrics.
- Detailed analysis have been presented with the specific focus on the best and worst performing cases that will allow to dissect method success and failure modes required to accelerate algorithm development.

The rest of the paper is organized as follows: In Section II, we present related work in the field. In Section III, we present the material. Section IV presents both detection, localization, and segmentation methods. Result are presented in Section V. Discussion on the best performing detection, localization, and semantic segmentation approaches are presented in Section VI and finally a conclusion is provided in the Section VII.

II. RELATED WORK

Automated polyp detection has been an active topic for research over the last two decades and considerable work has been done to develop efficient methods and algorithms. Earlier works were especially focused on polyp color and texture, using handcrafted descriptors-based feature learning [27], [28]. More recently, methods based on Convolutional Neural Networks (CNNs) have received significant attention [29], [30], and have been the go to approach for those competing in public challenges [31], [32].

Wang *et al.* [33] designed algorithms and developed software modules for fast polyp edge detection and polyp shot detection, including a polyp alert software system. Shin *et al.* [34] have used region-based CNN for automatic polyp detection in colonoscopy videos and images. They used Inception ResNet as a transfer learning approach and post-processing techniques for reliable polyp detection in colonoscopy. Later on, Shin *et al.* [14] used generative adversarial network [35], where they showed that the generated polyp images are not qualitatively realistic; however, they can help to improve the detection performance. Lee *et al.* [15] used YOLO-v2 [36], [37] for the development of polyp detection and localization algorithm. The algorithm produced high sensitivity and near real-time performance. Yamada *et al.* [38] developed an artificial intelligence system that can automatically detect the sign of CRC during colonoscopy with high sensitivity and specificity. They claimed that their system could aid endoscopists in real-time detection to avoid abnormalities and enable early disease detection.

In addition to the work related to automatic detection and localization, pixel-wise classification (segmentation) of the disease provides an exact polyp boundary and hence is also of high significance for clinical surveillance and procedures. Bernel *et al.* [31] presented the results of the automatic polyp detection subchallenge, which was the part of the endoscopic vision challenge at the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2015 conference. This work compared the performance of eight teams and provided an analysis of various detection methods applied on the provided polyp challenge data. Wang *et al.* [16] proposed a DL-based SegNet [39] that had a real-time performance with an inference of more than 25 frames per second. Geo and Matuszewski [40] used fully convolution dilation networks on the Gastrointestinal Image ANAlysis (GIANA) polyp segmentation dataset. Jha *et al.* [3] proposed ResUNet++ demonstrating 10% improvement compared to the widely used UNet baseline on Kvasir-SEG dataset. They also further applied the trained model on the CVC-ClinicDB [23] dataset showing more than 15% improvement over UNet. Ali *et al.* [32] did a comprehensive evaluation for both detection and segmentation approaches for the artifacts present clinical endoscopy including colonoscopy data [41]. Wang *et al.* [42] proposed a boundary-aware neural network (BA-Net) for medical image segmentation. BA-Net is an encoder-decoder network that is capable of capturing the high-level context and preserving the spatial information. Later on, Jha *et al.* [43] proposed DoubleUNet for the segmentation, which was applied to four biomedical imaging datasets. The proposed DoubleUNet is the combination of two UNet stacked on top of each other with some additional blocks. Experimental results on CVC-Clinic and ETIS-Larib polyp datasets show the state-of-the-art (SOTA) performances. In addition to the related work on polyp segmentation, there are studies on segmentation approaches [44]–[47].

Datasets has been instrumental for medical research. Table 1 shows the list of the available endoscopic image and video datasets. Kvasir-SEG, ETIS-Larib, and CVC-ClinicDB contain colonoscopy images, whereas Kvasir, Nerthus, and HyperKvasir contain the images from the whole GI. KvasirCapusle contains images from video capsule endoscopy. All the dataset contains images acquired from conventional White Light (WL) imaging technique except the EDD dataset, where it contains images from both WL imaging and Narrow Band Imaging (NBI) techniques. All of these datasets contain at least a polyp class. Out of nine available datasets, Kvasir-SEG [17], ETIS-Larib [22], and CVC-ClinicDB [23] has manually labeled ground truth masks. Among them, Kvasir-SEG offers the most number of annotated samples providing both ground truth masks and bounding boxes offering detection, localization, and segmentation task. All of the datasets are publicly available.

TABLE 1. Available endoscopic datasets.

Dataset	Organ	Source	Findings	Dataset content	Task type
Kvasir-SEG [17]	Large bowel	WL$^{\diamond}$	Polyp	1000 images	Detection, localization & segmentation
Kvasir [19]	Whole GI	WL$^{\diamond}$	Polyps, esophagitis, ulcerative colitis, z-line, pylorus, cecum, dyed polyp, dyed resection margins, stool	8,000 images	Classification
Nerthus [20]	Large bowel	WL$^{\diamond}$	Stool - categorization of bowel cleanliness	21 videos	Classification
HyperKvasir [21]	Whole GI	WL$^{\diamond}$	16 different classes from upper GI & 24 different classes from lower GI tract	110,079 images & 373 videos	Classification
ETIS-Larib [22]	Colonoscopy	WL$^{\diamond}$	Polyp	196 images	Segmentation
CVC-Clinic [23]	Colonoscopy	WL$^{\diamond}$	Polyp	612 images	Segmentation
KvasirCapsule [24]	Whole GI	VCE	13 different classes of GI anomalies	4,820,739 images & 118 videos	Classification
EDD 2020 [25]	Entire GI	NBI†, WL$^{\diamond}$	Polyp, Barrett's esophagus, high-grade dysplasia, suspicious (low-grade), cancer	386 images	Detection, localization & segmentation
Kvasir-Instrument [26]	Large Bowel	WL$^{\diamond}$	Tools and instruments	590 images	Detection, localization, Segmentation

† Narrow band imaging $^{\diamond}$White light imaging

Dataset development, benchmarking of the methods, and evaluation are critical in the medical imaging domain. It inspires the community to build clinically transferable methods on a well-curated and standardised dataset. Due to the lack of benchmark papers, it becomes utmost difficult to understand the clear strength of methods in the literature. New algorithm developments demonstrating its translational abilities in clinics is thus very minimal. Data science challenges do offer some insight, however, a comprehensive analysis on various different aspects such as detection, localization, segmentation, and inference time estimation are still not covered by the most.

Inspired by the previous benchmark for polyp detection [31], endoscopic artifact detection [41], endoscopic disease detection and segmentation [25], endoluminal scene object segmentation [48], and endoscopic instrument segmentation [49], we introduce a new benchmark for the automatic polyp detection, localization and segmentation using publicly available Kvasir-SEG dataset.

III. MATERIALS – DATASET

We have used the Kvasir-SEG [17] for detection, localization, and segmentation tasks. Figure 1 shows the image, ground truth information, and their detection (their localised bounding boxes in red). This dataset is the outcome of an initiative for open and reproducible results. It contains 1000 polyp images acquired by high-resolution electromagnetic imaging system, i.e., ScopeGuide, Olympus Europe, their corresponding masks and bounding box information. The images and their ground truths can be used for the segmentation task, whereas the bounding box information provides an opportunity for the detection task. The resolution of the images in this dataset ranges from 332 × 487 to 1920 × 1072 pixels. The dataset can be downloaded at https://datasets.simula.no/kvasir-seg/. The dataset includes

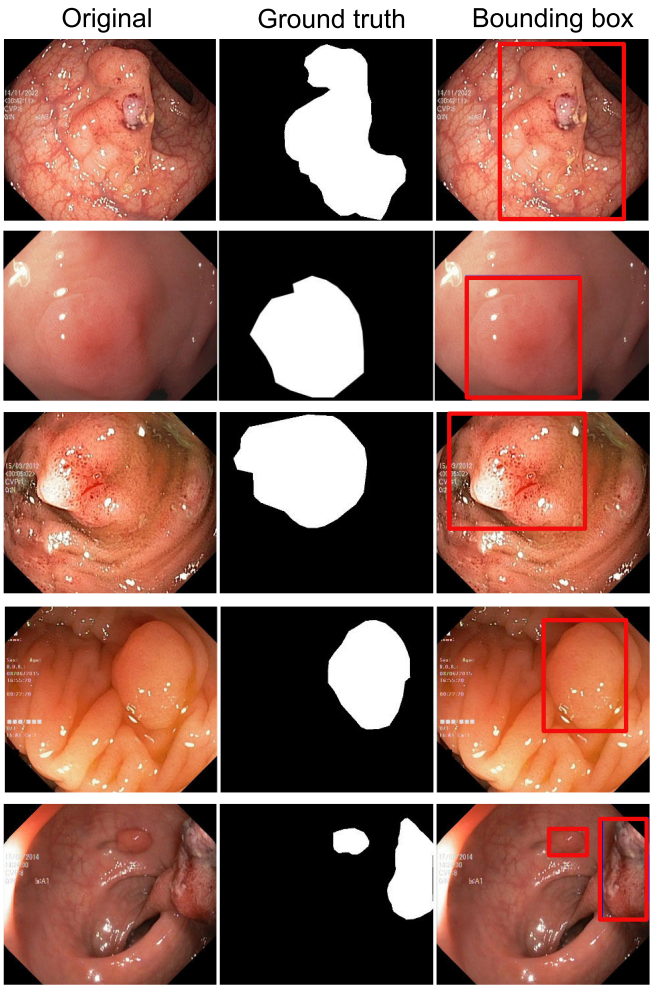


FIGURE 1. Sample images from Kvasir-SEG dataset: Annotated masks (2nd column) and bounding boxes (3rd column) for selected samples.

images of 700 large polyps ($> 160 \times 160$ pixels), 323 medium sized polyps ($> 64 \times 64$ pixels and $\leq 160 \times 160$ pixels)

and 48 small polyps ($\leq 64 \times 64$ pixels). In total, the dataset consists of 1072 images of polyps with segmentation masks and bounding boxes.

IV. METHOD

Detection methods aim to predict the object class and regress bounding boxes for localization, while segmentation methods aim to classify the object class for each pixel in an image. In Figure 1, ground truth masks for segmentation task are shown in 2nd column while corresponding bounding boxes for the detection task are in 3rd column. This section describes the baseline methods for detection, localization and segmentation methods used for the automated detection and segmentation of polyp in the Kvasir-SEG dataset.

A. DETECTION AND LOCALIZATION BASELINE METHODS

Detection methods consist of input, backbone, neck, and head. The input can be images, patches, or image pyramids. The backbone can be different CNN architectures such as VGG16, ResNet50, ResNext-101, and Darknet. The neck is the subset of the backbone network, which could consist of FPN, PANet, and Bi-FPN. The head is used to handle the prediction boxes that can be one stage detector for dense prediction (e.g., YOLO, RPN, and RetinaNet [50]), and two-stage detector with the sparse prediction (e.g., Faster R-CNN [51] and RFCN [52]). Recently, one stage methods have attracted much attention due to their speed and ability to obtain optima accuracy. This has been possible because recent networks utilise feature pyramid networks or spatial-pyramid pooling layers to predict candidate bounding boxes which are regressed by optimising loss functions (see Figure 2).

In this paper, we use EfficientDet [53] which uses EfficientNet [54], as the backbone architecture, bi-directional feature pyramid network (BiFPN) as the feature network, and shared class/box prediction network. Additionally, we also use Faster R-CNN [51], which uses region proposal network (RPN), as the proposal network and Fast R-CNN [55] as the detector network. Moreover, we use YOLOv3 [56] that utilises multi-class logistic loss (*binary cross-entropy* for classification loss and *mean square error* for regression loss) modeled with regularizers such as objectness prediction scores. Furthermore, we also used YOLOv4 [57], which utilises an additional bounding box regressor based on the Intersection over Union (IoU) and a cross-stage partial connections in their backbone architecture. Additionally, YOLOv4 allows on fly data augmentation, such as mosaic and cut-mix.

RetinaNet [50] takes into account the data driven property that allows the network to focus on "hard" samples for improved accuracy. The easy to adapt backbones for feature extraction at the beginning of the network provides the opportunity to experiment with deeper and varied architectures such as ResNet50, and ResNet101 for RetinaNet and 53 layered Darknet53 backbone for YOLOv3 and YOLOv4 architecture. To tackle the different aspect ratio problem, for both one stage networks, optimal anchor

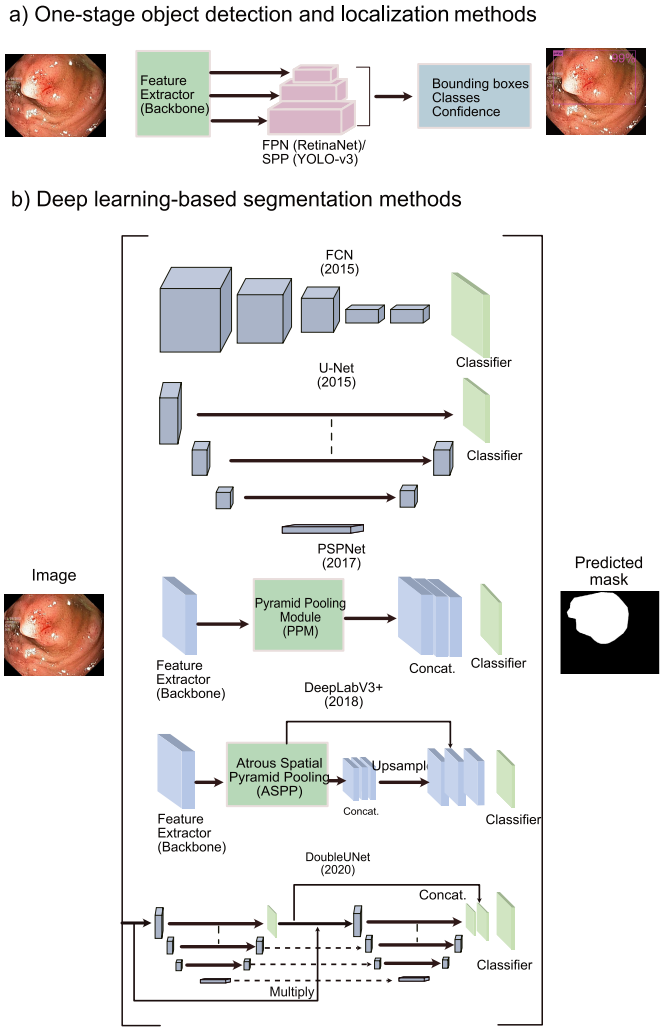


FIGURE 2. Baseline detection, localization and semantic segmentation method summary.

boxes [51] are searched and pre-defined for the provided data to tackle large variance of scale and aspect ratio of boxes. Table 2 shows the hyperparameter used by each of the object detection methods for the detection task.

B. SEGMENTATION BASELINE METHODS

In the past years, data-driven approaches using CNNs have changed the paradigm of computer vision methods, including segmentation. An input image can be directly be fed to convolution layers to obtain feature maps, which can be later upsampled to predict pixel-wise classification providing object segmentation. Such networks learn from available ground truth labels and can be used to predict labels from other similar data. A Fully Convolutional Network (FCN) based segmentation was first proposed by Long *et al.* [58] that can be trained end-to-end. Ronneberger *et al.* [59] modified and extended the FCN architecture to a UNet architecture. The UNet consist of an analysis (*encoder*) and a synthesis (*decoder*) path. In the analysis path of the network, deep features are learnt, whereas in the synthesis path segmentation is performed on the basis of the learnt features.

Pyramid Scene Parsing Network (PSPNet) [60] introduced a pyramid pooling module aimed at aggregating global context information from different regions which are upsampled and concatenated to form the final feature representation. A final per-pixel prediction is obtained after a convolution layer (see Figure 2, third architecture). For feature extraction, we have used the ResNet50 architecture pretrained on imageNet. Similar to the UNet architecture, DeepLabV3+ [61] is an encoder-decoder network. However, it utilizes atrous separable convolutions and spatial pyramid pooling (see Figure 2, last architecture) for fast inference and improved accuracy. Atrous convolution controls the resolution of features computed and adjust the receptive field to effectively capture multi-scale information. In this paper, we have used an output stride of 16 for both encoder and decoder networks of DeepLabV3+ and have experimented on both ResNet50 and ResNet101 backbones.

ResUNet [62] integrates the power of both UNet and residual neural network. ResUNet++ [3] is the improved version of ResUNet architecture. It has additional layers including squeeze-and-excite block, Atrous Spatial Pyramid Pooling (ASPP), and attention block. These additional layers helps learning the deep features that are capable of improved prediction of pixels for object segmentation tasks. DoubleU-Net [43] consists of two modified UNet architecture. It uses VGG-19 pretrained on ImageNet [63] as the first encoder. The main reason behind using VGG-19 (similar to UNet [64]) was that it is a lightweight model. The additional component in the DoubleUNet are squeeze-and-excite block, and ASPP block. High-Resolution Network (HRNet) [65] maintains high-resolution representation convolution in parallel and interchange the information across the resolution continuously. This is one of the most recent and popular method in the literature. Furthermore, we have used UNet with ResNet34 as a backbone network and trained the model to compare with the other state-of-the-art semantic segmentation networks.

Table 4 shows the hyperparameters used for each of the semantic segmentation based benchmark methods used. From the table, we can see that number of trainable parameters of the baseline methods are large. A high number of trainable parameters in the network makes it complex, leading to a lower frame rate. It is therefore essential to design an efficient, lightweight architecture that can provide a higher frame rate and better performance. In this regard, we propose a novel architecture, ColonSegNet, that requires only few number of training parameters, which can save training and inference time. More details about the architecture can be found in the below section.

C. COLONSEGNET

Figure 3 shows the block diagram of the proposed ColonSegNet. It is an encoder-decoder that uses residual block [66] with squeeze and excitation network [67] as the main component. The network is designed to have very few trainable parameters as compared to other baseline networks such as

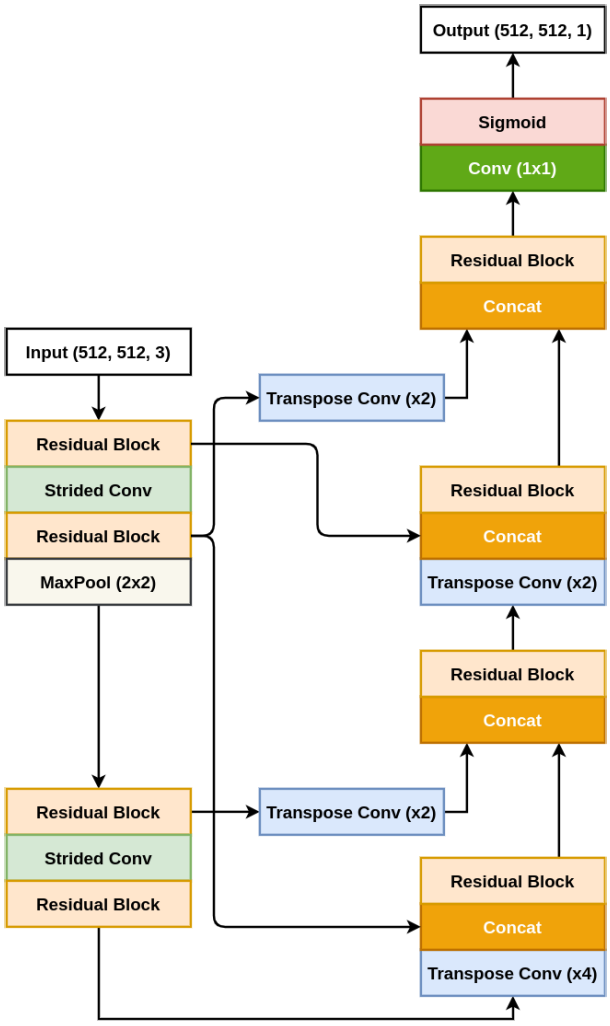


FIGURE 3. Block diagram of ColonSegNet.

U-Net [59], PSPNet [60], DeepLabV3+ [61], and others. The use of fewer trainable parameters makes the proposed architecture a very light-weight network that leads to real-time performance.

The network consists of two encoder blocks and two decoder blocks. The encoder network learns to extract all the necessary information from the input image, which is then passed to the decoder. Each decoder block consists of two skip connections from the encoder. The first is a simple concatenation, and the second skip connection passed through a transpose convolution to incorporates multi-scale features in the decoder. These multi-scale features help the decoder to generate more semantic and meaningful information in the form of a segmentation mask.

The input image is fed to the first encoder, which consists of two residual blocks and a 3×3 strided convolution in between them. This layer is followed by a 2×2 max-pooling. Here, the output feature map spatial dimensions are reduced to $\frac{1}{4}$ of the input image. The second encoder consists of two residual blocks and a 3×3 strided convolution in between them.

The decoder starts with a transpose convolution, where the first decoder uses a stride value 4, which increases the feature map spatial dimensions by 4. Similarly, the second decoder uses a stride value of 2, increasing the spatial dimensions by 2. Then, the network follows a simple concatenation and a residual block. Next, it is concatenated with the second skip connection and again followed by a residual block. The output of the last decoder block passes through a 1×1 convolution and a sigmoid activation function, generating the binary segmentation mask.

1) DATA AUGMENTATION

Supervised learning methods are data voracious and require large amount of data to obtain reliable and well-performing models. Acquiring such training data through data collection, curation, and annotation is a manual process that needs significant resources and man-hours from both clinical experts and computational scientists.

Data augmentation is a common technique to computationally increase the number of training samples in a dataset. For our DL models, we use basic augmentation techniques such as horizontal flipping, vertical flipping, random rotation, random scale, and random cropping. The images used in all the experiments undergo normalization and are resized to a fixed size of 512×512. For the normalization, we subtract the image by mean and divide it by standard deviation.

V. RESULTS

In this section, we first present our evaluation metrics and experimental setup. Then, we present both quantitative and qualitative results.

A. EVALUATION METRICS

We have used standard computer vision metrics to evaluate polyp detection and localization, and semantic segmentation methods on the Kvasir-SEG dataset.

1) DETECTION AND LOCALIZATION TASK

For the object detection and localization task, the commonly used Average Precision (AP) and IoU have been used [68], [69].

- IoU: This metric measures the overlap between two bounding boxes A and B as the ratio between the overlapped area.

$$\text{IoU(A,B)} = \frac{A \cap B}{A \cup B} \tag{1}$$

- AP: AP is computed as the Area Under Curve (AUC) of the precision-recall curve of detection sampled at all unique recall values (r1, r2, ...) whenever the maximum precision value drops:

$$\text{AP} = \sum_{n} \left\{ (r_{n+1} - r_n)\, p_{\text{interp}}(r_{n+1}) \right\}, \tag{2}$$

with $p_{\text{interp}}(r_{n+1}) = \max_{\tilde{r} \geq r_{n+1}} p(\tilde{r})$. Here, $p(r_n)$ denotes the precision value at a given recall value. This definition ensures monotonically decreasing precision. AP was computed as an average APs for IoU from 0.25 to 0.75 with a step-size of 0.05 which means an average over 11 IoU levels are used (AP @[.25 : .05 : .75]).

2) SEGMENTATION TASK

For polyp segmentation task, we have used widely accepted computer vision metrics that include Dice Coefficient (DSC), Jaccard Coefficient (JC), precision *(p)*, and recall *(r)*, and overall accuracy *(Acc)*. JC is also termed as IoU. We have also included Frame Per Second (FPS) to evaluate the clinical applicability of the segmentation methods in terms of inference time during the test.

To define each metric, let *tp*, *fp*, *tn*, and fn represents true positives, false positives, true negatives, and false negatives, respectively.

$$\text{DSC} = \frac{2 \cdot tp}{2 \cdot tp + fp + fn} \tag{3}$$

$$\text{IoU} = \frac{tp}{tp + fp + fn} \tag{4}$$

$$r = \frac{tp}{tp + fn} \tag{5}$$

$$p = \frac{tp}{tp + fp} \tag{6}$$

$$\text{F2} = \frac{5p \times r}{4p + r} \tag{7}$$

$$Acc = \frac{tp + tn}{tp + tn + fp + fn} \tag{8}$$

$$\text{FPS} = \frac{\#frames}{sec} = \frac{1}{sec/frame} \tag{9}$$

B. EXPERIMENTAL SETUP AND CONFIGURATION

The methods such as UNet, ResUNet, ResUNet ++, DoubleUNet, and HRNet were implemented using Keras [70] with a Tensorflow [71] back-end and were run on a Volta 100 GPU and an Nvidia DGX-2 AI system. A PyTorch implementation for FCN8, PSPNet, DeepLabv3 +, UNet-ResNet34, and ColonSegNet networks were done. Training of these methods were conducted on NVIDIA Quadro RTX 6000. NVIDIA GTX2080Ti was used for test inference for all methods reported in the paper. All of the detection methods were implemented using PyTorch and used NVIDIA Quadro RTX 6000 hardware for training the network.

In all of the cases, we used 880 images for training and the remaining 120 images for the validation. Due to different image sizes in the dataset, we resized the images to 512×512. Hyperparamters are important for the DL algorithms to find the optimal solution. However, picking the optimal hyperparameter is difficult. There are algorithms such as grid search, random search, and advanced solutions such as Bayesian optimization for finding the optimal parameters. However, an algorithm such as Bayesian optimization is computationally costly, making it difficult to test several DL algorithms. We have done an extensive hyperparameter

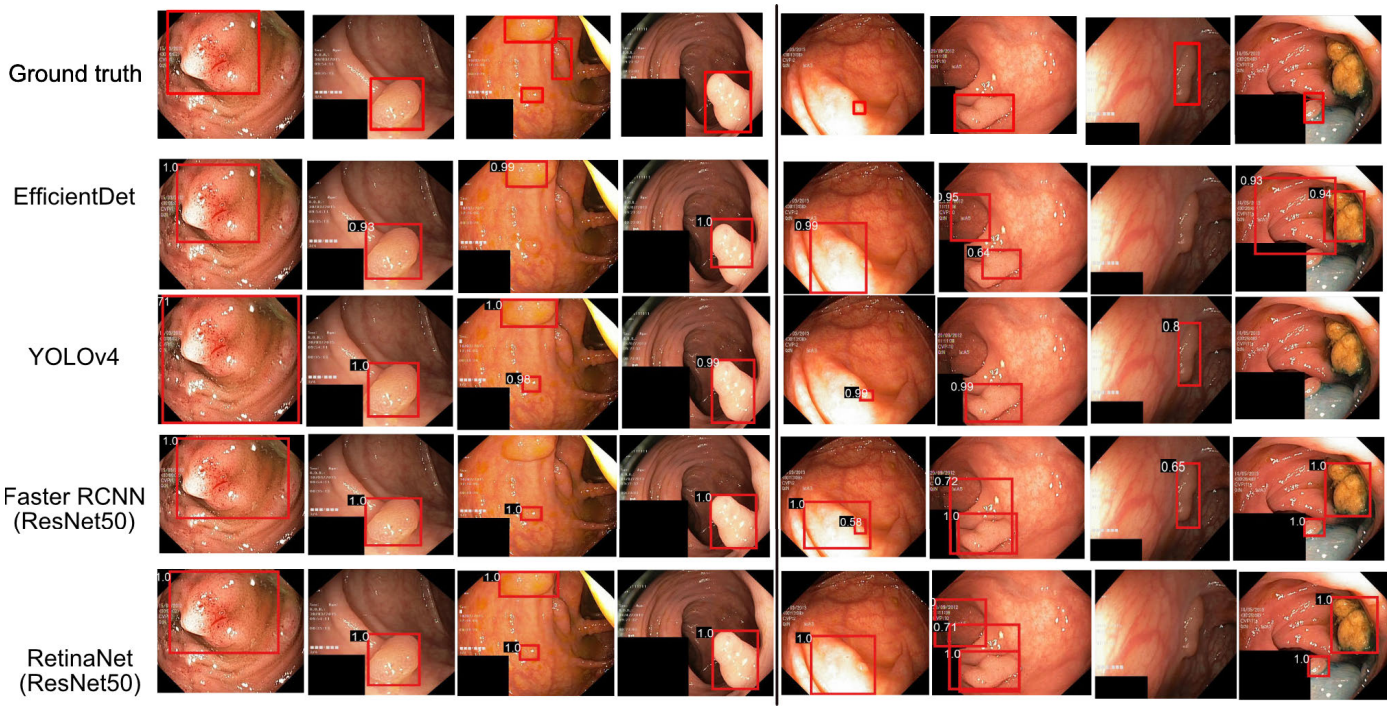


FIGURE 4. **Detection and localization results on test dataset: On right of the black solid line, images where EfficientDet-D0, YOLOv4, Faster R-CNN and RetinaNet (with ResNet50 backbone) have similar results and in most cases obtained highest IoU. On left, images with failed case (worse localization) for either of the method. Confidence scores are provided on the top-left of the red prediction boxes.**

TABLE 2. **Hyperparameters used for baseline methods for polyp detection and localization task on Kvasir-SEG. Here, CIoU: complete intersection-of-union loss, MSE: mean square error, CE: cross-entropy.**

Method	Learning rate	Optimizer	Batch size	Loss	Anchors	Threshold
Faster R-CNN [51]	$2.5e^{-4}$	Adam	8	$L1^{smooth}$, log-loss	256	0.4
RetinaNet [50]	$1e^{-5}$	SGD	8	$L1^{smooth}$, focal loss	15 (pyramid)	0.3
YOLOv3+spp [56]	$1e^{-3}$	SGD	16	MSE, CE	8	0.25
YOLOv4 [57]	$1e^{-3}$	SGD	16	CIoU, CE	8	0.25
EfficientDet-D0 [53]	$1e^{-4}$	Adam	8	Focal loss	default	0.4

TABLE 3. **Result on the polyp detection and localization task on the Kvasir-SEG dataset. Two best scores are highlighted in bold.**

Method	Backbone	AP	IoU	AP_{25}	AP_{50}	AP_{75}	FPS
EfficientDet-D0 [53]	EfficientNet-b0, biFPN	0.4756	0.4322	0.6846	0.5047	0.2280	35.00
Faster R-CNN [51]	ResNet50	0.7866	0.5621	0.8947	0.8418	0.5660	8.00
RetinaNet [50]	ResNet50	**0.8697**	0.7313	**0.9395**	**0.9095**	0.6967	16.20
RetinaNet [50]	ResNet101	**0.8745**	0.7579	**0.9483**	**0.9095**	0.7132	16.80
YOLOv3+spp [56]	Darknet53	0.8105	**0.8248**	0.8856	**0.8532**	**0.7586**	45.01
YOLOv4 [57]	Darknet53, CSP	0.8513	0.8025	0.9123	0.8234	**0.7594**	**48.00**
ColonSegNet (Proposed)	-	0.8000	**0.8100**	0.9000	0.8166	0.6706	**180.00**

search for finding the optimal hyperparameters for polyp detection, localization, and segmentation task. These sets of hyperparameters were chosen based on empirical evaluation. The used hyperparameters are for the Kvasir-SEG dataset and are reported in the Table 2, and Table 4.

C. QUANTITATIVE EVALUATION

1) DETECTION AND LOCALIZATION

Table 3 shows the detailed result for the polyp detection and localization task on the Kvasir-SEG dataset. It can be observed that RetinaNet shows improvement over YOLOv3 and YOLOv4 for mean average precision computed for multiple IoU thresholds and for average precision at IoU threshold 25 (AP_{25}) and 50 (AP_{50}). RetinaNet with ResNet101 backbone achieved an average precision of 0.8745, while YOLOv4 yielded 0.8513. However, for the IoU threshold of 0.75, YOLOv4 showed improvement over RetinaNet with (AP_{75}) of 0.7594 against 0.7132 for RetinaNet with ResNet101 backbone. Similarly, the average IoU of 0.8248 was observed for YOLOv3, which is nearly 8% improvement over RetinaNet. IoU determines the preciseness of the bounding box localization. EfficientDet-D0 obtained the least AP of 0.4756 and IoU of 0.4322. Faster R-CNN obtained an AP of 0.7866. However, it only

TABLE 4. Hyperparameters used for baseline methods for polyp segmentation task on Kvasir-SEG dataset.

Method	No. of parameters	Learning rate	Optimizer	Batch size	Loss	Momentum	Decay rate
UNet [58]	7,858,433	$1e^{-2}$	SGD	8	Cross-entropy	-	-
ResUNet [61]	8,420,077	$1e^{-4}$	Adam	8	Dice loss	-	-
ResUNet++ [3]	16,242,785	$1e^{-4}$	Adam	8	Dice loss	-	-
HRNet [64]	9,524,036	$1e^{-4}$	Adam	8	Dice loss	-	-
DoubleUNet [42]	29,303,426	$1e^{-4}$	Adam	8	Dice loss	-	-
PSPNet [59]	48,631,850	$1e^{-2}$	SGD	8	Cross-entropy	-	-
DeepLabv3+ [60]	ResNet50: 39,756,962	$1e^{-2}$	SGD	8	Cross-entropy	0.9	$1e^{-4}$
DeepLabv3+ [60]	ResNet101: 58,749,090	$1e^{-3}$	SGD	8	Cross-entropy	0.9	$1e^{-4}$
FCN8 [57]	134,270,278	$1e^{-2}$	SGD	8	Cross-entropy	0.9	$1e^{-4}$
UNet-ResNet34	33,509,098	$1e^{-5}$	Adam	8	Cross-entropy	0.9	$1e^{-4}$
ColonSegNet (Proposed)	5,014,049	1e-4	Adam	8	Cross-entropy + Dice loss	-	-

TABLE 5. Baseline methods for polyp segmentation on the Kvasir-SEG dataset. Two best scores are highlighted in bold. "-" shows that there is no backbone used in the network.

Method	Backbone	Jaccard C.	DSC	F2-score	Precision	Recall	Overall Acc.	FPS
UNet [58]	-	0.4713	0.5969	0.5980	0.6722	0.6171	0.8936	11.0161
ResUNet [61]	-	0.5721	0.6902	0.6986	0.7454	0.7248	0.9169	14.8204
ResUNet++ [3]	-	0.6126	0.7143	0.7198	0.7836	0.7419	0.9172	7.0193
FCN8 [57]	VGG 16	0.7365	0.8310	0.8248	0.8817	0.8346	0.9524	24.9100
HRNet [64]	-	0.7592	0.8446	0.8467	0.8778	0.8588	0.9524	11.6970
DoubleUNet [42]	VGG 19	0.7332	0.8129	0.8207	0.8611	0.8402	0.9489	7.4687
PSPNet [59]	ResNet50	0.7444	0.8406	0.8314	0.8901	0.8357	0.9525	16.8000
DeepLabv3+ [60]	ResNet50	0.7759	0.8572	0.8545	0.8907	**0.8616**	**0.9614**	27.9000
DeepLabv3+ [60]	ResNet101	**0.7862**	**0.8643**	**0.8570**	**0.9064**	0.8592	0.9608	16.7500
UNet [58]	ResNet34	**0.8100**	**0.8757**	**0.8622**	**0.9435**	**0.8597**	**0.9681**	**35.0000**
ColonSegNet (Proposed)	-	0.7239	0.8206	0.8206	0.8435	0.8496	0.9493	**182.3812**

obtained an FPS of 8. YOLOv4 with Darknet53 as backbone obtained a FPS of 48, which is 6× faster than Faster R-CNN. The other competitive network was YOLOv3, with an average FPS of 45.01. However, its average precision value is 5% less than YOLOv4. Thus, the quantitative results show that the YOLOv4 with Darknet can detect different types of polyps at a real-time speed of 48 FPS and average precision of 0.8513. Therefore, from the evaluation metrics comparison, YOLOv4 with Darknet53 is the best model for detection and localization of polyp. The results suggest that the model can help gastroenterologists find missed polyps and decrease the polyp miss-rate. Even though, the proposed ColonSegNet is primary built for real-time segmentation of polyps, we compared the bounding box predictions of the proposed network with SOTA detection methods. It can be observed that the inference of the proposed method is nearly four times faster (180 FPS) than YOLOv4. Additionally, it is also obtaining competitive scores on both AP and IoU metrics (IoU of 0.81 and AP of 0.80). Therefore, it can also be considered as one of the best detection and localization techniques.

2) SEGMENTATION

Table 5 shows the obtained results on the polyp segmentation task. It can be observed that the UNet with ResNet34 backbone performs better than the other SOTA segmentation methods in terms of DSC, and IoU. However, the proposed ColonSegNet outperforms in terms of processing speed. ColonSegNet is faster than UNet-ResNet34 by more than four times in processing colonoscopy frames. The complexity of the network is six times smaller than the UNet-ResNet34 network. The proposed network is even smaller than the conventional UNet, with its size only being around 0.75 times that of the UNet with higher scores on evaluation metrics compared to the classical UNet and its derivates such as ResUNet and ResUNet ++. Additionally, the recall and overall accuracy metrics of ColonSegNet are close to the highest performing UNet-ResNet34 network, which shows the proposed method's efficiency.

The original implementation of UNet obtained the least DSC of 0.5969, whereas the UNet with ResNet34 as the backbone model obtained the highest DSC of 0.8757. The second and third best DSC scores of 0.8643 and 0.8572 were obtained for DeepLabv3+ with ResNet101 and DeepLabv3+ with ResNet50 as the backbone, respectively. From the table, it is seen that DeepLabv3+ with ResNet101 performs better than Deeplabv3+ with ResNet50. This may be because of the top-5 accuracy (i.e., the validation results on the ImageNet model) of ResNet101 is slightly better than ResNet50.[1] Despite of DeepLabv3+ with ResNet101

[1] https://keras.io/api/applications/

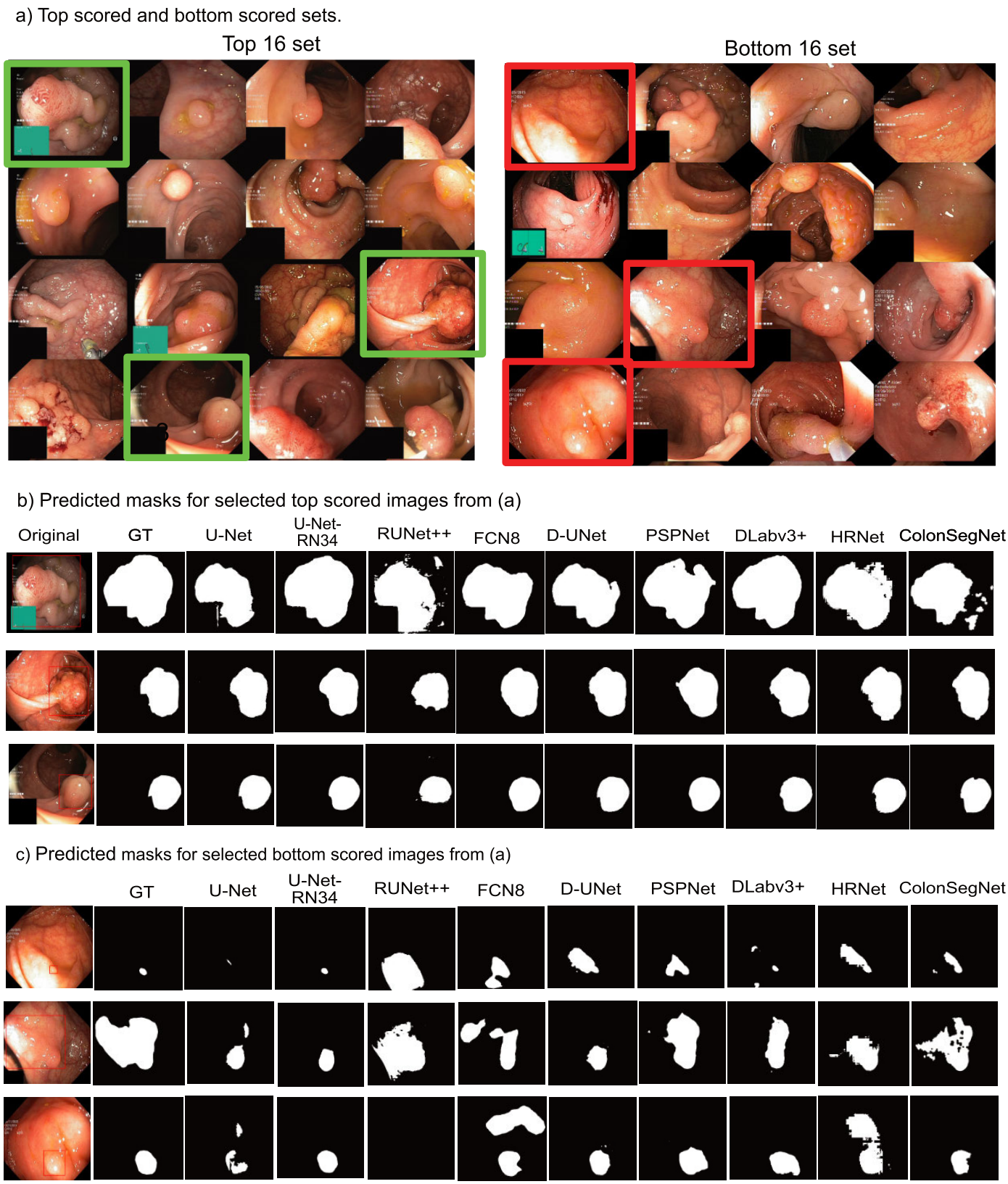


FIGURE 5. Best and worse performing samples for polyp segmentation: a) Top (left) and bottom (right) scored sets, b) predicted masks for top scored images and c) bottom scored images for all methods compared to the ground truth (GT) masks. Green rectangles represent the selected images from top scored set and red rectangle represent those from bottom set. Here, UNet-RN34: UNet-ResNet34, RUNet ++: ResUNet ++, D-UNet: Double UNet, DLabv3 +: DeepLabv3 + (ResNet50).

backbone having the total number of trainable parameters more than 11 times and DeepLabv3+ with ResNet34 being nearly eight times computational complexity, the DSC of ColonSegNet is competitive compared to both of these networks. However in terms, of processing speed, it is almost 11 times faster than DeepLabv3 + with ResNet101 and

nearly seven times faster than DeepLabv3 with ResNet34 backbone.

FCN8, HRNet and DoubleUNet provided similar results with DSC of 0.8310, 0.8446, and 0.8129 while ResUNet ++ achieved DSC of only 0.7143. A similar trend can be observed for F2-score for all methods. For precision, UNet with ResNet34 backbone achieved the maximum score of $p = 0.9435$, and DeepLabv3 + with ResNet50 backbone achieved the highest scores of $r = 0.8616$, while UNet scored the worst with $p = 0.6722$ and $r = 0.6171$. The overall accuracy was outstanding for most methods, with the highest for UNet and ResNet34 as the backbone. IoU is also provided in the table for each segmentation method for scientific completion. Again, UNet and ResNet34 surpassed others with a mIoU score of 0.8100. Also, UNet and ResNet34 achieved the highest FPS rate of 35 fps, which is acceptable in terms of speed and is relatively faster as compared to DeepLabv3+ with ResNet50 (27.9000) and DeepLabv3+ with ResNet101 (16.7500) and other SOTA methods. Additionally, when we consider the number of parameter uses (see Table 4), UNet with ResNet34 backbone uses less number of the parameters as compared to that of FCN8 or DeepLabv3 + network. Due to the low number of trainable parameters and fastest inference time, ColonSegNet is computationally efficient and becomes the best choice while considering the need for real-time segmentation (182.38 FPS on NVIDIA GTX2080Ti) of polyps with deployment possible on even low-end hardware devices making it feasible for many clinical settings. Whereas, UNet with ResNet34 backbone seems the best choice while taking DSC metric into account, however, with speed of only 35 FPS on NVIDIA GTX2080Ti.

D. QUALITATIVE EVALUATION

Figure 4 shows the qualitative result for the polyp detection and localization task along with their corresponding confidence scores. It can be observed that for most images on the left side of the vertical line, both YOLOv4 and RetinaNet are able to detect and localise polyps with higher confidence, except for the third column sample where most of these methods can identify only some polyp areas. Similarly, on the right side of the vertical line, the detected bounding boxes for 5th and 6th column images are too wide for the RetinaNet, while YOLOv4 has the best localization of polyp (observe the bounding box). Also, in the seventh column, RetinaNet and EfficientDet D0 misses the polyp. In the eighth column, YOLOv4 and EfficientDet D0 misses the small polyp completely while stool and polyp is detected as polyp by the Faster R-CNN and RetinaNet.

Figure 5 shows the result for the top-scored and bottom scored sets selected based on their dice similarity coefficient values for the semantic segmentation methods. It can be seen that all the algorithms are able to detect large polyps and produce high-quality masks (see Figure 5(b).

Here, the best obtained segmentation results can be observed for DeepLabv3+ and UNet-ResNet34. However, as shown in Figure 5(c), the segmentation results are affected for flat polyps (very small), images with a certain degree of inclined view, and for the images with saturated areas. The proposed ColonSegNet is able to achieve similar shapes compared to these of the ground truth with some outliers for the predictions which can be seen in Figure 5(b), while for the prediction on worse performing images in Figure 5(c), our proposed network provides comparatively improved predictions on almost all samples.

VI. DISCUSSION

It is evident that there is a growing interest in the investigation of computational support systems for decision making through endoscopic images. For the first time, we are using Kvasir-SEG for detection and localization tasks, and comparing segmentation methods with most recent SOTA methods. We provide a reproducible benchmarking of the DL methods using standard computer vision metrics in object detection and localization, and semantic segmentation. The choice of methods are based their popularity in the medical image domain for detection and segmentation (e.g., UNet, Faster R-CNN), speed (e.g., UNet with ResNet34, YOLOv3), and accuracy (e.g., PSPNet, FCN8, or DoubleUNet) or a combination of all (e.g., DeepLabv3+, YOLOv4).

From the experimental results in Table 3, we can observe that the combination of YOLOv3 with Darknet53 backbone shows improvement over other methods in terms of mIoU, which means a better localization compared to counterpart RetinaNet. However, YOLOv4 is 3× faster than RetinaNet and has a good trade-off between the average precision and IoU. This is because of their Cross-Stage-Partial-Connections (CSP) and CIoU loss for bounding box regression. However, RetinaNet with the backbone ResNet101 shows competitive results surpassing other methods on average precision but nearly 5% less IoU compared to YOLOv4 and nearly 5% less than YOLOv3-spp. Similarly, state-of-the-art methods Faster R-CNN and EffecientDet-D0 provided the least AP and IoU.

A choice between computational speed, accuracy and precision is vital in object detection and localization tasks, especially for colonoscopy video data where speed is a vital element to achieve real-time performance. Therefore, we consider YOLOv4 with Darknet53 and CSP backbone as the best approach in the table for the polyp detection and localization task.

For the semantic segmentation tasks, ColonSegNet showed improvement over all the methods. The method obtained the highest FPS of 182.38. The quantitative results in Figure 5 (b) showed the most accurate delineation of polyp pixels compared to other SOTA methods considered in this paper. The most competitive method to ColonSegNet was UNet with ResNet34 backbone. The other comparable method was DeepLabv3 +, which accuracy can be due to its ability to navigate the semantically meaningful regions with its atrous convolution and spatial-pyramid pooling mechanism. Additionally, the feature concatenation from previous feature maps may have helped to compute more accurate maps for object semantic representation and hence segmentation.

The other competitor was PSPNet, which is also based on similar idea but on aggregating the global context information from different regions rather than the use of dilated convolutions. The computational speed for DeepLabv3+ with the same ResNet50 backbone as used in PSPNet in our experiments comes from the fact that the 1D separable convolutions and SPP network is used in DeepLabv3+. We evaluated the most recent popular SOTA method in segmentation ''HRNet'' [65]. While HRNet produced competitive results compared to other SOTA methods, UNet with ResNet34 backbone and DeepLabv3+ outperformed for most evaluation metrics with ColonSegNet being competitive in the recall, and overall accuracy and outperforming other SOTA method significantly.

Figure 5 shows an example for the 16 top scored and 16 bottom scored images on DSC for segmentation. From the results in Figure 5(c), it can be observed that there are polyps whose appearance under the given lighting conditions is very similar to healthy surrounding gastrointestinal skin texture. We suggest that including more samples with variable texture, different lighting conditions, and different angular views (refer to the samples in Figure 5(a) on the right, and (c)) can help to improve the DSC and other metrics of segmentation. We also observed that the presence of sessile or flat polyps were major limiting factors for algorithm robustness. Thus, including smaller polyps with respect to image size can help algorithm to generalise better thereby making these methods more usable for early detection of hard-to find polyps. In this regard, we also suggest the use of spatial pyramid layers to handle small polyps and using context-aware methods such as incorporation of artifacts or shape information to improve the robustness of these methods.

The possible limitation of the study is its retrospective design. Clinical studies are required for the validation of the approach in a real-world setting [72]. Additionally, in the presented study design we have resized the images, which can lead to loss of information and affect the algorithm performance. Moreover, we have optimized all the algorithms based on the empirical evaluation. Even though, optimal hyper-parameters have been set after experiments, we acknowledge that these can be further adjusted. Similarly, meta-learning approaches can be exploited to optimize the hyper-parameters that can work even in resource constraint settings.

VII. CONCLUSION

In this paper, we benchmark deep learning methods on the Kvasir-SEG dataset. We conducted thorough and extensive experiments for polyp detection, localization, and segmentation tasks and shown how different algorithms performs on variable polyp sizes and image resolutions. The proposed ColonSegNet detected and localised polyps at 180 frames per second. Similarly, ColonSegNet segmented polyps at the speed of 182.38 frames per second. The automatic polyp detection, localization, and segmentation algorithms showed good performance, as evidenced by high average precision, IoU, and FPS for the detection algorithm and DSC, IoU, precision, recall, F2-score, and FPS for the segmentation algorithm. While algorithms investigated in this paper show a clear strength to be used in clinical settings to help gastroenterologists for the polyp detection, localization, and segmentation task, computational scientists can build upon these methods to further improve in terms of accuracy, speed and robustness.

Additionally, the qualitative results provide insight for failure cases. This gives an opportunity to address the challenges present in the Kvasir-SEG dataset. Moreover, we have provided experimental results using well-established performance metrics along with the dataset for a fair comparison of the approaches. We believe that further data augmentation, fine tuning, and more advanced methods can improve the results. Additionally, incorporating artifacts [73] (e.g., saturation, specularity, bubbles, and contrast) issues can help improve the performance of polyp detection, localization, and segmentation. In the future, research should be more focused on designing even better algorithms for detection, localization, and segmentation tasks, and models should be build taking the number of parameters into consideration as required by most clinical systems.

ACKNOWLEDGMENT

Debesh Jha is funded by the Research Council of Norway project number 263248 (Privaton). The computations in this paper were performed on equipment provided by the Experimental Infrastructure for Exploration of Exascale Computing (eX3), which is financially supported by the Research Council of Norway under contract 270053. Parts of computational resources were also used from the research supported by the National Institute for Health Research (NIHR) Oxford BRC with additional support from the Wellcome Trust Core Award Grant Number 203141/Z/16/Z. Sharib Ali is supported by the NIHR Oxford Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. *(Debesh Jha and Sharib Ali contributed equally to this work.)*

REFERENCES

[1] J. Asplund, J. H. Kauppila, F. Mattsson, and J. Lagergren, ''Survival trends in gastric adenocarcinoma: A population-based study in Sweden,'' *Ann. Surgical Oncol.*, vol. 25, no. 9, pp. 2693–2702, Sep. 2018.

[2] Ø. Holme, M. Bretthauer, A. Fretheim, J. Odgaard-Jensen, and G. Hoff, ''Flexible sigmoidoscopy versus faecal occult blood testing for colorectal cancer screening in asymptomatic individuals,'' *Cochrane Database Systematic Rev.*, vol. 9. Munich, Germany: Zuckschwerdt, Oct. 2013.

[3] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. D. Lange, P. Halvorsen, and H. D. Johansen, ''ResUNet++: An advanced architecture for medical image segmentation,'' in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2019, pp. 225–2255.

[4] R. G. Holzheimer and J. A. Mannick, *Surgical Treatment: Evidence-Based Problem-Oriented*. 2001.

[5] J. Lee, ''Resection of diminutive and small colorectal polyps: What is the optimal technique?'' *Clin. Endoscopy*, vol. 49, no. 4, p. 355, 2016.

[6] P. L. Ponugoti, O. W. Cummings, and D. K. Rex, ''Risk of cancer in small and diminutive colorectal polyps,'' *Digestive Liver Disease*, vol. 49, no. 1, pp. 34–37, Jan. 2017.

[7] C. V. Tranquillini, W. M. Bernardo, V. O. Brunaldi, E. T. D. Moura, S. B. Marques, and E. G. H. D. Moura, "Best polypectomy technique for small and diminutive colorectal polyps: A systematic review and meta-analysis," *Arquivos de Gastroenterologia*, vol. 55, no. 4, pp. 358–368, Dec. 2018.
[8] O. Kronborg and J. Regula, "Population screening for colorectal cancer: Advantages and drawbacks," *Digestive Diseases*, vol. 25, no. 3, pp. 270–273, 2007.
[9] M. F. Kaminski, J. Regula, U. Wojciechowska, E. Kraszewska, M. Polkowski, J. Didkowska, M. Zwierko, M. Rupinski, M. P. Nowacki, and E. Butruk, "Quality indicators for colonoscopy and the risk of interval cancer," *New England J. Med.*, vol. 362, no. 19, pp. 1795–1803, May 2010.
[10] D. Castaneda, V. B. Popov, E. Verheyen, P. Wander, and S. A. Gross, "New technologies improve adenoma detection rate, adenoma miss rate, and polyp detection rate: A systematic review and meta-analysis," *Gastrointestinal Endoscopy*, vol. 88, no. 2, pp. 209–222, 2018.
[11] M. Matyja, A. Pasternak, M. Szura, M. Wysocki, M. Pędziwiatr, and K. Rembiasz, "How to improve the adenoma detection rate in colorectal cancer screening? Clinical factors and technological advancements," *Arch. Med. Sci., AMS*, vol. 15, no. 2, p. 424, 2019.
[12] M. Riegler, "Eir—A medical multimedia system for efficient computer aided diagnosis," Ph.D. dissertation, Dept. Inform., Univ. Oslo, Oslo, Norway, 2017.
[13] T. D. Lange, P. Halvorsen, and M. Riegler, "Methodology to develop machine learning algorithms to improve performance in gastrointestinal endoscopy," *World J. Gastroenterol.*, vol. 24, no. 45, p. 5057, 2018.
[14] Y. Shin, H. A. Qadir, and I. Balasingham, "Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance," *IEEE Access*, vol. 6, pp. 56007–56017, 2018.
[15] J. Y. Lee, J. Jeong, E. M. Song, C. Ha, H. J. Lee, J. E. Koo, D.-H. Yang, N. Kim, and J.-S. Byeon, "Real-time detection of colon polyps during colonoscopy using deep learning: Systematic validation with four independent datasets," *Sci. Rep.*, vol. 10, no. 1, pp. 1–9, Dec. 2020.
[16] P. Wang, X. Xiao, J. R. Glissen Brown, T. M. Berzin, M. Tu, F. Xiong, X. Hu, P. Liu, Y. Song, D. Zhang, X. Yang, L. Li, J. He, X. Yi, J. Liu, and X. Liu, "Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy," *Nature Biomed. Eng.*, vol. 2, no. 10, pp. 741–748, Oct. 2018.
[17] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. D. Lange, D. Johansen, and H. D. Johansen, "Kvasir-SEG: A segmented polyp dataset," in *Proc. Int. Conf. Multimedia Modeling (MMM)*, 2020, pp. 451–462.
[18] D. Jha, P. H. Smedsrud, D. Johansen, T. D. Lange, H. Johansen, P. Halvorsen, and M. Riegler, "A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation," *IEEE J. Biomed. Health Inform.*, early access, Jan. 5, 2021, doi: 10.1109/JBHI.2021.3049304.
[19] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. D. Lange, D. Johansen, C. Spampinato, D. T. Dang-Nguyen, M. Lux, P. T. Schmidt, and M. Riegler, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proc. 8th ACM Multimedia Syst. Conf.*, 2017, pp. 164–169.
[20] K. Pogorelov, K. R. Randel, T. D. Lange, S. L. Eskeland, C. Griwodz, D. Johansen, C. Spampinato, M. Taschwer, M. Lux, P. T. Schmidt, and M. Riegler, "Nerthus: A bowel preparation quality video dataset," in *Proc. ACM Multimedia Syst. Conf. (MMSys)*, 2017, pp. 170–174.
[21] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen, D. Johansen, C. Griwodz, H. K. Stensland, E. Garcia-Ceja, P. T. Schmidt, H. L. Hammer, M. A. Riegler, P. Halvorsen, and T. D. Lange, "HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy," *Sci. Data*, vol. 7, no. 1, pp. 1–14, Dec. 2020.
[22] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 9, no. 2, pp. 283–293, Mar. 2014.
[23] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. Saliency maps from physicians," *Computerized Med. Imag. Graph.*, vol. 43, pp. 99–111, Jul. 2015.
[24] P. H. Smedsrud, H. L. Gjestang, O. O. Nedrejord, E. Næss, V. Thambawita, S. Hicks, H. Borgli, D. Jha, T. J. Berstad, S. L. Eskeland, and M. Lux, "Kvasir-capsule, a video capsule endoscopy dataset," *Sci. Data*, 2021.
[25] S. Ali *et al.*, "Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 102002.
[26] D. Jha, S. Ali, K. Emanuelsen, S. A. Hicks, V. Thambawita, E. Garcia-Ceja, M. A. Riegler, T. D. Lange, P. T. Schmidt, H. D. Johansenm, D. Johansen, and P. Halvorsen, "Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy," in *Proc. Int. Conf. Multimedia Modeling (MMM)*, 2021, pp. 218–229.
[27] S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. A. Karras, and M. Tzivras, "Computer-aided tumor detection in endoscopic video using color wavelet features," *IEEE Trans. Inf. Technol. Biomed.*, vol. 7, no. 3, pp. 141–152, Sep. 2003.
[28] S. Ameling, S. Wirth, D. Paulus, G. Lacey, and F. Vilarino, "Texture-based polyp detection in colonoscopy," in *Bildverarbeitung für die Medizin 2009. Informatik aktuell*, H. P. Meinzer, T. M. Deserno, H. Handels, and T. Tolxdorff, Eds. Berlin, Germany: Springer, 2009, pp. 346–350, doi: 10.1007/978-3-540-93860-6_70.
[29] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, May 2016.
[30] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.
[31] J. Bernal *et al.*, "Comparative validation of polyp detection methods in video colonoscopy: Results from the MICCAI 2015 endoscopic vision challenge," *IEEE Trans. Med. Imag.*, vol. 36, no. 6, pp. 1231–1249, Jun. 2017.
[32] S. Ali *et al.*, "An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy," *Sci. Rep.*, vol. 10, no. 1, pp. 1–15, Dec. 2020.
[33] Y. Wang, W. Tavanapong, J. Wong, J. H. Oh, and P. C. D. Groen, "Polyp-alert: Near real-time feedback during colonoscopy," *Comput. Methods Programs Biomed.*, vol. 120, no. 3, pp. 164–179, Jul. 2015.
[34] Y. Shin, H. A. Qadir, L. Aabakken, J. Bergsland, and I. Balasingham, "Automatic colon polyp detection using region based deep CNN and post learning approaches," *IEEE Access*, vol. 6, pp. 40950–40962, 2018.
[35] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[36] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[37] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.
[38] M. Yamada, Y. Saito, H. Imaoka, M. Saiko, S. Yamada, H. Kondo, H. Takamaru, T. Sakamoto, J. Sese, A. Kuchiba, T. Shibata, and R. Hamamoto, "Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy," *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, Dec. 2019.
[39] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[40] Y. Guo and B. Matuszewski, "GIANA polyp segmentation with fully convolutional dilation neural networks," in *Proc. 14th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2019, pp. 632–641.
[41] S. Ali, F. Zhou, C. Daul, B. Braden, A. Bailey, S. Realdon, J. East, G. Wagnières, V. Loschenov, E. Grisan, W. Blondel, and J. Rittscher, "Endoscopy artifact detection (EAD 2019) challenge dataset," 2019, *arXiv:1905.03209*. [Online]. Available: http://arxiv.org/abs/1905.03209
[42] R. Wang, S. Chen, C. Ji, J. Fan, and Y. Li, "Boundary-aware context neural network for medical image segmentation," 2020, *arXiv:2005.00966*. [Online]. Available: http://arxiv.org/abs/2005.00966
[43] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "DoubleU-Net: A deep convolutional neural network for medical image segmentation," in *Proc. IEEE 33rd Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jul. 2020, pp. 558–564.

[44] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," 2020, *arXiv:2001.05566*. [Online]. Available: http://arxiv.org/abs/2001.05566

[45] M. Baldeon-Calisto and S. K. Lai-Yuen, "AdaResU-Net: Multiobjective adaptive convolutional neural network for medical image segmentation," *Neurocomputing*, vol. 392, pp. 325–340, Jun. 2020.

[46] N. Saeedizadeh, S. Minaee, R. Kafieh, S. Yazdani, and M. Sonka, "COVID TV-UNet: Segmenting COVID-19 chest CT images using connectivity imposed U-Net," 2020, *arXiv:2007.12303*. [Online]. Available: http://arxiv.org/abs/2007.12303

[47] Y. Meng, M. Wei, D. Gao, Y. Zhao, X. Yang, X. Huang, and Y. Zheng, "CNN-GCN aggregation enabled boundary regression for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2020, pp. 352–362.

[48] D. Vázquez, A. M. López, F. J. Sánchez, J. Bernal, A. Romero, G. Fernández-Esparrach, M. Drozdzal, and A. Courville, "A benchmark for endoluminal scene segmentation of colonoscopy images," *J. Healthcare Eng.*, vol. 2017, Jul. 2017, Art. no. 4037190.

[49] T. Roß *et al.*, "Comparative validation of multi-instance instrument segmentation in endoscopy: Results of the ROBUST-MIS 2019 challenge," *Med. Image Anal.*, vol. 70, Nov. 2020, Art. no. 101920.

[50] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[51] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[52] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 379–387.

[53] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 10781–10790.

[54] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*. [Online]. Available: http://arxiv.org/abs/1905.11946

[55] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[56] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[57] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*. [Online]. Available: http://arxiv.org/abs/2004.10934

[58] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[59] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 234–241.

[60] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[61] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[62] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[63] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[64] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[65] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 1, 2020, doi: 10.1109/TPAMI.2020.2983686.

[66] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[67] J. Hu, L. Shen, and G. Sun, "Squeeze- and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[68] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[69] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.

[70] F. Chollet *et al.*, "Keras," Tech. Rep., 2015.

[71] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and M. Kudlur, "Tensorflow: A system for large-scale machine learning," in *Proc. USENIX Symp. Operating Syst. Design Implement. (OSDI)*, 2016, pp. 265–283.

[72] Y. Mori *et al.*, "Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy: A prospective study," *Ann. Internal Med.*, vol. 169, no. 6, pp. 357–366, 2018.

[73] S. Ali, F. Zhou, A. Bailey, B. Braden, J. E. East, X. Lu, and J. Rittscher, "A deep learning framework for quality assessment and restoration in video endoscopy," *Med. Image Anal.*, vol. 68, Feb. 2021, Art. no. 101900.

DEBESH JHA received the master's degree in information and communication engineering from Chosun University, Gwangju, Republic of Korea. He is currently pursuing the Ph.D. degree with SimulaMet, Oslo, Norway, and UiT—The Arctic University of Norway, Tromsø, Norway. His research interests include computer vision, machine learning, deep learning, and medical image analysis.

SHARIB ALI received the Ph.D. degree from the University of Lorraine, France. He worked as a Postdoctoral Researcher at the Biomedical Computer Vision Group and the German Cancer research Center (DKFZ), University of Heidelberg, Heidelberg, Germany. He is currently working at the Department of Engineering Science, Institute of Biomedical Engineering, University of Oxford, Oxford, U.K. His research interests include computer vision and medical image analysis.

NIKHIL KUMAR TOMAR received the bachelor's degree in computer application from Indira Gandhi Open University, New Delhi, India. He is currently doing collaborative research at SimulaMet. His research interests include computer vision, artificial intelligence, parallel processing, and medical image segmentation.

HÅVARD D. JOHANSEN received the Ph.D. degree from the UiT—The Arctic University of Norway. He is currently a Professor with the Department of Informatics, UiT—The Arctic University of Norway. His major research interests include computing networks, cloud computing, network security, information security, and network architecture.

DAG JOHANSEN is currently a Full Professor with the Department of Computer Science, UiT—The Arctic of Norway. He is exploring interdisciplinary research problems at the intersection of sport science, medicine, and computer science. A usecase receiving special attention is elite soccer performance development and quantification technologies as basis for evidence-based decisions. His research interests include intervention technologies where privacy is a first-order concern and design principle.

JENS RITTSCHER received the Ph.D. degree from the University of Oxford, in 2001. He has worked extensively in the areas of video surveillance, the automatic annotation of video, and understanding of volumetric seismic data. He previously worked at the GE Global Research, Niskayuna, NY, USA, where he led the Computer Vision Laboratory. He is currently a Professor at with the Institute of Biomedical Engineering, Department of Engineering science, University of Oxford, Oxford, U.K. He acts as an elected member of the IEEE SPS Technical Committee on Bio Image and Signal Processing.

MICHAEL A. RIEGLER received the Ph.D. degree from the Department of Informatics, University of Oslo, Oslo, Norway, in 2015. He is currently working as a Chief Research Scientist at SimulaMet, Oslo, Norway. His research interests include machine learning, video analysis and understanding, image processing, image retrieval, crowdsourcing, social computing, and user intentions.

PÅL HALVORSEN is currently a Chief Research Scientist at SimulaMet, Oslo, Norway, a Full Professor with the Department of Computer Science, Oslo Metropolitan University, and an Adjunct Professor at the Department of Informatics, University of Oslo, Norway. His research interest includes distributed multimedia systems, including operating systems, processing, storage and retrieval, communication, and distribution.

A.6 Paper VI: NanoNet: Real-Time Polyp Segmentation in Endoscopy

Authors: D. Jha, N. K. Tomar, S. Ali, M. A. Riegler, H. D. Johansen, D. Johansen, and P. Halvorsen

Abstract: Deep learning in gastrointestinal endoscopy can assist to improve clinical performance and be helpful to assess lesions more accurately. To this extent, semantic segmentation methods that can perform automated real-time delineation of a region-of-interest, e.g., boundary identification of cancer or precancerous lesions, can benefit both diagnosis and interventions. However, accurate and real-time segmentation of endoscopic images is extremely challenging due to its high operator dependence and high-definition image quality. To utilize automated methods in clinical settings, it is crucial to design lightweight models with low latency such that they can be integrated with low-end endoscope hardware devices. In this work, we propose NanoNet, a novel architecture for the segmentation of video capsule endoscopy and colonoscopy images. Our proposed architecture allows real-time performance and has higher segmentation accuracy compared to other more complex ones. We use video capsule endoscopy and standard colonoscopy datasets with polyps, and a dataset consisting of endoscopy biopsies and surgical instruments, to evaluate the effectiveness of our approach. Our experiments demonstrate the increased performance of our architecture in terms of a trade-off between model complexity, speed, model parameters, and metric performances. Moreover, the resulting model size is relatively tiny, with only nearly 36,000 parameters compared to traditional deep learning approaches having millions of parameters.

Published: Proceedings of IEEE International Symposium on Computer-Based Medical Systems (CBMS), 2021.

Candidate contributions: D. Jha contributed to conceptualizing and designing the work. He also contributed to algorithm development and method implementation. He annotated, prepared, and released a new dataset and provided baseline results on this dataset. The dataset is made open access to the community. Additionally, he prepared and revised the manuscript with input from all of the co-authors. Finally, he presented the paper at the conference.

Thesis objectives: Objective I, Objective III

NanoNet: Real-Time Polyp Segmentation in Video Capsule Endoscopy and Colonoscopy

Debesh Jha*†, Nikhil Kumar Tomar*, Sharib Ali§‡‡, Michael A. Riegler*,
Håvard D. Johansen†, Dag Johansen†, Thomas de Lange¶‖**††, Pål Halvorsen*‡

*SimulaMet, Norway †UiT The Arctic University of Norway, Norway ‡Oslo Metropolitan University, Norway
§Institute of Biomedical Engineering, University of Oxford, Oxford, UK
¶Department of Medical Research, Bærum Hospital, Norway ‖Augere Medical AS, Norway
**Medical Department, Sahlgrenska University Hospital-Mölndal Hospital, Sweden
††Department of Molecular and Clinical Medicine, Sahlgrenska Academy, University of Gothenburg, Sweden
‡‡Oxford NIHR Biomedical Research Centre, Oxford, UK

***Abstract*—Deep learning in gastrointestinal endoscopy can assist to improve clinical performance and be helpful to assess lesions more accurately. To this extent, semantic segmentation methods that can perform automated real-time delineation of a region-of-interest, e.g., boundary identification of cancer or pre-cancerous lesions, can benefit both diagnosis and interventions. However, accurate and real-time segmentation of endoscopic images is extremely challenging due to its high operator dependence and high-definition image quality. To utilize automated methods in clinical settings, it is crucial to design lightweight models with low latency so that they can be integrated with low-end endoscope hardware devices. In this work, we propose *NanoNet*, a novel architecture for the segmentation of video capsule endoscopy and colonoscopy images. Our proposed architecture allows real-time performance and has higher segmentation accuracy compared to other more complex ones. We use video capsule endoscopy and standard colonoscopy datasets with polyps, and a dataset consisting of endoscopy biopsies and surgical instruments, to evaluate the effectiveness of our approach. Our experiments demonstrate the increased performance of our architecture in terms of a trade-off between model complexity, speed, model parameters, and metric performances. Moreover, the resulting models´ size is relatively tiny, with only nearly 36,000 parameters compared to traditional deep learning approaches having millions of parameters.**

***Index Terms*—Video capsule endoscopy, colonoscopy, deep learning, segmentation, tool segmentation**

I. Introduction

Gastrointestinal (GI) endoscopy is a widely used technique to diagnose and treat anomalies in the upper (esophagus, stomach, and duodenum) and the lower (large bowel and anus) GI tract. Among the other GI tract organs, colorectal cancer (CRC) has the highest cancer incidences and mortality rate [1]. There are several CRC screening options. Theses are usually divided into two categories, namely, invasive (visual examination-based test) and non-invasive based tests (stool, blood, and radiological test). *Colonoscopy*, the gold standard for examining the large bowel (colon and rectum), is an invasive examination used to detect, observe, and remove abnormalities (such as polyps). It detects colorectal cancer with both high sensitivity and specificity. *Sigmoidscopy* is another invasive test. *Computed Tomography(CT) Colonoscopy*, *Fecal Occult Blood Test (FOBT), Fecal Immunochemical Test (FIT)*, and Video Capsule Endoscopy (VCE) are non-invasive tests. VCE is a technology for capturing the video inside the GI tract. It has evolved as an important tool for detecting small bowel diseases [2].

Deep Learning (DL) methods have made a significant breakthrough in several medical domain such as lung cancer detection [3], diabetic retinopathy progression [4], and obstructive hypertrophic cardiomyopathy detection [5]. It has provided new opportunities to solve challenges such as bleeding, light over/underexposure, smoke, and reflections [6]. However, DL normally needs a large annotated dataset for the implementation of methods. It is difficult to obtain a labeled medical dataset. First, it needs collaborations with the hospitals. For data collection, the doctors require approval from various authorities and patient consent. They need to set protocols for the collection, and the collected data must be anonymized and cleaned with the help of data engineers. Domain experts must label raw data, and after labeling, the annotations must be done depending upon the need of the task. The whole process requires a significant amount of expert time and is costly. Additionally, it is an operator-dependent process. The quality of the data labeling and annotation depends on the expertise of the clinicians. Therefore, it is challenging to curate a larger dataset.

One way of solving the dataset issue is to create synthetic images using a Genearative Adversarial Network (GAN) [7]. However, generated synthetic images may not always capture all the properties and characteristics of real endoscopic images. Consequently, the model may only learn to predict the properties from the synthetic images and may not perform well on a real endoscopic dataset. Another solution could be domain adaptation from a similar endoscopic dataset. However, we lack large publicly available labeled endoscopic datasets. Thus, a viable and compelling approach to solve the semantic segmentation task is to reuse ImageNet pre-trained encoders in the segmentation model [8]. The predicted masks from the algorithm can provide reliable information to the endoscopic model.

A lightweight Convolutional Neural Network (CNN) model can be essential for the development of real-time and efficient semantic segmentation methods. Usually, lightweight models are computationally efficient and require less memory. A smaller number of parameters makes the network less redundant. Lightweight CNN models are mainly being deployed in mobile applications [9]. A lightweight model can play a crucial role from a system perspective with a limited resource constraint for real-time prediction in clinics. Consequently, we propose a novel architecture, NanoNet, optimized for faster inference and high accuracy. An extremely lightweight model with very few trainable parameters, faster inference, and higher performance would require less memory footprint to be incorporated with any devices. Therefore, we put forward this approach to address the challenges in endoscopy.

The main contributions of this work include the following:

1) We proposed a novel architecture, named NanoNet, to segment video capsule endoscopy and colonoscopy images in real-time with high accuracy. The proposed architecture is very lightweight, and the model size is relatively small, requiring less computational cost.
2) VCE datasets are difficult to obtain with pixel-wise annotations. In this context, we have annotated 55 polyps from the "polyp" class of the Kvasir-Capsule dataset with the help of an expert gastroenterologist. We have made this dataset public and provided the benchmark.
3) NanoNet achieves promising performance on the KvasirCapsule-SEG, Kvasir-SEG [10], 2020 Medico automatic polyp segmentation challenge [11], 2020 EndoTect challenge [12], and Kvasir-Instrument [13] datasets. All experiments conform with state-of-the-art (SOTA) in terms of parameter uses (size), speed, computation, and performance metrics.
4) The model can be integrated with mobile and embedded devices because of fewer parameters used in the network.

II. Related work

A. Semantic segmentation of endoscopic images

Semantic segmentation of endoscopic images has been a well-established topic in medical image segmentation. Earlier work mostly relied on the handcrafted descriptors for feature learning [14], [15]. The handcrafted features such as color, shape, texture, and edges were extracted and fed to the Machine Learning (ML) classifier, which separates lesions from the background. However, the traditional ML methods based on handcrafted features suffer from low performance [16]. The recent works on polyp segmentation using both video capsule endoscopy and colonoscopy mostly relied on Deep Neural Network (DNN) [17]–[23].

With the DNN methods, there is progress in the performance for segmenting endoscopic images (for example, polyps). However, the network architectures are often complex and requires high-end GPUs for training, and is computationally expensive [24], [25]. Additionally, real-time lesion segmentation has often been ignored. Although there is some recent initiation for the real-time detection of endoscopic images, they have mostly used private datasets [26]–[28] for the experimentation. It is difficult to compare the new methods on these datasets and extend the benchmark. Therefore, there is a need for a benchmark on publicly available datasets to minimize the research gap towards building a clinically relevant model.

B. Lightweight model

There are few works in the literature that have proposed lightweight models for image segmentation. Ni et al. [29] presented a novel bilinear attention network-based approach with an adaptive receptive field for the segmentation of surgical instruments. Wang et al. [30] proposed a lightweight encoder-decoder network (LEDNet), an encoder-decoder network that uses ResNet50 in the encoder block and attention pyramidal network in the decoder block. Beheshti et al. [31] proposed SqueezeNet. The architecture of the SqueezeNet is inspired by UNet [32]. The proposed model obtained a $12\times$ reduction in model size and showed efficient performance in multiplication accumulation (mac) and memory uses.

From the above-related work, we identify a need for a real-time polyp segmentation method. A real-time polyp segmentation method can be achieved by building a lightweight network architecture by designing an efficient network with blocks that require fewer parameters. A lower number of network parameters will reduce the network complexity, leading to real-time or faster inference. In this respect, we propose NanoNet, which uses a lightweight pre-trained network MobileNetV2 [33], and simple convolutional blocks such as residual block and squeeze and excite block.

III. Network architecture

The architecture of NanoNet follows an encoder-decoder approach as shown in Figure 1. As depicted in Figure 1, the network architecture uses a pre-trained model as an encoder, followed by the three decoder blocks. Using pre-trained ImageNet [34] models for transfer learning has become the best choice for many CNN architectures [8], [25]. It helps the model converge much faster and achieves high performance compared to the non-pre-trained model. The proposed architecture uses a MobileNetV2 [33] model pre-trained on the ImageNet [34] dataset as the encoder. The decoder is built using a modified version of the residual block, which was initially introduced by He et al. [35]. The encoder is used to capture the required contextual information from the input, whereas the decoder is used to generate the final output by using the contextual information extracted by the encoder.

A. MobileNetV2

The MobileNetV2 [33] is an architecture that is primarily designed for mobile and embedded devices. The architecture performed well on a variety of different datasets while maintaining high accuracy, despite having fewer parameters. The architecture of MobileNetV2 is based on the architecture of MobileNetV1, which uses depth-wise separable convolutions

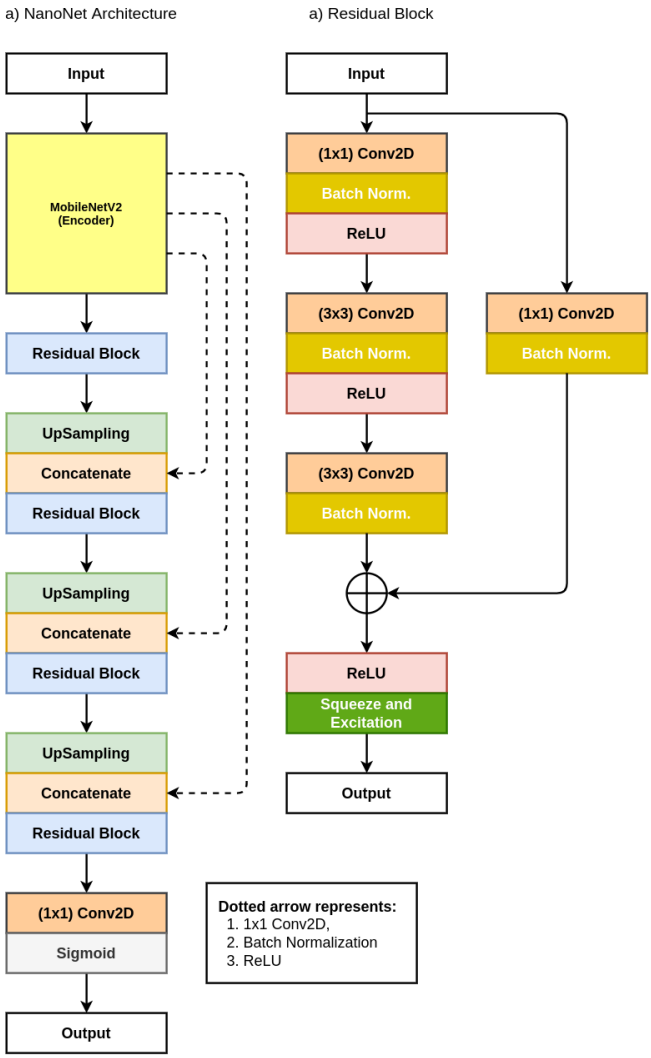


Fig. 1: Overview of the proposed NanoNet architecture

as the main building block. A depth-wise separable convolution consists of depth-wise convolution followed by a point-wise convolution. The MobileNetV2 introduces two main ideas: inverted residual block and linear bottleneck block [33].

The inverted residual block is based on the bottleneck residual block as described in [35], which consists of three standard convolutions, which are 1×1, 3×3, and 1×1. Every convolution layer is followed by a Rectified Linear Unit (ReLU) non-linearity. In the first 1×1 standard convolution, the number of feature channels are reduced, and in the last 1×1 standard convolution, the number of feature channels are expanded. After that, an element-wise addition with the identity mapping is performed. The inverted residual block also has three convolution layers: a 1×1 standard convolution, a 3×3 depth-wise convolution, and a 1×1 standard convolution. Every convolution has a ReLU activation function. Here, the exact opposite of the bottleneck residual block is performed. The first 1×1 standard convolution expands the number of feature channels, and the last 1×1 standard convolution reduces the number of feature channels. Due to this opposite functionality, it is referred to as an inverted residual block. The linear bottleneck block is the same as the inverted residual block, except the last 1×1 standard convolution has a linear activation before an element-wise addition is performed with the identity mapping.

B. Modified Residual Block

The original residual block uses two 3×3 standard convolutions, where the first convolution is followed by a batch-normalization and a ReLU activation function. After that, the second convolution is followed only by a batch-normalization. An element-wise addition is performed between the output of the batch-normalization and the identity mapping, followed by another ReLU activation function. An identity mapping consists of a 1×1 standard convolution and a batch-normalization over the original input.

We have modified the residual block for our network. The modified residual block starts with a 1×1 convolution followed by a 3×3 convolution. In both of these convolutions, we reduce the number of filters by $\frac{1}{4}$, which are then followed by the batch normalization and the ReLU activation function. We have a 3×3 convolution with batch normalization. Now, we perform an element-wise addition with the identity mapping. Finally, we apply a ReLU activation function followed by the squeeze and excitation block. The squeeze and excitation block improves the quality of feature maps by increasing their sensitivity towards essential features.

C. The NanoNet architecture

Figure 1 shows the block diagram of the NanoNet architecture. The NanoNet architecture starts with a pre-trained MobileNetV2 as an encoder followed by a decoder. There is a modified residual block between the encoder and the decoder, which acts like a bridge that connects the encoder and the decoder. In the first step, we feed the image data into the pre-trained encoder. The pre-trained encoder starts with a standard convolution with 32 feature channels, followed by the bottleneck layer with ReLU6 as the activation function. All the convolution operations use a standard 3×3 kernel size. The entire encoder network progressively downsamples the feature maps by using strided convolution and slowly increases the number of feature channels alternatively.

The output from the pre-trained encoder passes through the modified residual block, which is fed to the decoder. Every step in the decoder uses a bilinear upsampling to increase the spatial dimension (height and width) of the input feature maps. After that, it is concatenated with the appropriate feature maps from the pre-trained encoder using the skip connections. These skip connections pass information that may be lost sometimes between the layers and are used to improve the quality of the feature maps. These concatenated feature maps are passed through the modified residual block, which further increases the generalization capacity of the decoder. After the feature maps pass through all the three decoder blocks, the output of the last decoder block is fed to a 1×1 convolution with a number of classes as the feature channels. This is followed by

TABLE I: Publicly available endoscopic datasets used in our experiments

Dataset	No. of Images	Imaging Type	Availability
KvasirCapsule-SEG	55	Video capsule endoscopy	https://www.dropbox.com/sh/hr46vieykbmvmkk/AAAs_V8ECG0wq51Fpw3rYU_5a?dl=0
Kvasir-SEG [10]	1000	Colonoscopy	https://datasets.simula.no/kvasir-seg/
2020 Medico automatic polyp segmentation challenge [11]	160$^{\diamond}$	Colonoscopy	https://multimediaeval.github.io/editions/2020/tasks/medico/
Endotect Challenge Dataset [12]	200$^{\diamond}$	Colonoscopy	https://endotect.com/
Kvasir-Instrument [36]	590	Colonoscopy	https://datasets.simula.no/kvasir-instrument/

$^{\diamond}$test images

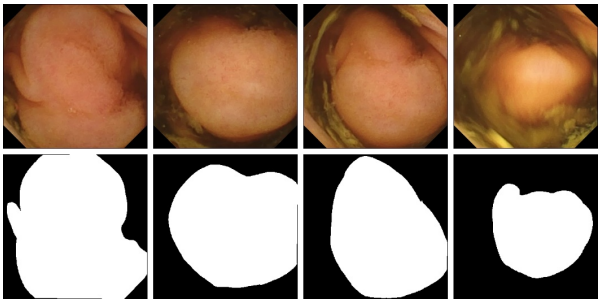


Fig. 2: Polyps and corresponding masks from KvasirCapsule-SEG

the sigmoid activation if it is a binary segmentation task, else we use the softmax activation function.

We have investigated three different NanoNet architectures: NanoNet-A, NanoNet-B, and NanoNet-C. Each architecture consists of different feature channels in its decoder block. NanoNet-A consists of 32, 64 and 128 feature channels. In NanoNet-B, the number of feature channels is reduced to 32, 64, and 96. In NanoNet-C, these feature channels are further reduced to 16, 24, and 32. The reduction in the number of feature channels leads to less trainable parameters, which simplifies the model complexity leading to a light-weight network.

IV. Experimental setup

In this section, we will describe the dataset, evaluation metrics, implementation details, and data augmentation techniques used.

A. Datasets

To address the polyp segmentation problem from video capsule endoscopy images, we have selected the polyp class from the labelled images folder of the Kvasir-Capsule dataset [37] and annotated it with the help of an expert gastroenterologist. The Kvasir-Capsule is an open-access dataset that contains 13 classes of labelled anomalies and findings. It only includes 55 polyp frames out of 44,228 medically verified video capsule frames present in the Kvasir-Capsule. We have annotated the polyp class of Kvasir-Capsule and generated corresponding ground truth masks. Examples of polyps and their corresponding masks from KvasirCapsule-SEG can be found in Figure 2. Furthermore, we also provide bounding box information to be used for video capsule endoscopy detection and localization tasks. The Kvasir-Capsule can be downloaded from here [1] and KvasirCapsule-SEG can be downloaded from here [2].

Table I shows the detailed information about the open imaging dataset used in our experiments. Each of the datasets presented in Table I also has the corresponding ground truth. The link for each of the datasets is provided in the table. The standard setting for the "Medico automatic polyp segmentation challenge" and "Endotect challenge" is that they use the Kvasir-SEG for training. The challenge organizers have provided unseen 160 images in the "Medico automatic polyp segmentation challenge" and released 200 images in the "Endotect challenge" to test the participant's approaches. For the Kvasir-instrument dataset, we experimented with the official split provided by the organizers. The detail explanation of these datasets and the baseline results can be found in [10]–[13].

B. Evaluation metrics

For evaluation purposes, we have chosen standard computer vision metrics such as Dice Coefficient (DSC), mean Intersection over Union (mIoU), Precision, Recall, Specificity, Accuracy, and Frame-per-second (FPS). More explanation of these metrics can be found in [10]–[13].

C. Implementation details

We have implemented the NanoNet using Keras[3] with TensorFlow [40] as backend. The experiments were run on the Experimental Infrastructure for Exploration of Exascale Computing (eX3), NVIDIA DGX-2 machine. The code implementation of NanoNet can be found here[4]. As the model has very few low trainable parameters, we have set a batch size of 16. We have resized the dataset images to 256×256 pixels for better utilization of the GPU, and it also helps to reduce the training time. The model is trained on 200 epochs with the Nadam optimizer [41] and dice coefficient as the loss function. The learning rate for the optimizer is set to $1e^{-4}$. We prefer to choose a low learning rate to update the parameters slowly and carefully. The learning rate is reduced by a factor of 0.1 when the validation loss does not decrease in 10 consecutive epochs. It helps to improve model performance. Additionally, we have used an early stopping mechanism to prevent over-fitting.

[1] https://osf.io/dv2ag/

[2] https://www.dropbox.com/sh/hr46vieykbmvmkk/AAAs_V8ECG0wq51Fpw3rYU_5a?dl=0

[3] https://keras.io/

[4] https://github.com/DebeshJha/NanoNet

TABLE II: Performance evaluation of the proposed networks and recent SOTA methods on KvasirCapsule-SEG

Method	Parameters	DSC	mIoU	Recall	Precision	F2	Accuracy	FPS
ResUNet (GRSL'18) [38]	8,227,393	**0.9532**	**0.9137**	**0.9785**	**0.9325**	**0.9677**	**0.9386**	17.96
ResUNet++ (ISM'19) [24]	4,070,385	0.9499	0.9087	0.9762	0.9296	0.9648	0.9334	15.39
NanoNet-A (Ours)	235,425	0.9493	0.9059	0.9693	**0.9325**	0.9609	0.9351	28.35
NanoNet-B (Ours)	132,049	0.9474	0.9028	0.9682	0.9308	0.9593	0.9324	27.39
NanoNet-C (Ours)	36,561	0.9465	0.9021	0.9754	0.9238	0.9629	0.9297	**29.48**

TABLE III: Performance evaluation of the proposed networks and recent SOTA methods on Kvasir-SEG [10]

Method	Parameters	DSC	mIoU	Recall	Precision	F2	Accuracy	FPS
ResUNet (GRSL'18) [38]	8,227,393	0.7203	0.6106	0.7602	0.7624	0.7327	0.9251	17.72
ResUNet++ (ISM'19) [24]	4,070,385	0.7310	0.6363	0.7925	0.7932	0.7478	0.9223	19.79
NanoNet-A (Ours)	235,425	**0.8227**	**0.7282**	**0.8588**	**0.8367**	**0.8354**	**0.9456**	26.13
NanoNet-B (Ours)	132,049	0.7860	0.6799	0.8392	0.8004	0.8067	0.9365	29.73
NanoNet-C (Ours)	36,561	0.7494	0.6360	0.8081	0.7738	0.7719	0.9290	**32.17**

TABLE IV: Performance evaluation of the proposed networks and recent SOTA methods on the Medico 2020 dataset [11]

Method	Parameters	DSC	mIoU	Recall	Precision	F2	Accuracy	FPS
ResUNet (GRSL'18) [38]	8,227,393	0.6846	0.5599	0.7235	0.7236	0.6961	**0.9231**	18.54
ResUNet++ (ISM'19) [24]	4,070,385	0.6925	0.5849	0.8249	0.6840	0.7434	0.8995	19.47
NanoNet-A (Ours)	235,425	0.7364	**0.6319**	**0.8566**	0.7310	**0.7804**	0.9166	28.07
NanoNet-B (Ours)	132,049	**0.7378**	0.6247	0.8283	**0.7373**	0.7685	0.9223	29.04
NanoNet-C (Ours)	36,651	0.7070	0.5866	0.8095	0.7089	0.7432	0.9148	**32.66**

TABLE V: Performance evaluation of the proposed networks and recent SOTA methods on the Endotect 2020 dataset [12]

Method	Parameters	DSC	mIoU	Recall	Precision	F2	Accuracy	FPS
ResUNet (GRSL'18) [39]	8,227,393	0.6640	0.5408	0.7510	0.6841	0.6943	0.9075	26.55
ResUNet++ (ISM'19) [24]	4,070,385	0.6940	0.5838	**0.8797**	0.6591	0.7597	0.8841	18.58
NanoNet-A (Ours)	235,425	**0.7508**	**0.6466**	0.8238	**0.7744**	**0.7773**	**0.9255**	27.19
NanoNet-B (Ours)	132,049	0.7362	0.6238	0.8109	0.7532	0.7646	0.9252	29.91
NanoNet-C (Ours)	36,651	0.7001	0.5792	0.8000	0.7159	0.7380	0.9091	**32.98**

TABLE VI: Performance evaluation of the proposed networks and recent SOTA methods on Kvasir-Instrument [13]

Method	Parameters	DSC	mIoU	Recall	Precision	F2	Accuracy	FPS
UNet (Baseline) [39]	-	0.9158	0.8578	**0.9487**	0.8998	**0.9320**	0.9864	20.46
DoubleUNet (Baseline) [25]	-	0.9038	0.8430	0.9275	0.8966	0.9147	0.9838	10.00
ResUNet++ (ISM'19) [24]	4,070,385	0.9140	0.8635	0.9103	0.9348	0.9140	0.9866	17.87
NanoNet-A (Ours)	235,425	0.9251	0.8768	0.9142	**0.9540**	0.9251	**0.9887**	28.00
NanoNet-B (Ours)	132,049	**0.9284**	**0.8790**	0.9205	0.9482	0.9284	0.9875	29.82
NanoNet-C (Ours)	36,561	0.9139	0.8600	0.9037	0.9452	0.9139	0.9863	**32.18**

D. Data augmentation

We use data-augmentation on the training set to increase diversity and to improve the generalization of our model. Data augmentation techniques such as random cropping, random rotation, horizontal flipping, vertical flipping, grid distortion, and many more are used. We have used an offline data augmentation technique. The validation and testing set is not augmented and is directly resized into 256×256.

V. Result and Discussion

In this section, we provide the experimental results for the segmentation task of the endoscopic image dataset. We have used performance metrics such as DSC and mIoU, and FPS as the main evaluation metrics. We also calculate recall, precision, F2, and overall accuracy to support a complete set of metrics. Table II, Table III, Table IV, Table V, and Table VI show the results of the NanoNet model experiments using different parameters. The results are compared with the recent SOTA computer vision methods.

The quantitative results in these tables show that NanoNet consistently outperforms or performs nearly equal to its competitors in terms of performance. The quantitative results also show that NanoNet can produce real-time segmentation (i.e., produces at least close to 30 FPS for each dataset present in the Tables). This is one of the major contributions of the work. The other strength of the work lies in the parameter use. From Table II, we can observe that the best performing NanoNet (i.e., NanoNet-A) uses nearly 35 times less parameters as ResUNet [38]. Similarly, NanoNet-C uses 225 times less parameters as compared to that of ResUNet and also produces better DSC, mIoU and FPS with the Kvasir-SEG.

The qualitative results are displayed in Figure 3. The first, second, and third columns show the image, ground truth, and prediction masks, respectively. Similarly, the name of the dataset is provided on the left side. One example image for each dataset is shown. The qualitative results with diversified classes of medical datasets show that NanoNet can produce accurate segmentation results with different types of lesions

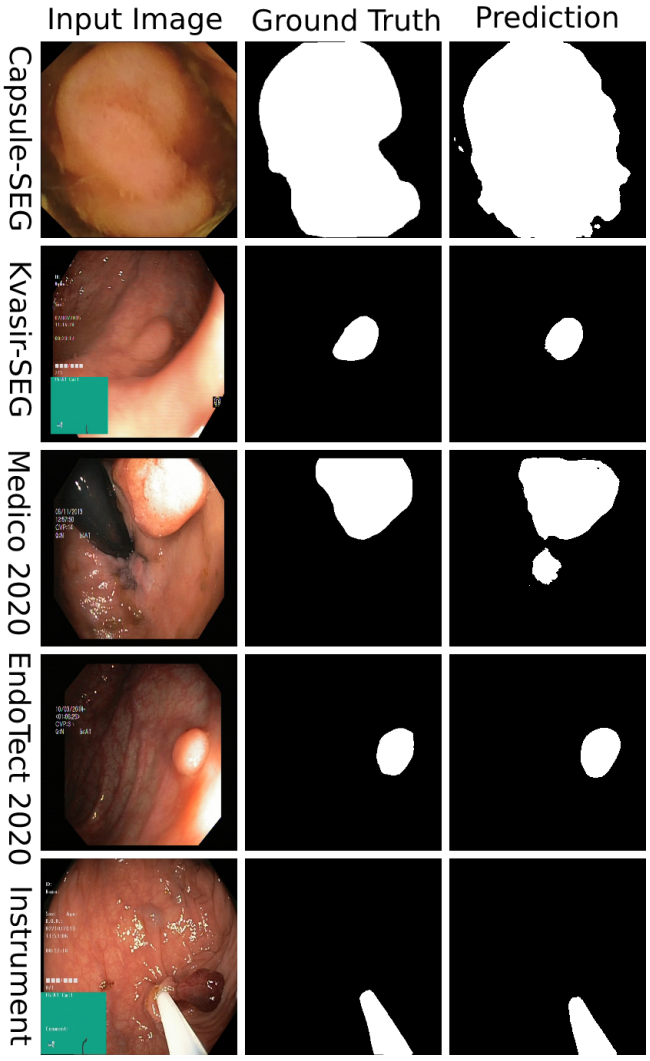


Fig. 3: Qualitative results of NanoNet-A on five different datasets

(polyps) and therapeutic tools. The example images and the prediction also show that NanoNet produces good segmentation masks for large, medium, and small polyps (see Figure 3). From the qualitative results, we can derive and conclude that NanoNet produces good results with small-sized polyps but produces over-segmentation for the large-sized lesions upon detail dissection. For future work, one could create a specific dataset consisting of a set of small and large-sized polyps to explore this further.

From both evaluation metrics and qualitative results, the improvement is remarkable. Thus, the proposed NanoNet architecture is simple, compact, and provides a robust solution for real-time applications, as it produces satisfactory performance despite having fewer parameters.

VI. Conclusion

In this paper, we proposed a novel lightweight architecture for real-time video capsule endoscopy and colonoscopy image segmentation. The proposed NanoNet architecture utilizes a pre-trained MobileNetV2 model and a modified residual block. The depthwise separable convolution is the main building block of the network and allows the model to achieve high performance with minuscule trainable parameters. The experimental results on varied endoscopy datasets demonstrate the strength of our model compared to SOTA models with respect to their speed and performance. The presented model has the potential to enable easier roll out of deep learning models in clinical systems due to fewer parameters, competitive accuracy, and low-latency. In addition, the model does not require any sort of initialization, post-processing, or temporal regularization, considered as another strength of this work. In the future, we will design an encoder lighter than the currently used pre-trained MobilNetV2. Moreover, we aspire to utilize the currently built segmentation module in the clinic and study the efficacy of our designed model.

Acknowledgment

The research is partially funded by the PRIVATON project (263248) and the Autocap project (282315) from the Research Council of Norway (RCN). Our experiments were performed on the Experimental Infrastructure for Exploration of Exascale Computing (eX3) system, which is financially supported by RCN under contract 270053.

References

[1] H. Sung *et al.*, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, 2021.

[2] A. Kornbluth, P. Legnani, and B. S. Lewis, "Video capsule endoscopy in inflammatory bowel disease: past, present, and future," *Inflammatory Bowel Diseases*, vol. 10, no. 3, pp. 278–285, 2004.

[3] D. Ardila *et al.*, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature medicine*, vol. 25, no. 6, pp. 954–961, 2019.

[4] F. Arcadu *et al.*, "Deep learning algorithm predicts diabetic retinopathy progression in individual patients," *NPJ digital medicine*, vol. 2, no. 1, pp. 1–9, 2019.

[5] E. M. Green *et al.*, "Machine learning detection of obstructive hypertrophic cardiomyopathy using a wearable biosensor," *NPJ digital medicine*, vol. 2, no. 1, pp. 1–4, 2019.

[6] S. Bodenstedt *et al.*, "Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery," *arXiv preprint arXiv:1805.02475*, 2018.

[7] I. J. Goodfellow *et al.*, "Generative adversarial networks," *arXiv preprint arXiv:1406.2661*, 2014.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[9] Y.-D. Kim *et al.*, "Compression of deep convolutional neural networks for fast and low power mobile applications," *arXiv preprint arXiv:1511.06530*, 2015.

[10] D. Jha *et al.*, "Kvasir-seg: A segmented polyp dataset," in *Proc. of International Conference on Multimedia Modeling (MMM)*, 2020, pp. 451–462.

[11] D. Jha, S. A. Hicks, K. Emanuelsen, H. Johansen, D. Johansen, T. de Lange, M. A. Riegler, and P. Halvorsen, "Medico multimedia task at mediaeval 2020: Automatic polyp segmentation," in *CEUR Proceedings of MediaEval Workshop*, 2020.

[12] S. A. Hicks *et al.*, "The endotect 2020 challenge: Evaluation and comparison of classification, segmentation and inference time for endoscopy," in *Proceedings of ICPR 2020 Workshops and Challenges*, 2020.

[13] D. Jha *et al.*, "Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy," in *Proc. of Multimedia Modeling (MMM)*, 2021.

[14] S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. A. Karras, and M. Tzivras, "Computer-aided tumor detection in endoscopic video using color wavelet features," *IEEE transactions on information technology in biomedicine*, vol. 7, no. 3, pp. 141–152, 2003.

[15] S. Ameling, S. Wirth, D. Paulus, G. Lacey, and F. Vilarino, "Texture-based polyp detection in colonoscopy," in *Bildverarbeitung für die Medizin 2009*, 2009, pp. 346–350.

[16] J. Bernal, J. Sánchez, and F. Vilarino, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognition*, vol. 45, no. 9, pp. 3166–3182, 2012.

[17] X. Jia, X. Xing, Y. Yuan, L. Xing, and M. Q.-H. Meng, "Wireless capsule endoscopy: A new tool for cancer screening in the colon with deep-learning-based polyp recognition," *Proceedings of the IEEE*, vol. 108, no. 1, pp. 178–197, 2019.

[18] V. Prasath, "Polyp detection and segmentation from video capsule endoscopy: A review," *Journal of Imaging*, vol. 3, no. 1, p. 1, 2017.

[19] N. K. Tomar *et al.*, "Fanet: A feedback attention network for improved biomedical image segmentation," *arXiv preprint arXiv:2103.17235*, 2021.

[20] S. Ali *et al.*, "Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy," *Medical Image Analysis*, p. 102002, 2021.

[21] D. Jha *et al.*, "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning," *Ieee Access*, vol. 9, pp. 40 496–40 510, 2021.

[22] Y. Guo, J. Bernal, and B. J Matuszewski, "Polyp segmentation with fully convolutional deep neural networks—extended evaluation study," *Journal of Imaging*, vol. 6, no. 7, p. 69, 2020.

[23] D. Jha *et al.*, "A Comprehensive Study on Colorectal Polyp Segmentation with ResUNet++, Conditional Random Field and Test-Time Augmentation," *IEEE Journal of Biomedical and Health Informatics*.

[24] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. D. Johansen, "ResUNet++: An Advanced Architecture for Medical Image Segmentation," in *Proc. of IEEE International Symposium on Multimedia (ISM)*, 2019, pp. 225–2255.

[25] D. Jha, , M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "DoubleU-Net: A Deep Convolutional Neural Network for Medical Image Segmentation," in *Proc. of International Conference on Multimedia Modeling (MMM)*, 2020, pp. 451–462.

[26] J. Y. o. Lee, "Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets," *Scientific reports*, vol. 10, no. 1, pp. 1–9, 2020.

[27] M. Yamada *et al.*, "Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy," *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.

[28] C. C. Poon *et al.*, "Ai-doscopist: a real-time deep-learning-based algorithm for localising polyps in colonoscopy videos with edge computing devices," *NPJ Digital Medicine*, vol. 3, no. 1, pp. 1–8, 2020.

[29] Z.-L. Ni *et al.*, "Barnet: Bilinear attention network with adaptive receptive field for surgical instrument segmentation," *arXiv preprint arXiv:2001.07093*, 2020.

[30] Y. Wang *et al.*, "Lednet: A lightweight encoder-decoder network for real-time semantic segmentation," in *Proc. of IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 1860–1864.

[31] N. Beheshti and L. Johnsson, "Squeeze u-net: A memory and energy efficient image segmentation network," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020, pp. 364–365.

[32] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2017.

[33] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. of IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[34] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *Proc. of IEEE conference on computer vision and pattern recognition (CVPR)*, 2009, pp. 248–255.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 770–778.

[36] J. Bernal *et al.*, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.

[37] P. H. Smedsrud *et al.*, "Kvasir-capsule, a video capsule endoscopy dataset," *Springer Nature Scientific Data*, 2021.

[38] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[39] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. of International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, 2015, pp. 234–241.

[40] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proc. of USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2016, pp. 265–283.

[41] T. Dozat, "Incorporating nesterov momentum into adam," in *Proc. of International Conference on Learning Representations*, 2016.

A.7 Paper VII : Kvasir-Instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy

Authors: D. Jha, S. Ali, K. Emanuelsen, S. A. Hicks, E. Garcia-Ceja, M. A. Riegler, T. d. Lange, P. T Schmidt, H. Johansen, D. Johansen and P. Halvorsen

Abstract: Gastrointestinal (GI) pathologies are periodically screened, biopsied, and resected using surgical tools. Usually, the procedures and the treated or resected areas are not specifically tracked or analysed during or after colonoscopies. Information regarding disease borders, development, amount, and size of the resected area get lost. This can lead to poor follow-up and bothersome reassessment difficulties post-treatment. To improve the current standard and also to foster more research on the topic, we have released the "Kvasir-Instrument" dataset, which consists of 590 annotated frames containing GI procedure tools such as snares, balloons, and biopsy forceps, etc. Besides the images, the dataset includes ground truth masks and bounding boxes and has been verified by two expert GI endoscopists. Additionally, we provide a baseline for the segmentation of the GI tools to promote research and algorithm development. We obtained a dice coefficient score of 0.9158 and a Jaccard index of 0.8578 using a classical U-Net architecture. A similar dice coefficient score was observed for DoubleUNet. The qualitative results showed that the model did not work for the images with specularity and the frames with multiple tools, while the best result for both methods was observed on all other types of images. Both qualitative and quantitative results show that the model performs reasonably good, but there is potential for further improvements. Benchmarking using the dataset provides an opportunity for researchers to contribute to the field of automatic endoscopic diagnostic and therapeutic tool segmentation for GI endoscopy.

Published: Proceedings of International Conference on Multimedia Modeling (MMM 2021)

Candidate contributions: D. Jha contributed to conceptualizing and design of the work. He collected, annotated, and prepared the dataset with the help of two expert gastroenterologists. Additionally, he also performed all the experiments and analyzed the results. He wrote and revised the manuscript with input from all the

authors. He also released the dataset on the webpage and made it publicly available for research and academic purposes.

Thesis objectives: Objective I

Kvasir-Instrument: Diagnostic and Therapeutic Tool Segmentation Dataset in Gastrointestinal Endoscopy

Debesh Jha[1,2( )], Sharib Ali[9], Krister Emanuelsen[3], Steven A. Hicks[1,5], Vajira Thambawita[1,5], Enrique Garcia-Ceja[10], Michael A. Riegler[1], Thomas de Lange[4,6,7], Peter T. Schmidt[8], Håvard D. Johansen[2], Dag Johansen[2], and Pål Halvorsen[1,5]

[1] SimulaMet, Oslo, Norway
debesh@simula.no
[2] UIT The Arctic University of Norway, Tromsø, Norway
[3] Simula Research Laboratory, Oslo, Norway
[4] Augere Medical AS, Oslo, Norway
[5] Oslo Metropolitan University, Oslo, Norway
[6] Medical Department, Sahlgrenska University Hospital-Mölndal, Gothenburg, Sweden
[7] Department of Medical Research, Bærum Hospital, Gjettum, Norway
[8] Karolinska University Hospital, Solna, Sweden
[9] Department of Engineering Science, University of Oxford, Oxford, UK
[10] Sintef Digital, Oslo, Norway

Abstract. Gastrointestinal (GI) pathologies are periodically screened, biopsied, and resected using surgical tools. Usually, the procedures and the treated or resected areas are not specifically tracked or analysed during or after colonoscopies. Information regarding disease borders, development, amount, and size of the resected area get lost. This can lead to poor follow-up and bothersome reassessment difficulties post-treatment. To improve the current standard and also to foster more research on the topic, we have released the "Kvasir-Instrument" dataset, which consists of 590 annotated frames containing GI procedure tools such as snares, balloons, and biopsy forceps, etc. Besides the images, the dataset includes ground truth masks and bounding boxes and has been verified by two expert GI endoscopists. Additionally, we provide a baseline for the segmentation of the GI tools to promote research and algorithm development. We obtained a dice coefficient score of 0.9158 and a Jaccard index of 0.8578 using a classical U-Net architecture. A similar dice coefficient score was observed for DoubleUNet. The qualitative results showed that the model did not work for the images with specularity and the frames with multiple tools, while the best result for both methods was observed on all other types of images. Both qualitative and quantitative results show that the model performs reasonably good, but there is potential for further improvements. Benchmarking using the dataset provides an opportunity for researchers to contribute to the field of automatic endoscopic diagnostic and therapeutic tool segmentation for GI endoscopy.

J. Lokoč et al. (Eds.): MMM 2021, LNCS 12573, pp. 218–229, 2021.
https://doi.org/10.1007/978-3-030-67835-7_19

Keywords: Gastrointestinal endoscopy · Tool segmentation · Endoscopic tools · Convolutional neural network · Benchmarking

1 Introduction

Minimally Invasive Surgery (MIS) is a commonly used technique in surgical procedures. The advantage of MIS is that small surgical incisions are made in the patient for endoscopy that causes less pain, reduced time of the hospital stay, fast recovery, reduced blood loss, and less scaring process as compared to the traditional open surgery. The nature of the operation is complex, and the surgeons have to precisely tackle hand-eye coordination, which may lead to restricted mobility and a narrow field of view [5].

However, unlike the treatment of accessory organs such as liver and pancreas, no incision is required forGastrointestinal (GI) tract organs (*oesophagus, stomach, duodenum, colon, and rectum*). GI procedures also include both minimally invasive surveillance and treatment (*including surgery*) procedures. A varied number of tools are used as per the requirement of these procedures. For example, balloon dilatation to help open the GI surface, biopsy forceps for tissue sample collection, polyp removal with snares, and submucosal injections.

A computer and robotic-assisted surgical system can enhance the capability of the surgeons [9]. It can provide the opportunity to gain additional information about the patient, which can be useful for decision making during surgery [6]. However, it is difficult to understand the spatial relationship between surgical instruments, cameras, and anatomy for the patient [11]. In GI endoscopy, it is vital to track and guide surgeons during tumor resection or biopsy collection from a defined site and help to correlate the biopsied samples and treatment locations post-diagnostic and therapeutic or surgical procedures. While most datasets and automated-algorithm developments for instrument segmentation are mostly focused on laparoscopy-based surgical removal, automatic guidance of tools for GI surgery has not been addressed before.

New developments in the area of robot-assisted systems show that there is potential for developing a fully automated robotic surgeon [14]. The da Vinci robot is a surgical system that is considered the de-facto standard-of-care for certain urological, gynecological, and general procedures [4]. Thus, it is critical to have information regarding intra-operative guidance, which plays an essential role in decision making. However, there are specific challenges, such as limited field of view and difficulties with the surgeons handling the instruments during surgery [13]. Therefore, image-based instrument segmentation and tracking are gaining more and more attention in both robotic and non-robotic minimally invasive surgery. Previous work targeting instrument segmentation, detection, and tracking on endoscopic video images failed on challenging images such as images with blood, smoke, and motion artifacts [13]. Other reasons that make semantic segmentation of surgical instruments a challenging task are the presence of images containing shadows, specular reflections, blood, camera lens fogging, and the complex background tissue [14]. The segmentation masks of these images can be useful for instrument detection and tracking.

Similarly, in the GI tract procedures, from tissue sample collection to surgical removal of pathologies is performed in low field-of-view areas. Visual clutter such as artifacts, moving objects, and fluid, hinders the localisation of the target site during surgical procedures. Additionally, currently, there is no way of correlating the tissue sample collection with biopsied location and assessing surgical procedure effectiveness or even post-treatment recovery analysis. Automated localisation and tracking of tools can help guide the endoscopists and surgeons to perform their tasks more effectively. Also, post-procedure video analysis can be done using these automated methods to track such tools, thus enabling improved surgical procedures or surveillance and their post-assessment. Currently, this is an open problem in the research community, where most procedures are not automated in GI tract endoscopy.

While there is an open research question for automated tool detection and guidance in GI procedures, there is a lack of available public datasets. We aim to initiate the development of automated systems for the segmentation of GI tract diagnostic and therapeutic endoscopy tools. This research direction will enable tracking and localisation of essential tools used in endoscopy and help to improve targeted biopsies and surgeries in complex GI tract organs. To accomplish this, and to address the lack of publicly available labeled datasets, we have publicly released 590 pixel-level annotated frames that comprise of tools such as balloon dilation for facilitating the opening of GI organs, biopsy forceps for tissue sample collection, polyp removal with snares, submucosal injections, radiofrequency ablation of dysplastic mucosa using probes and some other related surgical/diagnostic procedures. The released video frames will allow for building automated Machine Learning (ML) algorithms that can be applied during clinical procedures or post-analyses. To commence this effort, we provide a baseline benchmark on this dataset. U-Net [12] is a common semantic segmentation based architecture for medical image segmentation tasks. In this paper, we therefore present results utilising two U-Net based architectures. The provided dataset is open and can be used for research and development, and we invite medical imaging, computer vision, ML and multimedia researchers to develop novel algorithms on the provided dataset. The main contributions of this paper are:

- The release of 590 annotated images with bounding boxes and segmentation masks of GI diagnostic and surgical tool dataset. To the best of our knowledge, this is the first dataset of segmented tools used in the GI tract.
- A benchmark of the provided dataset using the U-Net [12] and DoubleUNet [10] architectures for semantic segmentation is provided.

2 Related Work

Surgical vision is evolving as a promising technique to segment and track instruments using endoscopic images [6]. To gather researchers on a single platform, the *Endoscopic vision (EndoVis) challenge* has been organized since 2015 at Medical Image Computing and Computer Assisted Intervention Society (MICCAI)

Table 1. Similar available datasets

Dataset	Content	Task type	Procedure
Instrument segmentation and tracking (2015) [6]	Rigid and robotic instruments	Segmentation and tracking	Laparoscopy
Robotic Instrument Segmentation (2017) [4]	Robotic surgical instruments	Binary segmentation, part based segmentation, instrument segmentation	Abdominal porcine
Robotic Scene Segmentation (2018) [3]	Surgical instruments and other	Multi-instance segmentation	Robotic nephrectomy
Robust Medical instrument segmentation (2019) [13]	Laparoscopic instrument	Binary segmentation, multiple instance detection, multiple instance segmentation	Laparoscopy
Kvasir-Instrument (Ours)	Diagnostic and therapeutic tools in endoscopic images	Binary segmentation, detection and localization	Gastroscopy & colonoscopy

with an exception in 2016. The EndoVis challenge hosts different sub-challenges. The year-wise information about the hosted sub-challenge can be found on the challenge website[1].

Bodenstedt et al. [6] organized "EndoVis 2015 Instrument sub-challenge" for developing new techniques and benchmarking ML algorithms for segmentation and tracking of the instruments on a common dataset. The organizers challenged on two different tasks, i.e., (1) Segmentation and (2) Tracking. The goal of the challenge was to address the problem related to segmentation and tracking of articulated instruments in both laparoscopic and robotic surgery[2]. A comprehensive evaluation of the methods used in instrument segmentation and tracking task for minimally invasive surgery is summarized in this work [6]. The extensive evaluation showed that deep learning works well for instrument segmentation and tracking tasks.

In 2017, a follow up to the previous 2015 challenge was organized called "Robotic Instrument Segmentation Sub-Challenge"[3]. The challenge was part of the Endoscopic vision challenge that was organized at MICCAI 2017. This challenge offered three tasks: (1) Binary segmentation, (2) Parts based segmentation, and (3) Instrument type segmentation. The goal of the binary segmentation

[1] https://endovis.grand-challenge.org/.
[2] https://endovissub-instrument.grand-challenge.org/EndoVisSub-Instrument/.
[3] https://endovissub2017-roboticinstrumentsegmentation.grand-challenge.org/.

task was to separate the image into an instrument and background. Parts segmentation challenged the participants to divide the binary instrument into a shaft, wrist, and jaws. Type segmentation challenged the participants to identify different instrument types. A detailed description of the challenge tasks, dataset, methodologies used by ten participating teams in different tasks, challenge design, and limitation of the challenge can be found in the challenge summary paper [4].

In 2019, a similar challenge called "Robust Medical Instrument Segmentation Challenge 2019"[4] was organized by Roß et al. [13]. This challenge offered three tasks (1) Binary segmentation, (2) Multiple instance detection, and (3) Multiple instance segmentation. The challenge was focused on addressing two key issues in surgical instruments, *Robustness* and *Generalization*, and benchmark medical instrument segmentation and detection on the provided surgical instrument dataset. Endoscopic artefact detection challenge (EAD2019) challenge focused on endoscopic artifact detection primarily but also included instrument class in their detection, segmentation, and "out-of-sample" generalisation tasks. The challenge outcome revealed that most methods performed well for instrument detection and segmentation class [2]. However, this dataset mostly consisted of large biopsy forceps.

In Table 1, we present available instrument datasets in the field of tool segmentation. All of the datasets were designed for hosting challenges. The training dataset is released for all the datasets (except ROBUST-MIS); however, the test dataset is not provided by the challenge organizers. Thus, it makes it difficult to calculate and compare the results on the test dataset. However, experiments are still possible by splitting the training dataset into train, validation, and testing sets. The Robust Medical instrument segmentation dataset is yet not public. However, the participants who have participated in the challenge have the opportunity to download the training dataset. Usually, there are certain practicalities to download the dataset, such as signing the agreement and getting permission from the owner, which takes time, and it is inconvenient. Moreover, to participate in the challenge, the participants have to signup in a particular year, and usually, it often takes a very longtime before they publish the dataset. Thus, the significance of the datasets becomes less as the technology is changing rapidly. More information on available instrument datasets, contents, and offered tasks by the organizers and about the availability can be found from Table 1.

The literature review shows that there are only a few open-access datasets for MIS instrument segmentation. Moreover, to the best of our knowledge, GI tract tools have never been explored. This is the first attempt to provide the community with a curated and annotated public dataset that comprises diagnostic and therapeutic tools in the GI tract. We believe that the presented dataset and the widely used U-Net based algorithm benchmark will encourage the researchers to develop robust and efficient algorithms using the provided dataset that can help clinical procedures in endoscopy.

[4] https://robustmis2019.grand-challenge.org/.

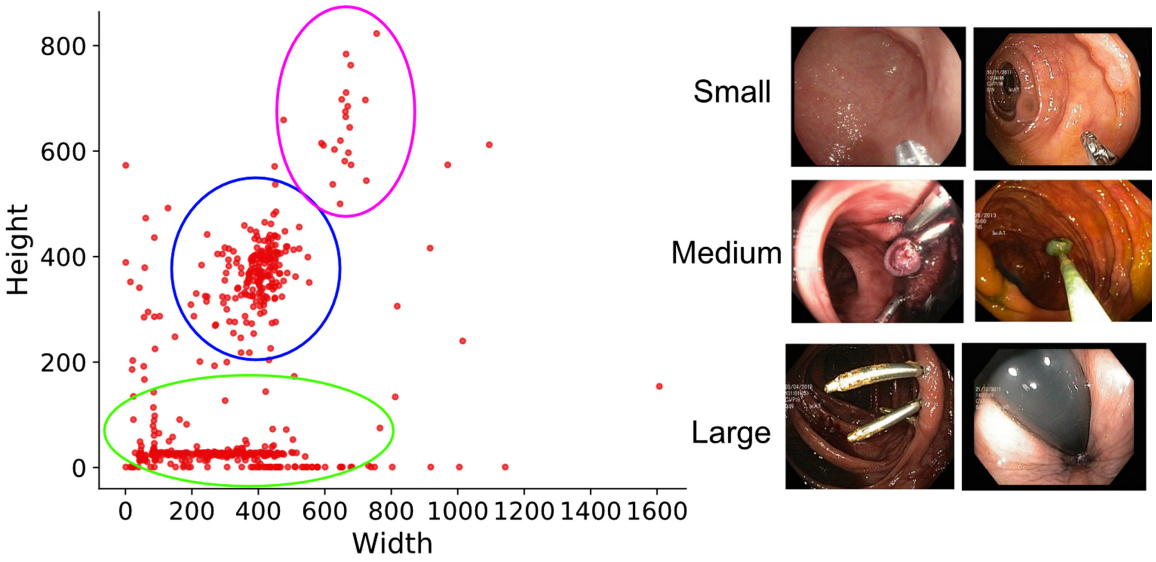


Fig. 1. Distribution of Kvasir-Instrument dataset. On left: Small (green), medium (blue) and large (pink) sized tool clusters. On right: sample images with variable tool size in images. (Color figure online)

3 Kvasir-Instrument Dataset

In this section, we introduce the Kvasir-Instrument dataset with details on how the data was collected, the annotation protocol, and the dataset's structure. The dataset was collected from endoscopic examinations performed at Bærum Hospital in Norway. The unlabelled images' frames are selected from the HyperKvasir dataset [7].

HyperKvasir provides frame-level annotations for 10,662 frames for 23 different classes. However, the majority of the images (99,417 frames) are not labeled. We trained a model using the labeled samples of this dataset and tried to predict the classes of the unlabeled samples. Although our algorithm [15,16] could not classify all the images correctly; however, we were able to classify the presence of instrument or tool out of thousands of provided image frames. However, in order to perform segmentation, pixel-wise masks and bounding boxes were missing. This is what is provided in the proposed dataset, and below, we present the acquisition and annotation protocols used in the data preparation:

3.1 Data Acquisition

The images and videos were collected using standard endoscopy equipment from Olympus (Olympus Europe, Germany) and Pentax (Pentax Medical Europe, Germany) at Bærum Hospital, Vestre Viken Hospital Trust, Norway. All the data used in this study were obtained from videos for procedures that had followed the patient consenting protocol of Bærum Hospital. Additionally, no patient information was available. We have performed a random naming for each publicly released image for further effective annonymisation.

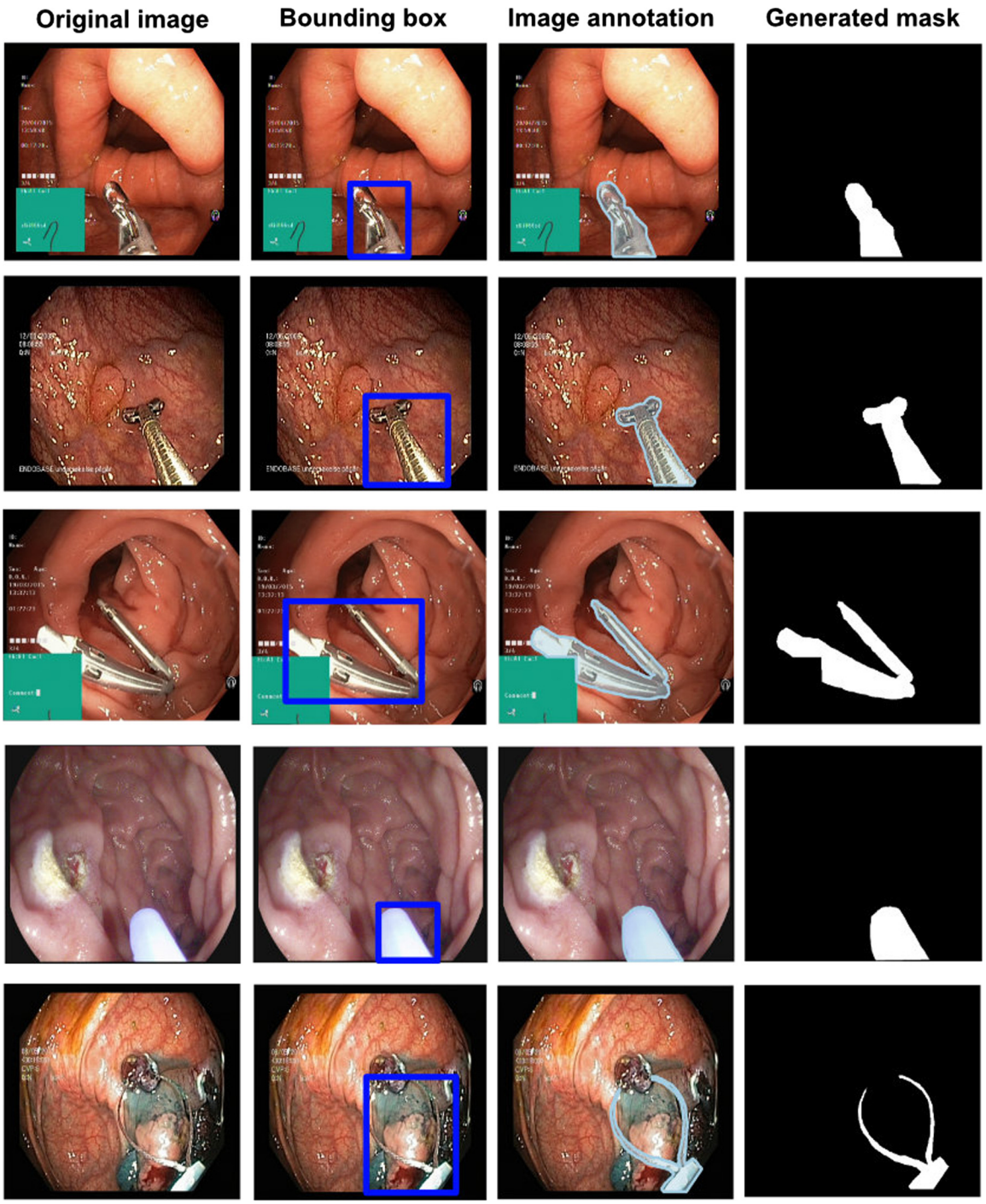


Fig. 2. Kvasir-Instrument dataset: first two rows represent frames with biopsy forceps, the middle row consist of metallic clip, the fourth row is a radio-frequency ablation probe and the last row depicts the crescent and hexagonal shaped snares for polyp removal.

3.2 Annotation Strategy

We have uploaded the Kvasir-Instrument dataset to labelbox[5] and labeled the Region of Interest (ROI) in the image frames, i.e., the ROI of diagnostic and

[5] https://www.labelbox.com/.

therapeutic tools in our cases, and generated all the ground truth masks. Figure 2 shows the example images, bounding box, image annotation, and generated masks for the Kvasir-Instrument dataset. All annotations were then exported in a JSON format, which was used to generate masks for each of the annotations. Related source codes and more information about the dataset can be found at https://github.com/DebeshJha/Kvasir-Instrument.

The exported file contained the information of the images along with the coordinate points that were used for mask and bounding box generation. All annotations were performed using a three-step strategy:

1. The selected samples were labeled by two experienced research assistants.
2. The annotated samples were cross-validated for their delineation quality by two experienced GI experts (more than 10 years of work experience in colonoscopy).
3. The suggested changes were incorporated using the comments from the experts.

The Kvasir-Instrument dataset includes 590 frames consisting of various GI endoscopy tools used during both endoscopic surveillance and therapeutic or surgical procedures. A thorough annotation strategy (detailed above) was used to create bounding boxes and segmentation masks. The dataset consists of variable tool size with respect to image height and width, as presented in Fig. 1. The majority of the tools are small and medium-sized. The sample bounding box annotation, precise area delineation, and extracted masks are shown in Fig. 2.

Our dataset is publicly available and can be accessed at https://datasets.simula.no/kvasir-instrument/. It consists of original image samples (in JPEG format), their corresponding masks (in PNG format), and bounding box information (in JSON format).

4 Benchmarking, Results and Discussion

In this section, we explore encoder-decoder based classical models for baseline algorithm benchmarking, their implementation details for reproducibility, details on evaluation metric used for quantitative analysis, and results and discussion.

4.1 Baseline Methods

U-Net [12] has been explored in the past through many biomedical segmentation challenges and has shown strength towards an effective supervised segmentation model. In this paper, we, therefore, use U-Net based architectures on our Kvasir-Instrument dataset to provide a baseline result for future comparisons. U-Net uses an encoder-decoder architecture, that is, a contractive feature extraction path and expansive path with a classifier to perform binary classification of each image pixel in an upsampled feature map. In our previous work, we have shown that the strength of supervised classification can be amplified by using the output mask from one U-Net [12] architecture to the other by proposing DoubleUNet [10]. In addition, the DoubleUNet architecture uses VGG-19 pretrained

on ImageNet as one of the encoder blocks, squeeze and excite block, and Atrous spatial pyramid pooling (ASPP) block. All other components in the network remain the same as the U-Net. For both networks, dice loss gives a $1 - DSC$, where DSC is the dice similarity coefficient (see Eq. 1 below).

4.2 Implementation Details

We have implemented the U-Net-based and DoubleUNet based architectures using the Keras framework [8] with TensorFlow [1] as backend running on the Experimental Infrastructure for Exploration of Exascale Computing (eX3), NVIDIA DGX-2 machine. We have resized the training dataset into 512×512. We set the batch size of 8 for training. Both architectures are optimized by using Adam optimizer. We have made use of dice loss as the loss function. We split the dataset using 80% of the dataset for training and the remaining 20% for the testing (evaluation). The same split is also provided in the dataset for the further research. We performed basic augmentation, such as horizontal flip, vertical flip, and random rotation. Moreover, we have also provided the train-test split so that others can improve the methods on the same dataset.

4.3 Evaluation Metrics

In this medical image segmentation approach, each pixel of the diagnostic and therapeutic tool either belongs to a tool or non-tool region. The Dice similarity coefficient (DSC) is the mainly used for result evaluation in medical image segmentation. Additionally, we calculate other standard metrics such as Jaccard similarity coefficient (JC) (also known as the intersection over union (IoU)), precision, recall, overall accuracy, F2, and frames per second (FPS). Using tp, fp, tn, and fn to represent the true positives, false positives, true negatives, and false negatives, respectively, the mathematical formulas for them are as follows:

$$\text{DSC} = \frac{2 \cdot tp}{2 \cdot tp + fp + fn} \tag{1}$$

$$\text{JC or IoU} = \frac{tp}{tp + fp + fn} \tag{2}$$

$$\text{Recall } (r) = \frac{tp}{tp + fn} \tag{3}$$

$$\text{Precision } (p) = \frac{tp}{tp + fp} \tag{4}$$

$$\text{F2} = \frac{5p \times r}{4p + r} \tag{5}$$

$$\text{Overall accuracy } (Acc.) = \frac{tp + tn}{tp + tn + fp + fn} \tag{6}$$

$$\text{Frame Per Second } (FPS) = \frac{\#frames}{sec} \tag{7}$$

Table 2. Baseline results for tool segmentation

Method	JC	DSC	F2-score	Precision	Recall	Acc.	FPS
U-Net [12]	**0.8578**	**0.9158**	**0.9320**	**0.8998**	**0.9487**	**0.9864**	**20.4636**
DoubleUNet [10]	0.8430	0.9038	0.9147	0.8966	0.9275	0.9838	10.0000

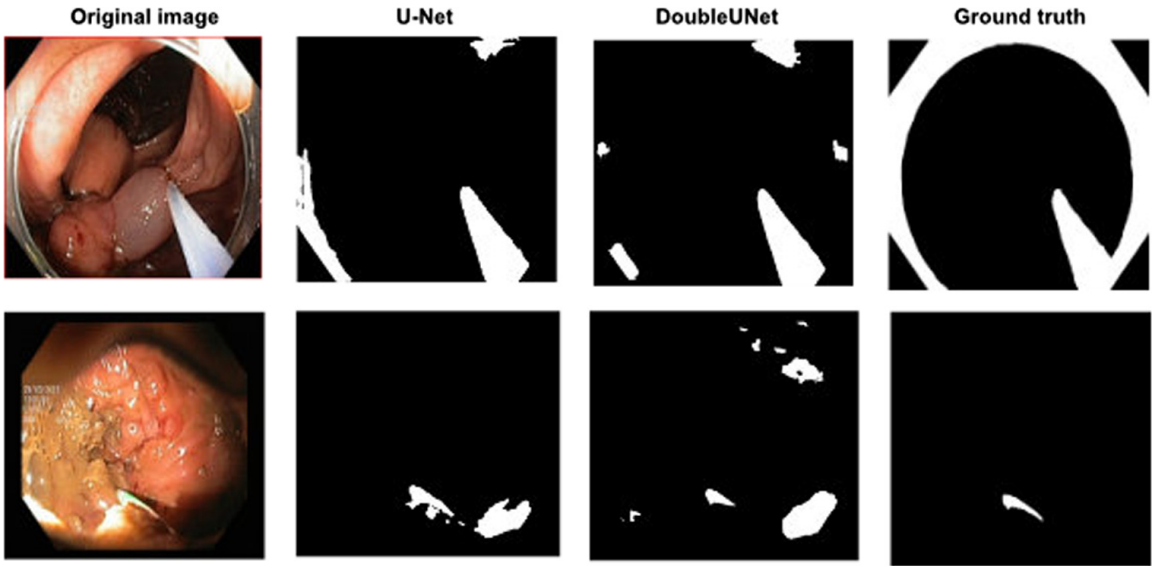


Fig. 3. Failed cases: cap region (top) is under-segmented and small clip area is over-segmented and consist of large number of false positives (bottom).

4.4 Quantitative and Qualitative Results

Table 2 shows the results of the baseline methods for the tool segmentation on the proposed Kvasir-Instrument dataset. From the table, we can observe that the UNet achieved a high JC of 0.8578 and DSC of 0.9158, which is slightly above than the DoubleUNet that yielded JC of 0.8430 and DSC of 0.9038. Also, UNet achieved a speed of 20.4636 FPS, whereas computational time is double for DoubleUNet with only 10 FPS. Similarly, both the recall and precision scores are very comparable for both U-Net ($p = 0.8998, r = 0.9487$) and DoubleUNet ($p = 0.8966, r = 0.9275$).

Figure 3 shows the qualitative result on two challenging sample images. It can be observed that both UNet and DoubleUNet are under-segmenting the cap region (top) and over-segmenting the small clip area (bottom). Some parts of these images are confused because of the presence of saturation areas. However, both models were able to segment well with most endoscopic tool samples in the dataset. This is also evident from the quantitative results. However, even better models are still needed to motivate further research.

4.5 Discussion

From the experimental results in Table 2, we can validate that the classical U-Net architecture outperforms DoubleUNet model. Additionally, U-Net is 2×

faster than the DoubleUNet. This is because U-Net uses basic convolution blocks, whereas DoubleUNet uses pre-trained encoders, ASPP, squeeze, and excite blocks, all of which increase the inference latency. Here, the UNet is optimized by dice loss instead of binary cross-entropy loss, which showed improved performance during our experiments.

Further, fine-tuning on other similar datasets, rigorous data augmentation, and applying more advanced Deep learning (DL) techniques can improve the baseline results - eventually achieving the detection, localisation, and segmentation performance needed to make the technology useful in a clinical environment. Additionally, the use of DL networks with fewer parameters could increase computational efficiency, thereby enabling real-time systems that can be used in clinical settings effectively.

5 Conclusion

We have curated, annotated, and publicly released a dataset that contains *endoscopic tools* used in GI examinations and surgical procedures. The dataset consists of images, bounding boxes, and segmentation masks of endoscopy tools used during different procedures in the GI tract. Additionally, we provided baseline segmentation methods for the automatic delineation of these tools and have compared them using standard computer vision metrics. In the future, we plan to continuously increase the amount of data and also call for multimedia challenges using the presented dataset.

Acknowledgements. This work is funded in part by the Research Council of Norway, project number 263248 (Privaton) and project number 282315 (AutoCap). We performed all computations in this paper on equipment provided by the Experimental Infrastructure for Exploration of Exascale Computing (eX^3), which is financially supported by the Research Council of Norway under contract 270053.

References

1. Abadi, M., et al.: TensorFlow: a system for large-scale machine learning. In: Proceedings of USENIX Symposium on Operating Systems Design and Implementation, pp. 265–283 (2016)
2. Ali, S., et al.: An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. Sci. Rep. **10**(1), 1–15 (2020)
3. Allan, M., Azizian, M.: Robotic scene segmentation sub-challenge. arXiv preprint arXiv:1902.06426 (2019)
4. Allan, M., et al.: 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)
5. Bernhardt, S., Nicolau, S.A., Soler, L., Doignon, C.: The status of augmented reality in laparoscopic surgery as of 2016. Med. Image Anal. **37**, 66–90 (2017)
6. Bodenstedt, S., et al.: Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. arXiv preprint arXiv:1805.02475 (2018)

7. Borgli, H., et al.: Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Sci. Data **7**(1), 1–14 (2020)
8. Chollet, F., et al.: Keras (2015)
9. Cleary, K., Peters, T.M.: Image-guided interventions: technology review and clinical applications. Annu. Rev. Biomed. Eng. **12**, 119–142 (2010)
10. Jha, D., Riegler, M., Johansen, D., Halvorsen, P., Håvard, J.: DoubleU-net: a deep convolutional neural network for medical image segmentation. In: Proceedings of 33rd International Symposium on Computer-Based Medical Systems, pp. 558–564 (2020)
11. Pakhomov, D., Premachandran, V., Allan, M., Azizian, M., Navab, N.: Deep residual learning for instrument segmentation in robotic surgery. In: Suk, H.-I., Liu, M., Yan, P., Lian, C. (eds.) MLMI 2019. LNCS, vol. 11861, pp. 566–573. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32692-0_65
12. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
13. Ross, T., et al.: Robust medical instrument segmentation challenge 2019. arXiv preprint arXiv:2003.10299 (2020)
14. Shvets, A.A., Rakhlin, A., Kalinin, A.A., Iglovikov, V.I.: Automatic instrument segmentation in robot-assisted surgery using deep learning. In: Proceedings of International Conference on Machine Learning and Applications, pp. 624–628 (2018)
15. Thambawita, V., et al.: The medico-task 2018: disease detection in the gastrointestinal tract using global features and deep learning. arXiv preprint arXiv:1810.13278 (2018)
16. Thambawita, V., et al.: An extensive study on cross-dataset bias and evaluation metrics interpretation for machine learning applied to gastrointestinal tract abnormality classification. arXiv preprint arXiv:2005.03912 (2020)

A.8 Paper VIII : Medico Multimedia Task at MediaEval 2020: Automatic Polyp Segmentation

Authors: D. Jha, S. A. Hicks, K. Emanuelsen, H. Johansen, D. Johansen, T. de. Lange, M. A Riegler, and P. Halvorsen

Abstract: Colorectal cancer is the third most common cause of cancer worldwide. According to Global cancer statistics 2018, the incidence of colorectal cancer is increasing in both developing and developed countries. Early detection of colon anomalies such as polyps is important for cancer prevention, and automatic polyp segmentation can play a crucial role for this. Regardless of the recent advancement in early detection and treatment options, the estimated polyp miss rate is still around 20%. Support via an automated computer-aided diagnosis system could be one of the potential solutions for the overlooked polyps. Such detection systems can help low-cost design solutions and save doctors time, which they could for example use to perform more patient examinations. In this paper, we introduce the 2020 Medico challenge, provide some information on related work and the dataset, describe the task and evaluation metrics, and discuss the necessity of organizing the Medico challenge.

Published: Proceedings of MediaEval2020

Candidate contributions: D. Jha proposed and led the "Medico automatic polyp segmentation challenge" held at MediaEval 2020. He collected, annotated, and prepared the dataset with the help of an expert gastroenterologist. He proposed the tasks and evaluation metrics in the challenge. At the end of the challenge, he evaluated all the teams' solutions, provided the scores to the teams, and led the workshop. Additionally, he wrote the manuscript with input from all of the co-authors.

Thesis objectives: Objective I, Objective III

Medico Multimedia Task at MediaEval 2020: Automatic Polyp Segmentation

Debesh Jha[1,2], Steven A. Hicks[1,3], Krister Emanuelsen[1], Håvard Johansen[2]
Dag Johansen[2], Thomas de Lange[4,5,6], Michael A. Riegler[1], Pål Halvorsen[1,3]
[1] SimulaMet, Norway [2] UiT The Arctic University of Norway [3] Oslo Metropolitan University, Norway
[4] Augere Medical AS, Norway [5] Sahlgrenska University Hospital, Sweden [6] Bærum Hospital, Norway

ABSTRACT

Colorectal cancer is the third most common cause of cancer worldwide. According to Global cancer statistics 2018, the incidence of colorectal cancer is increasing in both developing and developed countries. Early detection of colon anomalies such as polyps is important for cancer prevention, and automatic polyp segmentation can play a crucial role for this. Regardless of the recent advancement in early detection and treatment options, the estimated polyp miss rate is still around 20%. Support via an automated computer-aided diagnosis system could be one of the potential solutions for the overlooked polyps. Such detection systems can help low-cost design solutions and save doctors time, which they could for example use to perform more patient examinations. In this paper, we introduce the 2020 Medico challenge, provide some information on related work and the dataset, describe the task and evaluation metrics, and discuss the necessity of organizing the Medico challenge.

1 INTRODUCTION

The goal of *Medico automatic polyp segmentation challenge* the benchmarking of polyp segmentation algorithms on new test images for automatic polyp segmentation that can detect and mask out polyps (including irregular, small or flat polyps) with high accuracy. The main goal of the challenge is to benchmark different computer vision and machine learning algorithms on the same dataset that could promote to build novel methods which could be potentially useful in clinical settings. Moreover, we emphasize on robustness and generalization of the methods to solve the limitations related to data availability and method comparison. The detailed challenge description can be found here https://multimediaeval.github.io/editions/2020/tasks/medico/.

After three years of organizing the Medico Multimedia Task [6, 17, 18], we present the fourth iteration in the series. With a focus on assessing human semen quality last year [6], this year we build on the 2017 [18] and 2018 [17] challenges of automatically detecting anomalies in video and image data from the GI tract. We introduce a new task for automatic polyp *segmentation*. In the prior gastrointestinal (GI) challenges, we classified the images into various classes. We are now interested in identifying each pixel of the lesions from the provided polyp images in this challenge.

The task is important because colorectal cancer (CRC) is the third most leading cause of cancer and fourth most prevailing strain in terms of cancer incidence globally [2]. Regular screening through colonoscopy is a prerequisite for early cancer detection and prevention of CRC. Regardless of the achievement of colonoscopy examinations, the estimated polyp miss rate is still around 20% [12], and there are large inter-observer variabilities [13]. An automated computer-aided diagnosis (CADx) system detecting and highlighting polyps could be of great help to improve the average endoscopist performance.

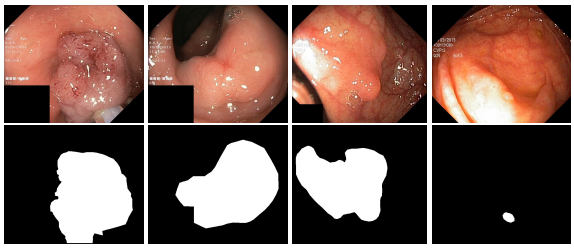


Figure 1: Polyps and corresponding masks from Kvasir-SEG

In recent years, convolutional neural networks (CNNs) have advanced medical image segmentation algorithms. However, it is essential to understand the strengths and weaknesses of the different approaches via performance comparison on a common dataset. There are a large number of available studies on automatic polyp segmentation [3–5, 8, 9, 11, 14, 20]. However, most of the conducted studies are performed on a restricted dataset which makes it difficult for benchmarking, algorithm development and reproducible results. Our challenge is utilizing the publicly available Kvasir-SEG dataset [10]. The entire Kvasir-SEG dataset is used for training and an additional and unseen test dataset for benchmarking the algorithms.

In summary, the Medico 2020 challenge can support building future systems and foster open, comparable and reproducible results where *the objective of the task is to find efficient solutions automatic polyp segmentation*, both in terms of pixel-wise accuracy and processing speed.

For the clinical translation of technologies, it is essential to design methods on multi-centered and multi-modal datasets. We have recently released several gastrointestinal endoscopy [1, 15, 16], wireless capsule endoscopy [19], endoscopic instrument [7], and polyp datasets [10]. Thus, we have put in significant effort to address the challenges related to lack of public available datasets in the field of GI endoscopy.

2 DATASET

The Kvaris-SEG [10] training dataset can be downloaded from https://datasets.simula.no/kvasir-seg/. It contains 1,000 polyp images and

MediaEval'20, December 14-15 2020, Online

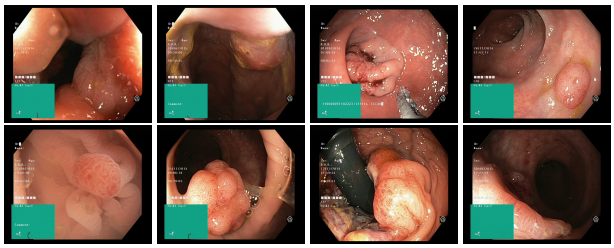


Figure 2: Examples polyps from the test images

their corresponding ground truth mask as shown in Figure 1. The dataset was collected from real routine clinical examinations at Bærum Hospital in Norway by expert gastroenterologists. The resolution of images varies from 332×487 to 1920×1072 pixels. Some of the images contain a green thumbnail in the lower-left corner of the images showing the scope position marking from the ScopeGuide (Olympus) (see Figure 2). We annotate another separate dataset consisting of 160 new polyp images and use the resulting dataset as the test set to benchmark the participants' approaches. Figure 2 shows some examples of test images used in the challenge.

3 TASK DESCRIPTION

The participants are invited to submit their solutions for the two following tasks: segmentation and efficiency (speed).

3.1 The automatic polyp segmentation task

This task invites participants to develop new algorithms for segmentation of polyps. The main focus is to develop an efficient system in terms of diagnostic ability and processing speed and accurately segment the maximum polyp area in a frame from the provided colonoscopic images.

There are several ways to evaluate the segmentation accuracy. The most commonly used metrics by the wider medical imaging community are the correct **Dice similarity coefficient (DSC)** or overlap index, and the **mean Intersection over Union (mIoU)**, also known as the Jaccard index. In clinical applications, the gastroenterologists are interested in pixel-wise detail information extraction from the potential lesions. The metrics such as DSC and mIoU are used to compare the pixel-wise similarity between the predicted segmentation maps and the original ground truth of the lesions.

The DSC is a metric for comparison of the similarities between two given samples. If *tp*, *tn*, *fp*, and *fn* represent the number of true positive, true negative, false positive and false negative per-pixel predictions for an image, respectively, then the DSC is given as

$$\text{DSC} = \frac{2 \cdot tp}{2 \cdot tp + fp + fn}$$

Furthermore, the IoU is then defined as the ratio of intersection of two metrics over a union of two corresponding metrics. The mean IoU computes IoU of each semantic class of an image and calculate the mean over each classes. The IoU is defined as:

$$\text{IoU} = \frac{tp}{tp + fp + fn}$$

Moreover, in the polyp image segmentation task (i.e., a binary segmentation task), **precision** (positive predictive value) shows over-segmentation, and **recall** (true positive rate) shows under-segmentation. Over-segmentation means that the predicted image covers more area than the ground truth in some part of the frame. The under-segmentation implies that the algorithm has predicted less polyp content in some portion of the image compared to its corresponding ground truth. We also encourage participants to calculate precision and recall, and these are given by:

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}.$$

The main metric for evaluation and ranking of the teams is **mIoU**. There is a direct correlation between mIoU and DSC. Therefore, we have only used one metric. If the teams have the same mIoU values, then the teams will be further evaluated on the basis of the higher value of the DSC. For the evaluation, we ask the participants to submit the predicted masks in a zip file. The resolution of the predicted masks must be equal to the test images.

3.2 The algorithm speed efficiency task

Real-time polyp detection is required for live patient examinations in the clinic. It can gain gastroenterologist attention to the region of interest. Thus, we also ask participants to participate in the efficiency task. The algorithm efficiency task is similar to the previous task, but it puts a stronger emphasis on the algorithm's speed in terms of frames-per-second.

Submissions for this task will be evaluated based on both the algorithm's speed and segmentation performance. The segmentation performance (the segmentation accuracy) will be measured using the same **mIoU** metric as described above for the first task, whereas speed will be measured by **frames-per-second (FPS)** according to the following formula:

$$FPS = \frac{\#frames}{sec}$$

For this task, we require participants to submit their proposed algorithm as part of a Docker image so that we can evaluate it on our hardware. We evaluate the performance of the algorithm on the Nvidia GeForce GTX 1080 system. For the team ranking, we set a certain mIoU as threshold for considering it as a valid efficient segmentation solution and rank according to the FPS.

4 DISCUSSION AND OUTLOOK

Currently, there is a growing interest in the development of CADx systems that could act as a second observer and digital assistant for the endoscopists. Algorithmic benchmarking is an efficient approach to analyze the results of different methods. A comparison of different approaches can help us to identify challenging cases in the data. We then can discriminate the image frames into simple, moderate, and challenging images. Later on, we can target to develop models on the challenging images that are usually missed out during a routine examination to design better CADx systems. We hope that this approach would help us to design better performing algorithms/models that may increase the efficiency of the health system.

REFERENCES

[1] Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, et al. 2020. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data* 7, 1 (2020), 1–14.

[2] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* 68, 6 (2018), 394–424.

[3] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. 2020. Pranet: Parallel reverse attention network for polyp segmentation. *arXiv preprint arXiv:2006.11392* (2020).

[4] Yunbo Guo, Jorge Bernal, and Bogdan J Matuszewski. 2020. Polyp Segmentation with Fully Convolutional Deep Neural Networks—Extended Evaluation Study. *Journal of Imaging* 6, 7 (2020), 69.

[5] Yun Bo Guo and Bogdan Matuszewski. 2019. GIANA Polyp Segmentation with Fully Convolutional Dilation Neural Networks. In *Proc. of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. 632–641.

[6] Steven Hicks, Michael Riegler, Pia Smedsrud, Trine B Haugen, Kristin Ranheim Randel, Konstantin Pogorelov, Håkon Kvale Stensland, Duc-Tien Dang-Nguyen, Mathias Lux, Andreas Petlund, et al. 2019. Acm multimedia biomedia 2019 grand challenge overview. In *Proc. of the ACM International Conference on Multimedia*. 2563–2567.

[7] Debesh Jha, Sharib Ali, Krister Emanuelsen, Steven Hicks, Vajira Thambawita, Riegler Michael A Garcia-Ceja, Enrique, Lange Thomas de, Peter T. Schmidt, Johansen Håvard, Dag Johansen, and Halvorsen Pål. 2021. Kvasir-Instrument: Diagnostic and Therapeutictool Segmentation Dataset in Gastrointestinal Endoscopy. In *Proc. of International Conference on Multimedia Modeling*.

[8] Debesh Jha, Sharib Ali, Håvard D. Johansen, Dag Johansen, Jens Rittscher, Michael A. Riegler, and Pål Halvorsen. 2020. Real-Time Polyp Detection, Localisation and Segmentation in Colonoscopy Using Deep Learning. *arXiv preprint arXiv:2006.11392* (2020).

[9] Debesh Jha, Michael A Riegler, Dag Johansen, Pål Halvorsen, and Håvard D Johansen. 2020. DoubleU-Net: A Deep Convolutional Neural Network for Medical Image Segmentation. In *Proc. of International Symposium on Computer-Based Medical Systems*. 558–564.

[10] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. 2020. Kvasir-SEG: A segmented polyp dataset. In *Proc. of International Conference on Multimedia Modeling*. 451–462.

[11] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D Johansen. 2019. ResUNet++: An Advanced Architecture for Medical Image Segmentation. In *Proc. of International Symposium on Multimedia*. 225–230.

[12] Michal F Kaminski, Jaroslaw Regula, Ewa Kraszewska, Marcin Polkowski, Urszula Wojciechowska, Joanna Didkowska, Maria Zwierko, Maciej Rupinski, Marek P Nowacki, and Eugeniusz Butruk. 2010. Quality indicators for colonoscopy and the risk of interval cancer. *New England Journal of Medicine* 362, 19 (2010), 1795–1803.

[13] Nadim Mahmud, Jonah Cohen, Kleovoulos Tsourides, and Tyler M Berzin. 2015. Computer vision and augmented reality in gastrointestinal endoscopy. *Gastroenterology report* 3, 3 (2015), 179–184.

[14] Tanvir Mahmud, Bishmoy Paul, and Shaikh Anowarul Fattah. 2020. PolypSegNet: A Modified Encoder-Decoder Architecture for Automated Polyp Segmentation from Colonoscopy Images. *Computers in Biology and Medicine* (2020), 104119.

[15] Konstantin Pogorelov, Kristin Ranheim Randel, Thomas de Lange, Sigrun Losada Eskeland, Carsten Griwodz, Dag Johansen, Concetto Spampinato, Mario Taschwer, Mathias Lux, Peter Thelin Schmidt, et al. 2017. Nerthus: A Bowel Preparation Quality Video Dataset. In *Proceedings of the ACM on Multimedia Systems Conference*. 170–174.

[16] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, et al. 2017. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proc. of the ACM on Multimedia Systems Conference*. 164–169.

[17] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Steven Hicks, Kristin Ranheim Randel, Duc Tien Dang Nguyen, Mathias Lux, Olga Ostroukhova, and Thomas de Lange. 2018. Medico multimedia task at MediaEval 2018. In *Proc. of MediaEval 2018 CEUR Workshop*.

[18] Michael Riegler, Konstantin Pogorelov, Pål Halvorsen, Carsten Griwodz, Thomas Lange, Kristin Randel, Sigrun Eskeland, Duc Tien Dang Nguyen, Mathias Lux, and Concetto Spampinato. 2017. Multimedia for medicine: the medico task at Mediaeval 2017. In *Proc. CEUR Worksh. Multim. Bench. Worksh.*

[19] Pia H Smedsrud, Henrik L Gjestang, Oda O Nedrejord, Espen Næss, Vajira Thambawita, Steven Hicks, Hanna Borgli, Debesh Jha, Tor Jan Derek Berstad, Sigrun L Eskeland, et al. 2020. Kvasir-Capsule, a video capsule endoscopy dataset. (2020).

[20] Pu Wang, Xiao Xiao, Jeremy R Glissen Brown, Tyler M Berzin, Mengtian Tu, Fei Xiong, Xiao Hu, Peixi Liu, Yan Song, Di Zhang, et al. 2018. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nature biomedical engineering* 2, 10 (2018), 741–748.

A.9 Paper IX: LightLayers: Parameter Efficient Dense and Convolutional Layers for Image Classification

Authors: D. Jha, A. Yazidi, M. A. Riegler, D. Johansen, H. D. Johansen, and P. Halvorsen

Abstract: Deep Neural Networks (DNNs) have become the de-facto standard in computer vision, as well as in many other pattern recognition tasks. A key drawback of DNNs is that the training phase can be very computationally expensive. Organizations or individuals that cannot afford purchasing state-of-the-art hardware or tapping into cloud-hosted infrastructures may face a long waiting time before the training completes or might not be able to train a model at all. Investigating novel ways to reduce the training time could be a potential solution to alleviate this drawback, and thus enabling more rapid development of new algorithms and models. In this paper, we propose LightLayers, a method for reducing the number of trainable parameters in deep neural networks (DNN). The proposed LightLayers consists of LightDense andLightConv2D layer that are as efficient as regular Conv2D and Dense layers, but uses less parameters. We resort to Matrix Factorization to reduce the complexity of the DNN models resulting into lightweight DNN models that require less computational power, without much loss in the accuracy. We have tested LightLayers on MNIST, Fashion MNIST, CI-FAR 10, and CIFAR 100 datasets. Promising results are obtained for MNIST, Fashion MNIST, CIFAR-10 datasets whereas CIFAR 100 shows acceptable performance by using fewer parameters.

Published: Proceedings of Joint conference of Parallel and Distributed Computing, Applications and Technologies and International Symposium on Parallel Architectures, Algorithms and Programming (PDCAT-PAAP)

Candidate contributions: D. Jha conceptualized and designed the work. He performed all the experiments and analyses in the paper. Additionally, D. Jha drew all the figures and conducted ablation study. Moreover, he wrote and revised the manuscript with input from all the co-authors.

Thesis objectives: Objective II

LightLayers: Parameter Efficient Dense and Convolutional Layers for Image Classification

Debesh Jha[1,2( )], Anis Yazidi[3], Michael A. Riegler[1], Dag Johansen[2], Håvard D. Johansen[2], and Pål Halvorsen[1,3]

[1] SimulaMet, Oslo, Norway
debesh@simula.no
[2] UIT The Arctic University of Norway, Tromsø, Norway
[3] Oslo Metropolitan University, Oslo, Norway

Abstract. Deep Neural Networks (DNNs) have become the de-facto standard in computer vision, as well as in many other pattern recognition tasks. A key drawback of DNNs is that the training phase can be very computationally expensive. Organizations or individuals that cannot afford purchasing state-of-the-art hardware or tapping into cloud hosted infrastructures may face a long waiting time before the training completes or might not be able to train a model at all. Investigating novel ways to reduce the training time could be a potential solution to alleviate this drawback, and thus enabling more rapid development of new algorithms and models. In this paper, we propose LightLayers, a method for reducing the number of trainable parameters in DNNs. The proposed LightLayers consists of LightDense and LightConv2D layers that are as efficient as regular Conv2D and Dense layers but uses less parameters. We resort to Matrix Factorization to reduce the complexity of the DNN models resulting in lightweight DNN models that require less computational power, without much loss in the accuracy. We have tested LightLayers on MNIST, Fashion MNIST, CIFAR 10, and CIFAR 100 datasets. Promising results are obtained for MNIST, Fashion MNIST, and CIFAR-10 datasets whereas CIFAR 100 shows acceptable performance by using fewer parameters.

Keywords: Deep learning · Lightweight model · Convolutional neural network · MNIST · Fashion MNIST · CIFAR-10 · CIFAR 100 · Weight decomposition

1 Introduction

Deep learning (DL) techniques have revolutionized the field of Machine Learning (ML) and gained immense research attention during the last decade. Deep neural networks provide state-of-the-art solution in several domains such as image recognition, speech recognition, and text processing [20]. One of the most popular techniques within deep learning is Convolutional Neural Network (CNN), which possesses a structure that is well-suitable specially for image and video processing. A CNN [16] comprises a convolution layer and dense layer. CNN has

Y. Zhang et al. (Eds.): PDCAT 2020, LNCS 12606, pp. 285–296, 2021.
https://doi.org/10.1007/978-3-030-69244-5_25

emerged as powerful techniques for solving many classification [14] and regression [12] tasks. Additionally, CNN has produced promising results in various applications areas, including in the medical domain, with applicability in diabetic retinopathy prediction [3], endoscopic disease detection [23], and breast cancer detection [19].

Recently, developing deeper and larger architectures has been a common trend in the development of state-of-the-art methods [4]. Most of the time, we can observe that deeper networks especially with large and complex datasets lead to better performance. One of the major drawbacks of CNNs are that they often require an immense amount of training time compared to other classical ML algorithms. Hyperparameter optimization for fine-tuning the model is another challenging task that increases dramatically the overall training time to achieve optimum results from any model. CNN models often require powerful Graphical Processing Units (GPUs) for training, which can span over days, weeks, and even months, with no guarantee that the model will produce satisfactory results. A long training process also consumes a lot of energy and is not considered environmentally friendly. Furthermore, long training is demanding in terms of resources as a large amount of memory is required which renders it difficult to deploy on low-power devices [11]. The requirements for the expensive hardware and high training time complicate the use of models with large number of trainable parameters to be deployed on portable devices or conventional desktops [20].

A potential way to address these issues is the introduction of lightweight models. A lightweight model can potentially be built by reducing the number of trainable parameters within the layers. In an effort towards reducing the training time and complexity of CNN models, we propose LightLayers, which is a combination of LightDense and LightConv2D layers, that focuses on CNNs and more particularly on creating both a lightweight convolutional layer and a lightweight dense layer that are both easy to train. Lightweight CNN models are computationally cheap and can be used in various applications for carrying out online estimation. Therefore, the main goal of this paper is to present a general model to reduce the number of parameters in a CNN model so that it can be used in various image processing or other applicable tasks in the future.

The main contributions of the paper are:

- LightLayers, a combination of LightConv2D and LightDense layers, is proposed. Both layers are based on matrix decomposition for reducing the number of trainable parameters of the layers.
- We have investigated and tested the proposed model with four different publicly available datasets: MNIST [16], Fashion MNIST [26], and CIFAR10 [13], CIFAR100 [13], and we have showed that the proposed method is competitive in terms of both accuracy and efficiency when the number of training parameters used are taken into consideration.
- We experimentally show that good accuracy can be achieved by using a relatively small number of trainable parameters with MNIST, Fashion MNIST,

and CIFAR 10 datasets. Moreover, we found there was a significant reduction in the number of trainable parameters as compared to Conv2D.

2 Related Work

In the context of reducing the cost of network model training, several approaches have been presented. For example, Xue et al. [27] presented a Deep Neural Network (DNN) technique for reducing the model size while maintaining the accuracy. For achieving this goal, they used singular value decomposition (SVD) on the weight matrix in DNN, and reconstructed the model based on inherent sparseness of the original matrices. The application of DNNs for mobile applications has become increasingly popular. The computational and storage limitation should be taken into account while deploying DNN on such devices.

To address this need, Li et al. [17] proposed two techniques for effectively learning from DNNs with a smaller number of hidden nodes and smaller number of senones set. The details about these techniques can be found in the literature [17]. Similarly, Xue et al. [28] introduced two SVD based techniques to solve the issue related to DNN personalization and adaptation. Garipov et al. [8] developed a tensor factorization framework for compressing fully connected layers. The focus of their work was to compress convolutional layers which would potentially excel in image recognition tasks by reducing the memory complexity and high computational cost. Later, Kim et al. [10] proposed an energy-efficient kernel decomposition architecture for binary-weight CNNs.

Ding et al. [7] proposed CIRCNN, an approach for representing the weights and processing neural networks using block-circulant matrices. CIRCNN utilizes Fast Fourier Transform based fast multiplication operation which simultaneously reduces the computational and storage complexity causing negligible loss in accuracy. Chai et al. [25] proposed a model for reducing the parameters in DNNs via product-of-sums matrix decomposition. They obtained good accuracy on the MNIST and Fashion MNIST datasets with a smaller number of trainable parameters. Another similar work is by Agrawal et al. [2], where they designed a lightweight deep learning model for human activity recognition that is sufficiently computationally efficient to be deployed on edge devices. For more recent works on matrix and tensor decomposition, we refer the reader to [6,15].

Kim et al. [11] proposed a method for compressing CNNs to be deployed as a mobile application. Mariet et al. [18] proposed another efficient neural network architecture that reduces the size of neural networks without hurting the overall performance. Novikov et al. [20] converted dense weight matrices of fully connected layers to Tensor Train [21] format such that the number of parameters are reduced by a huge factor by preserving the expressive power of the layer.

Lightweighted networks have gained attention in computer vision (for instance, in the area of real-time image segmentation [9,22,24,29]). Real-time applications are growing because the lightweight models can be an efficient

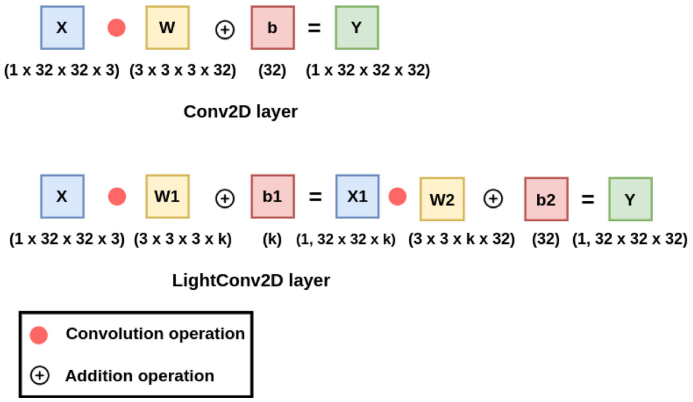


Fig. 1. Comparative diagram of Conv2D layer and LightConv2D layer.

solution for resource constraints and mobile devices. Only a lightweight model demands lower memory that leads to a lower computation and faster speed. Therefore, developing a lightweight model can be a good idea for achieving real-time solutions, and it can be used for other applications too.

The above studies show that there is great potential for lightweight networks for computer-vision tasks. With large amounts of training data, it is likely that a model with huge numbers of trainable parameters will outperform the smaller models—if one can afford the high training costs and resource demands at inference time. However, there is a need for models with low-cost computational power and small memory footprints [11], especially for mobile applications [11] and portable devices. In this respect, we propose LightLayers that is based on the concept of matrix decomposition. LightLayers uses fewer trainable parameters and shows the state-of-the-art tradeoff between parameter size and accuracy.

3 Methodology

In this section, we introduce the proposed layers. Figure 1 shows the comparison of a Conv2D and a LightConv2D layer. In the LightConv2D layer, we decompose the weight matrix W into $W1$ and $W2$ on the basis of hyperparameter k, which leads to a reduction of the total number of trainable parameters in the network. We follow the same strategy for the LightDense layer. The block diagram of the LightDense layer is shown in Fig. 2.

The main objective of building the model is to compare our LightLayers (i.e., the combination of LightConv2D and LightDense layers) with the conventional Conv2D and SeparableConv2D layers. For comparing the performance of the various layers, we have built a simple model from scratch. The block diagram of the proposed model is shown in Fig. 3. We used the same hyperparameters and setting for all the experiments. For the LightLayers experiments, we used LightConv2D and LightDense layers (see Fig. 3). For the other experiments, we replaced LightConv2D with Conv2D or SeparableConv2D and LightDense with a regular Dense layer.

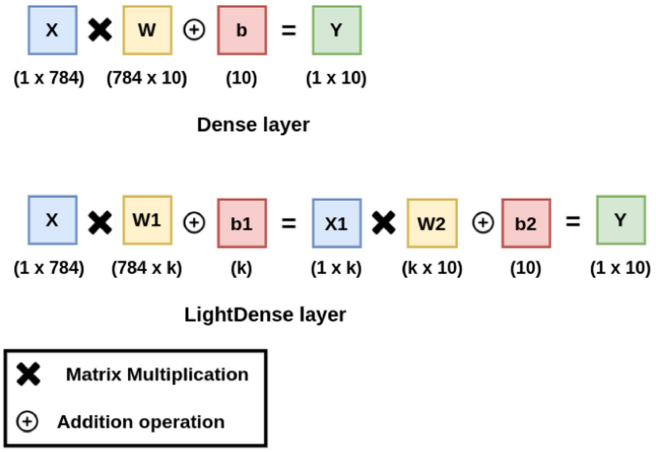


Fig. 2. Comparative diagram of Dense layer and LightDense layer.

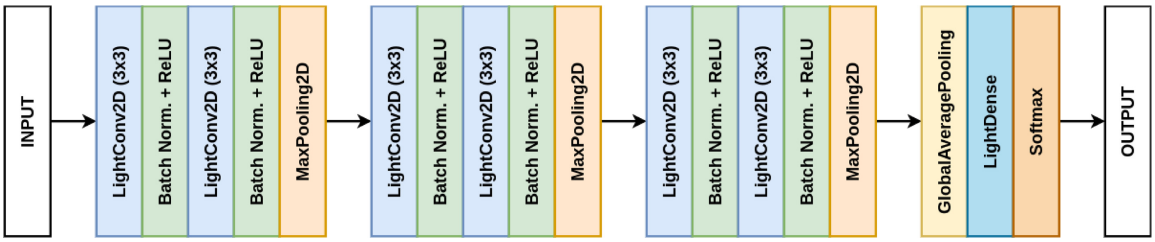


Fig. 3. Block diagram of the architecture used for comparison of the proposed Lightlayers with regular convolution and dense layers. In the case of regular layers, we use regular convolution and dense layers instead of Lightlayers.

The model architecture used for experimentation (see Fig. 3) comprises two 3×3 convolution layers, each followed by a batch-normalization and ReLU non-linearity as the activation function. We have introduced 2×2 max-pooling, which reduces the spatial dimension of the feature map. We have used three similar blocks of layers in the model followed by the GlobalAveragePooling, LightDense layers with $k = 8$, and a softmax activation function for classifying the input image.

3.1 Description of Convolution Layers

Conv2D. A convolution layer is the most common layer used in any computer vision task and is applied extensively. This layer uses a multidimensional kernel as the weight, which is used to perform convolution operation on the input to produce an output. If the bias is used, then a $1D$ vector is added to the output. Finally, the activation is applied to introduce the non-linearity into the neural network. In this paper, we worked on a $2D$ convolution layer, which uses a $4D$ tensor as the weight.

$$Output = Activation((Input \otimes Weight) + Bias) \tag{1}$$

In the above equation, $\otimes$ represents the convolution operation, and weight represents the kernel.

Dense Layer. A dense layer is the regular, deeply connected neural-network layer. It is the most common and frequently used layer. It is also known as a fully-connected layer as each neuron receives input from the previous layer.

$$Output = Activation((Input \oplus Weight) + Bias) \tag{2}$$

In the above equation, $\oplus$ represents the matrix multiplication instead of convolution operation as above.

Separable Conv2D. Separable convolution, also known as depth-wise convolution, is used in our experiment. We use depth-wise separable 2D convolution to compare the performance of our model. It first applies a depth-wise spatial convolution, i.e., performing a convolution operation on each input channel independently. After that, it is followed by a point-wise convolution, i.e., a 1×1 convolution. Pointwise, convolution controls the number of filters in the output feature maps.

4 Experimental Setup

For the experiments, we use the same number of layers, filters, filter sizes, and activation functions in every model for the individual dataset. We have modified the existing Dense and Conv2D layer in such a way that the number of trainable parameters decreases with some decrease in the accuracy of the model. In particular, we use three types of layers for this experiment, i.e., Conv2D, SeparableConv2D, and LightLayers. First, we run the model using Conv2D layers. The Conv2D layer is replaced by SeperableConv2D and run again. Again, we replace SeperableConv2D with the LightLayers and run the model.

In the modified layers, we introduced the hyperparameter k to control the number of trainable parameters in the LightDense and LightConv2D layer. In the LightDense layer, we set k to 8. In the LightConv2D layer, k varies between 1 to 6, and more could be set depending on the requirement. The values of the k are chosen empirically. We only replace the Conv2D layer with the LightConv2D layer and Dense layer with the LightDense layer of the proposed lightweight model. The rest of the network architecture remains the same.

4.1 Implementation Details

We have implemented the proposed layers using the Keras framework [5] and TensorFlow 2.2 [1] as backend. The implementation can be found at GitHub[1]. We performed all the experiments on an NVIDIA GEFORCE GTX 1080 system, which has 2560 NVIDIA CUDA Cores with 8 GB GDDR5X memory. The system was running on Ubuntu 18.04.3 LTS. We used a batch size of 64. All the experiments were run, keeping all the hyperparameters (i.e., learning rate, optimizer, batch size, number of filters, and filter size) the same. We have trained all the models for 20 epochs. After each convolution layer, batch normalization is used, which is activated by the Rectified linear unit (ReLU).

[1] https://github.com/DebeshJha/LightLayers.

4.2 Datasets

To evaluate LightConv2D layer and LightDense layer, we have performed experiments using various datasets.

MNIST Database. Modified National Institute of Standards and Technology (MNIST) [16] is the primary dataset for computer vision tasks introduced by LeCun et al. in 1998. MNIST comprises 10 classes of handwritten digits with 60,000 training and 10,000 testing images. The resolution of the images in the MNIST dataset is 28×28. There is a huge recent advancement in ML and DL algorithms. However, the MNIST remains a common choice for learners and beginners. The reason is that it is easy to deploy, test, and compare an algorithm on a publicly available dataset. The dataset can be downloaded from http://yann.lecun.com/exdb/mnist/.

Fashion MNIST Database. Fashion MNIST [26] is a 10 class of 70,000 grayscale images of size 28×28. Xiao et al. released a novel image dataset that could be used for benchmarking ML algorithms. Their goal was to replace the MNIST database with a new database. The images of the Fashion MNIST database are more challenging as compared to the MNIST database. It contains natural images such as t-shirt/top, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag, and ankle boot. The database can be downloaded from https://github.com/zalandoresearch/fashion-mnist.

CIFAR-10 Database. CIFAR-10 [13] is a commonly established dataset for computer-vision tasks. It is especially used for object recognition tasks. CIFAR-10 contains 60,000 color images of size 32×32. It also has 10 classes of images. Each class contains 6,000 images per class. The classes contain datasets of cars, birds, cats, deer, dogs, horses, and trucks. The dataset can be downloaded from https://www.cs.toronto.edu/~kriz/cifar.html.

CIFAR-100 Database. CIFAR-100 [13] is also collected by the team of Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. This database is similar to the previous CIFAR-10 database. The 100 classes of the database consist of images such as beaver, dolphin, flatfish, roses, clock, computer keyboard, bee, forest, baby, pine, tank, etc. Each class of the database contains 600 images each. This dataset contains 500 training examples and 100 testing examples per class. The dataset can be found on the same webpage as CIFAR-10.

5 Results

In this section, we present and compare the experimental results of the Conv2D, SeperableConv2D, and LightLayers models on the MNIST, Fashion MNIST, CIFAR-10, and CIFAR 100 datasets. Table 1 shows the summary of result

Table 1. Results on **MNIST** test dataset (Number of epochs = 10, Batch size = 64, Learning rate = 1e-3, Number of filters = $[8, 16, 32]$).

Method	Parameters	Test Accuracy	Test Loss
Conv2D	18,818	0.9887	0.018
SeparableConv2D	3,611	0.9338	0.2433
LightLayers ($K = 1$)	2,649	0.9418	0.1327
LightLayers ($K = 2$)	4,392	0.9749	0.0554
LightLayers ($K = 3$)	**6,135**	**0.9775**	**0.0513**
LightLayers ($K = 4$)	7,878	0.9720	0.0704

Table 2. Results on **Fashion MNIST** test dataset (Number of epochs = 10, Batch size = 64, Learning rate = 1e-3, Number of filters = $[8, 16, 32]$).

Method	Parameters	Test accuracy	Test loss
Conv2D	18,818	0.9147	0.1468
SeparableConv2D	3,611	0.8725	0.3175
LightLayers ($K = 1$)	2,649	0.789	0.6752
LightLayers ($K = 2$)	4,392	0.8452	0.4247
LightLayers ($K = 3$)	6,135	0.8695	0.3708
LightLayers ($K = 4$)	7,878	0.8623	0.6184
LightLayers ($K = 5$)	**9,621**	**0.8820**	**0.2810**
LightLayers ($K = 6$)	11,364	0.8733	0.3986

comparison of Conv2D, SeperableConv2D, and LightLayers on MNIST dataset. Based on Conv2D and SeperableConv2D, we propose Layers and show improvement over both layers. The concept of LightLayers is based on weight matrix decomposition. This is the main motivation behind comparison of the proposed layers with Conv2D and SeperableConv2D.

The hyperparameters used are described in the caption of the Table 1. We can see that the result of the proposed LightLayers is comparable to that of Conv2D and SeperableConv2D in terms of test accuracy. When we compare the LightLayers with Conv2D, in terms of the number of parameters used, it uses only $\frac{1}{3}$ of parameters of Conv2D, which is more efficient with only 1% drop in terms of test accuracy. LightLayers with hyperparameter $k = 3$ achieves the highest test accuracy. However, for the other values of k as well there is only minimal variation in test accuracy.

Table 2 shows the results for different layers for the model trained on the Fashion MNIST dataset. From the table, we can observe that the proposed model (LightLayers) with hyperparameter $k = 5$ uses only half of the parameters with around 3% drop in terms of test accuracy with the Fashion MNIST dataset. However, when we compare the quantitative results with SeperableConv2D, our

Table 3. Evaluation results on test set of **CIFAR10** dataset (Number of epochs = 20, Batch size = 64, Learning rate = 1e-4, Number of filters = $[8, 16, 32, 64]$). The 'Params' in the bold represents total number of parameters.

Method	Parameters	Test accuracy	**Test loss**
Conv2D	76,794	0.6882	0.9701
SeparableConv2D	14,440	0.5953	1.3263
LightLayers ($K = 1$)	5,937	0.3686	1.6723
LightLayers ($K = 2$)	9,592	0.4596	1.5372
LightLayers ($K = 3$)	13,247	0.4937	1.5287
LightLayers ($K = 4$)	16,902	0.5319	1.3214
LightLayers ($K = 5$)	**20,557**	**0.5576**	**1.2122**

Table 4. Evaluation on **CIFAR100** test set (Number of epochs = 20, Batch size = 64, Learning rate = $1e - 4$, Number of filters = $[8, 16, 32, 64]$).

Method	Parameters	Test accuracy	Test loss
Conv2D	82,644	0.3262	2.6576
SeparableConv2D	20,290	0.2207	3.2108
LightLayers ($K = 1$)	6,747	0.0275	4.2391
LightLayers ($K = 2$)	10,402	0.0398	4.1836
LightLayers ($K = 3$)	14,057	0.0559	4.0304
LightLayers ($K = 4$)	17,712	0.0551	3.9978
LightLayers ($K = 5$)	**21,367**	**0.0589**	**4.0009**

proposed LightLayers achieves better test accuracy with the trade-off in number of trainable parameters.

Table 3 shows the results on the CIFAR 10 dataset. On this dataset as well, the proposed method is 3.75 times computationally efficient in terms of parameters it uses. However, there is a drop in accuracy of around 13%. Nevertheless, for some tasks the efficiency can be more important than the reduced accuracy.

Similarly, we have trained and tested the proposed model on the CIFAR 100 dataset, where the test accuracy of the proposed layers is much lower as compared to the Conv2D. This is obvious because CIFAR 100 consists of 100 classes of images that are difficult to generalize with such a small number of trainable parameters. However, the total number of parameters used is still around 4 times less than that of Conv2D. The total number of trainable parameters for Conv2D is 82,644, and for LightLayers, it is only 21,367. We refer to Table 4 for more details on the test accuracy and test loss.

From the experimental results, we are convinced that LightLayers has the following advantages:

- It requires less trainable parameters than Conv2D, which is an important factor to implement in different applications where heavy trainable parameters could not be beneficial.
- Due to less parameters, the space taken by the weight file is smaller, which makes it more suitable to devices where storage space is limited.

6 Ablation Study

Let us consider that the input size is 784, and the number of output features is 10. Therefore, the weight matrix W is 784×10 resulting in $7,840$ trainable parameters. Now, in the LightDense layer, we decompose the weight matrix W into two smaller matrix $W1$ and $W2$ of lower dimension using the hyperparameter k.

Here, $W1 = [784, k]$ and $W2 = [k, 10]$ values from the above example, the total number of trainable parameters in the LightDense layer becomes $786 \times k + k \times 10$. Now, if $k = 1$, then trainable parameters are 796, and if $k = 2$ the number of trainable parameters becomes $1,588$, and so on.

Next, consider the weight decomposition in the LightConv2D layer. If the input is $32 \times 32 \times 3$, the number of filters is 32, and the kernel size is 3×3, then the filters size becomes $3 \times 3 \times 3 \times 32$. This means that the total number of trainable parameters is 864. Now, we will decompose the kernel W into $W1$ and $W2$ using hyperparameter k. Here, $W1$ is $3 \times 3 \times 3 \times k$ and $W2$ is $3 \times 3 \times k \times 32$. If k is 1, then the total number of trainable parameters becomes $27 + 288$, which is equal to 315.

From the ablation study, we deduce that the number of trainable parameters used is less in LightLayers compared to the Conv2D and Dense layers. Overall, we can argue that the proposed LightLayers approach has the potential to be a powerful solution to solve the problem of excessive parameter used by traditional DL approaches. However, our LightLayers model needs further improvement for successfully implementing it on a larger dataset with high resolution images. We can conclude that further investigating matrix weight decomposition is important and other similar studies are necessary to reach the goal of lightweight models in the near future.

7 Conclusion

In this paper, we propose the LightLayers model, which uses matrix decomposition to help to reduce the complexity of the DLN. With the extensive experiments, we observed that changing the value of hyperparameter k yields a tradeoff between model complexity in terms of the number of trainable parameters and performance. We compare the accuracy of the LightLayers model with Conv2D. An extensive evaluation shows the tradeoffs in terms of parameter uses, accuracy, and computation. In the future, we want to train LightLayers on other publicly available datasets. We also aim to develop efficient techniques for finding the optimal value of k automatically. Further research will be required to find suitable algorithms and implementations that will scale this approach to a biomedical dataset.

References

1. Abadi, M., Barham, P., et al.: Tensorflow: a system for large-scale machine learning. In: Proceedings of the Symposium on Operating Systems Design and Implementation (OSDI), pp. 265–283 (2016)
2. Agarwal, P., Alam, M.: A lightweight deep learning model for human activity recognition on edge devices. arXiv preprint arXiv:1909.12917 (2019)
3. Arcadu, F., Benmansour, F., Maunz, A., Willis, J., Haskova, Z., Prunotto, M.: Deep learning algorithm predicts diabetic retinopathy progression in individual patients. NPJ Digit. Med. **2**(1), 1–9 (2019)
4. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020)
5. Chollet, F., et al.: Keras (2015). https://keras.io
6. Denton, E.L., Zaremba, W., Bruna, J., LeCun, Y., Fergus, R.: Exploiting linear structure within convolutional networks for efficient evaluation. In: Advances in Neural Information Processing Systems, pp. 1269–1277 (2014)
7. Ding, C., et al.: Circnn: accelerating and compressing deep neural networks using block-circulant weight matrices. In: Proceedings of the IEEE/ACM International Symposium on Microarchitecture, pp. 395–408 (2017)
8. Garipov, T., Podoprikhin, D., Novikov, A., Vetrov, D.: Ultimate tensorization: compressing convolutional and fc layers alike. arXiv preprint arXiv:1611.03214 (2016)
9. Jiang, W., Xie, Z., Li, Y., Liu, C., Lu, H.: Lrnnet: A light-weighted network with efficient reduced non-local operation for real-time semantic segmentation. In: Proceedings of International Conference on Multimedia & Expo Workshops (ICMEW), pp. 1–6 (2020)
10. Kim, H., Sim, J., Choi, Y., Kim, L.S.: A kernel decomposition architecture for binary-weight convolutional neural networks. In: Proceedings of the Annual Design Automation Conference, pp. 1–6 (2017)
11. Kim, Y.D., Park, E., Yoo, S., Choi, T., Yang, L., Shin, D.: Compression of deep convolutional neural networks for fast and low power mobile applications. arXiv preprint arXiv:1511.06530 (2015)
12. Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M., Klein, M.: Logistic regression (2002)
13. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
15. Lebedev, V., Ganin, Y., Rakhuba, M., Oseledets, I., Lempitsky, V.: Speeding-up convolutional neural networks using fine-tuned cp-decomposition. arXiv preprint arXiv:1412.6553 (2014)
16. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
17. Li, J., Zhao, R., Huang, J.T., Gong, Y.: Learning small-size dnn with output-distribution-based criteria. In: Proceedings of the Conference of the International Speech Communication Association (2014)
18. Mariet, Z., Sra, S.: Diversity networks: Neural network compression using determinantal point processes. arXiv preprint arXiv:1511.05077 (2015)

19. McKinney, S.M., et al.: International evaluation of an AI system for breast cancer screening. Nature **577**(7788), 89–94 (2020)
20. Novikov, A., Podoprikhin, D., Osokin, A., Vetrov, D.P.: Tensorizing neural networks. In: Advances in Neural Information Processing Systems, pp. 442–450 (2015)
21. Oseledets, I.V.: Tensor-train decomposition. SIAM J. Sci. Comput. **33**(5), 2295–2317 (2011)
22. Paszke, A., Chaurasia, A., Kim, S., Culurciello, E.: Enet: A deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147 (2016)
23. Thambawita, V., et al.: An extensive study on cross-dataset bias and evaluation metrics interpretation for machine learning applied to gastrointestinal tract abnormality classification. ACM Trans. Comput. Healthcare 1(3), 1–29 (2020)
24. Wang, Y., et al.: Lednet: a lightweight encoder-decoder network for real-time semantic segmentation. In: Proceedings of International Conference on Image Processing (ICIP), pp. 1860–1864 (2019)
25. Wu, C.W.: Prodsumnet: reducing model parameters in deep neural networks via product-of-sums matrix decompositions. arXiv preprint arXiv:1809.02209 (2018)
26. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017)
27. Xue, J., Li, J., Gong, Y.: Restructuring of deep neural network acoustic models with singular value decomposition. In: Interspeech, pp. 2365–2369 (2013)
28. Xue, J., Li, J., Yu, D., Seltzer, M., Gong, Y.: Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6359–6363 (2014)
29. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 325–341 (2018)

A.10 Paper X : A comprehensive analysis of classification methods in gastrointestinal endoscopy imaging

Authors: D. Jha, S. Ali, S. Hicks, V. Thambawita, H. Borgli, P. H. Smedsrud, T. d. Lange, K. Pogorelov, X. Wang, P. Harzig, M. Tran, W. Meng, T. Hoang, D. Dias, T. H. Ko, T. Agrawal, O. Ostroukhova, Z. Khan, M. A. Tahir, Y. Liu, Y. Chang, M. Kirkerød, D. Johansen, M. Lux, H. D Johansen, M. A Riegler, and P. Halvorsen

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 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.

Published: Medical Image Analysis, Volume 70, May 2021, 102007

Candidate contributions: D. Jha conceptualized the work. This paper was written and revised mostly by D. Jha and S. Ali, with input from all the co-authors. He led the work and performed all the analysis with the input from S. Ali and all other

co-authors. He also led the revision and provided critical insight to the manuscript to prepare a final version.

Thesis objectives: Objective II

Medical Image Analysis 70 (2021) 102007

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 Pogorelov[i], Xiaowei Wang[j], Philipp Harzig[k], Minh-Triet Tran[l], Wenhua Meng[m], Trung-Hieu Hoang[l], Danielle Dias[n], Tobey H. Ko[o], Taruna Agrawal[p], Olga Ostroukhova[q], Zeshan Khan[r], Muhammad Atif Tahir[r], Yang Liu[s], Yuan Chang[t], Mathias Kirkerød[i], 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, Norway
[i] 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, China
[n] 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).

https://doi.org/10.1016/j.media.2021.102007

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.

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,

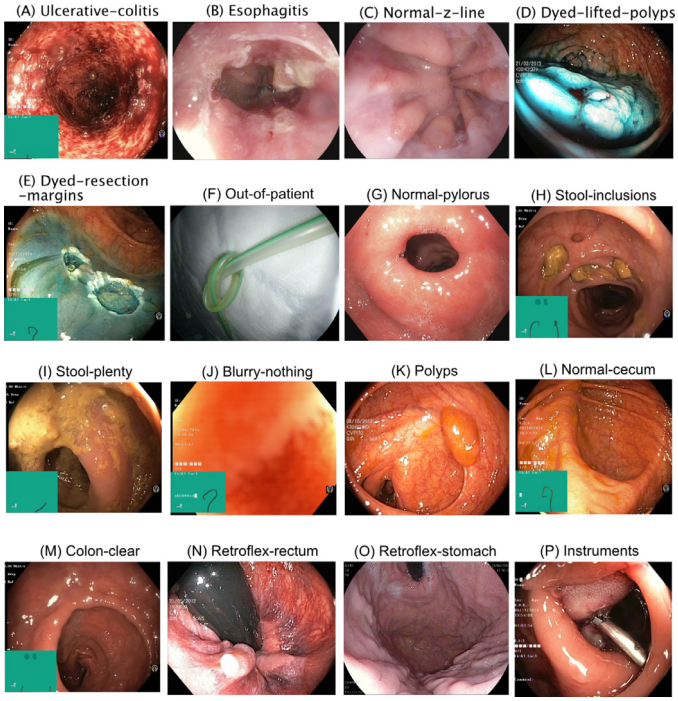


Fig. 1. Examples images from the 16 classes of Medico 2018 and BioMedia 2019 dataset.

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 Biomedia 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; Chheda 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/biomedia-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 detection[2]). 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 2015[3]. 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. Bernel 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 classes[5]. 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 challenge[6] 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 built 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 Biomedia 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,

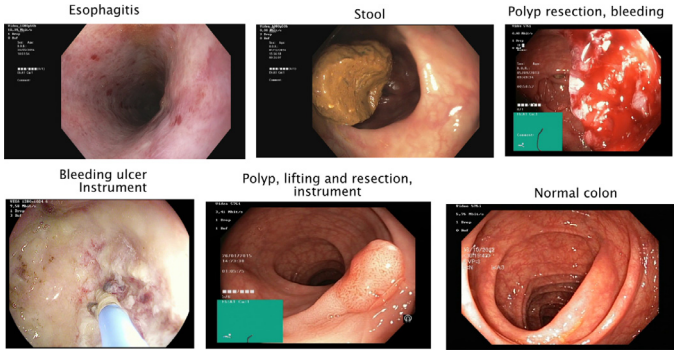


Fig. 2. Example of extracted frame from each of the 6 videos provided to the participants for for automatic report generation task.

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

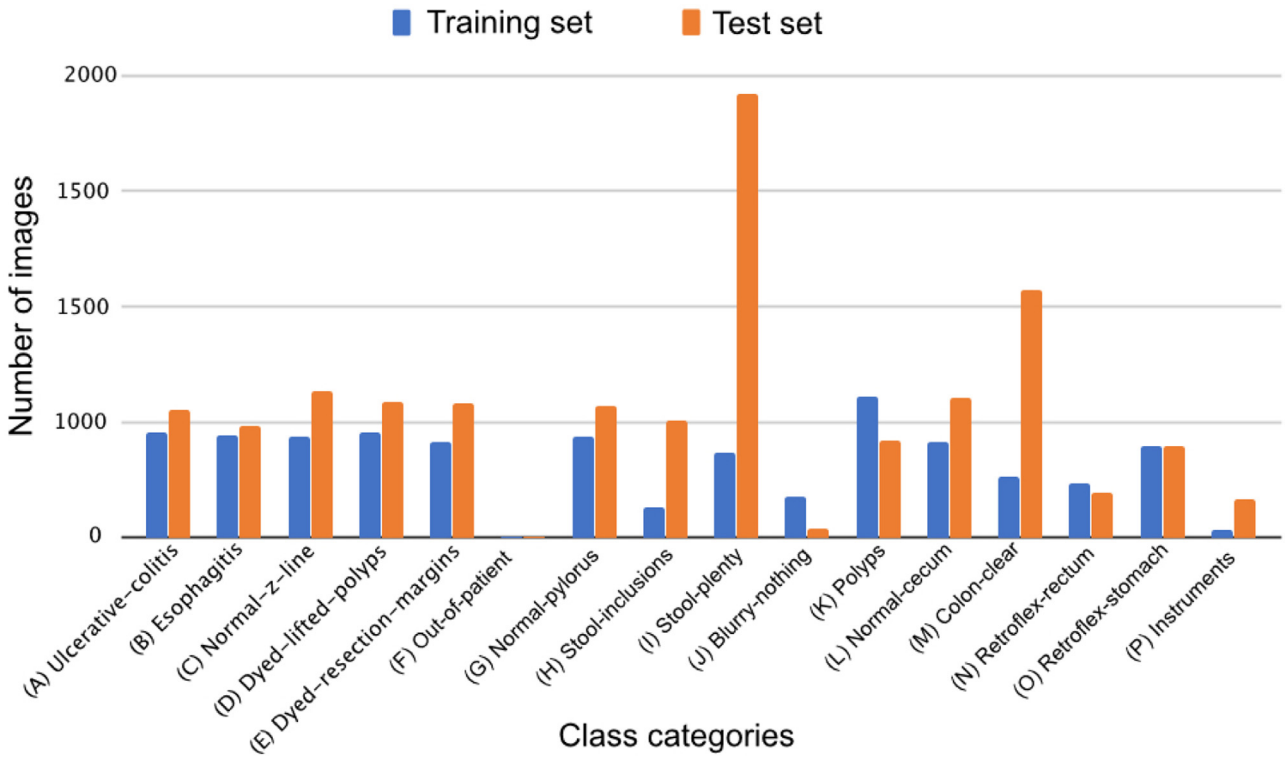


Fig. 3. Summary of the Medico 2018 and BioMedia 2019 dataset.

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)).

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

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

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

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

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

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

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

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

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 descriptions 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/biomedia-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	-	-	-
	HCMUS	X	X	-	-
2019	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 orgnaizers, 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 + νSVM	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
	ParaNoMundo (Dias and Dias, 2018a)	DenseNet	DenseNet-201	General	Accuracy	No	Cross-entropy	SGD	N/A
Medico 2018	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
	CIISR (Meng et al., 2019)	ResNet + Softmax	ResNet-50	General	Accuracy-speed	Yes	Cross-entropy	Adam	Tesla P4
Biomedia 2019	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 multilayer 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 comparison. 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
AAUITEC (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 discussions 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 a 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. Additionally, 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 be 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

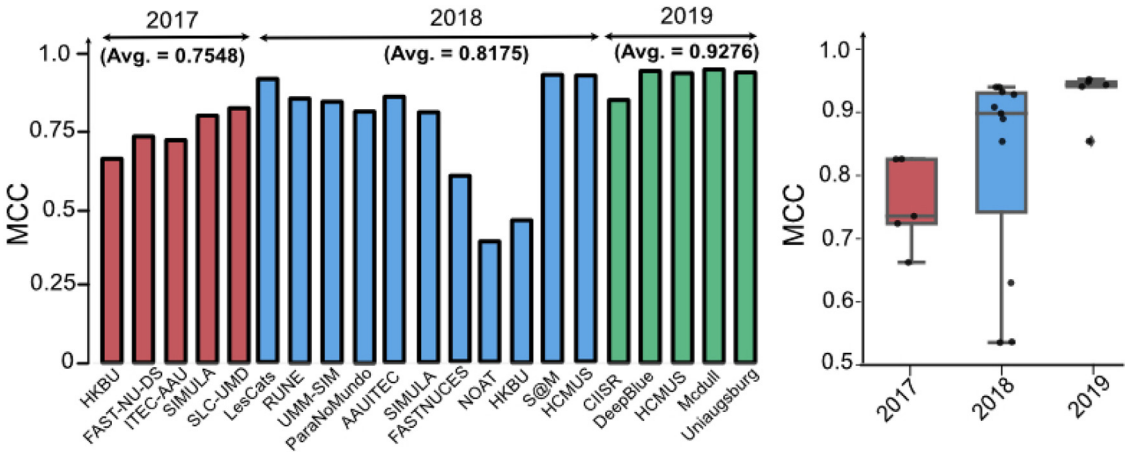


Fig. 5. MCC score comparison of different participating teams in Medico 2017, Medico 2018, and BioMedia 2019 challenges. On left individual team scores (bar plot), and on the right statistics of each year submission (box plot).

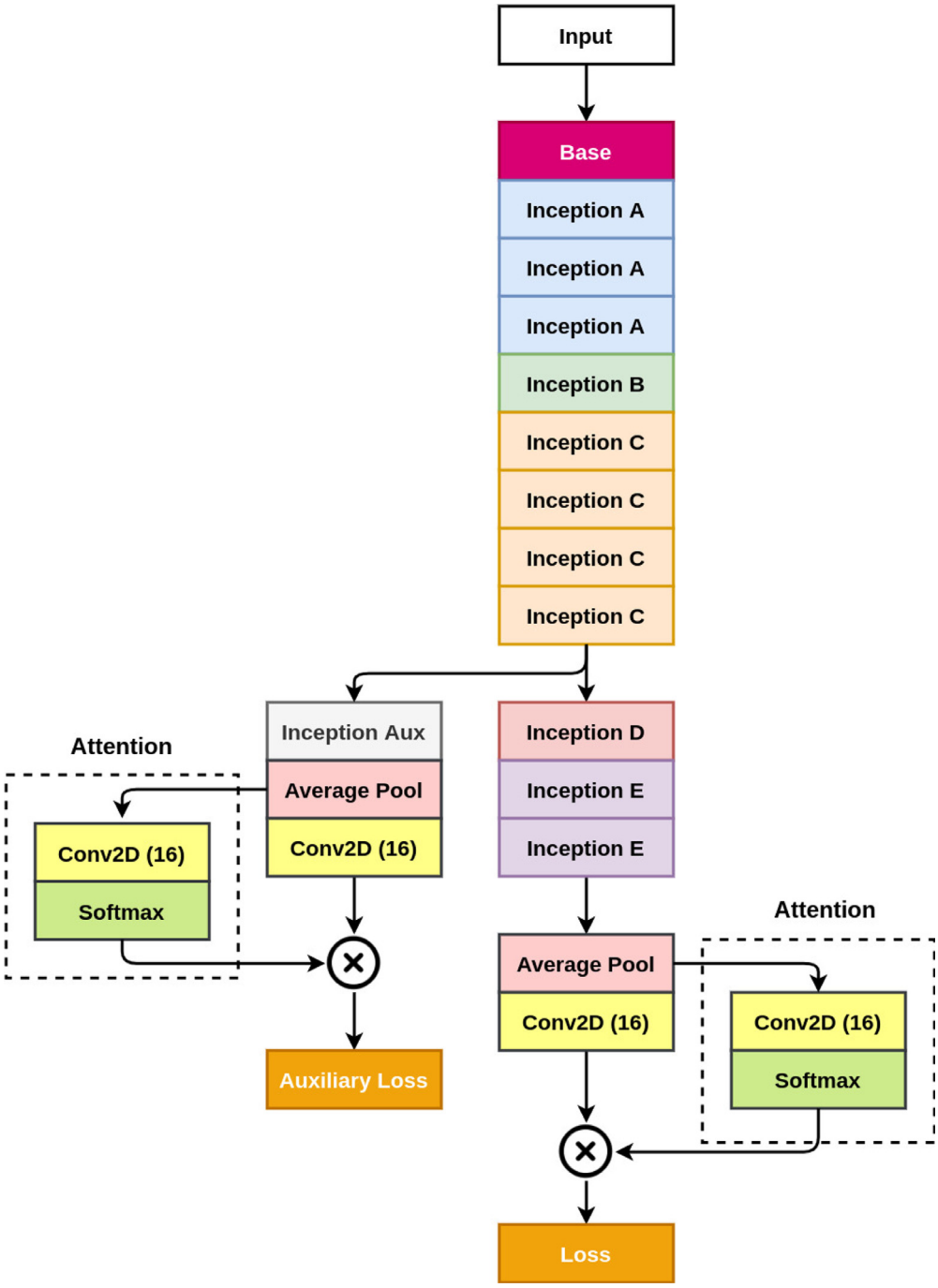


Fig. 6. Best architecture in Medico 2018 "classification task" **Team Mcdull** (Chang et al., 2019).

(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 SIMULA 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

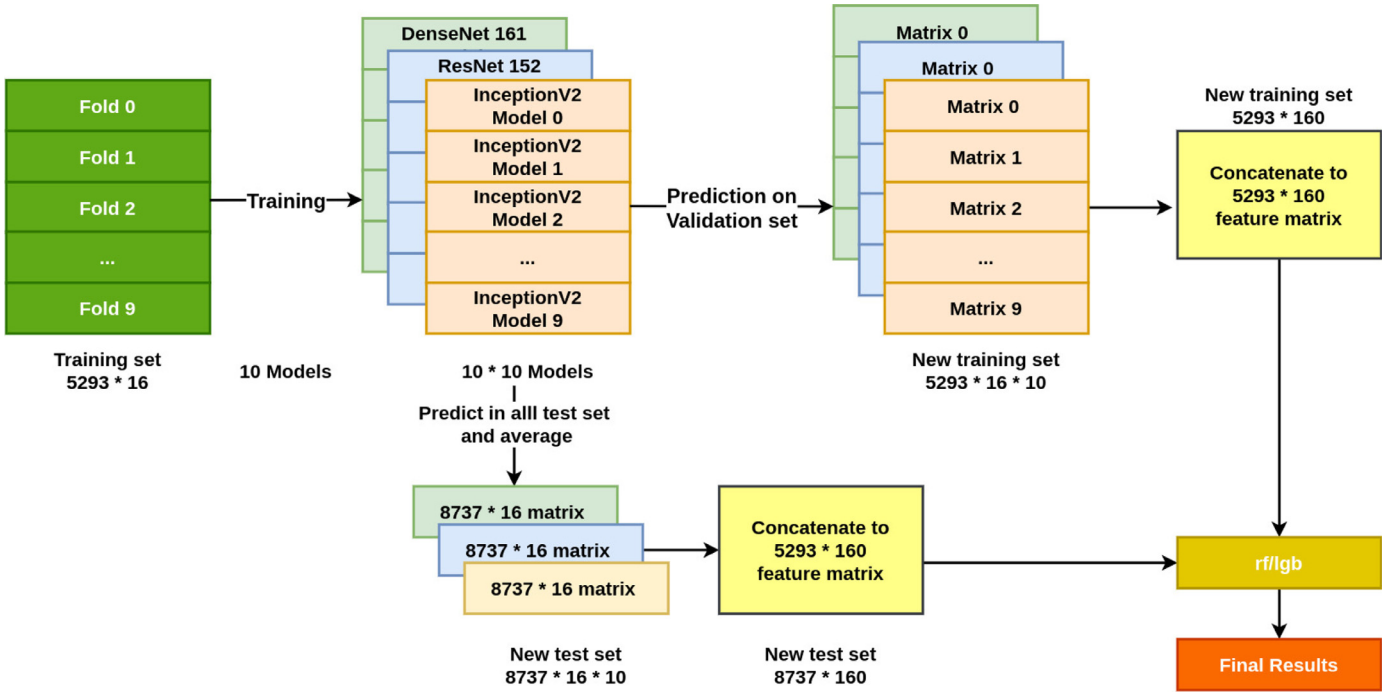


Fig. 7. The architecture of the best performing team in Biomedia 2019 challenge (Team DeepBlue (Luo et al., 2019)).

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 DeeepBlue 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.

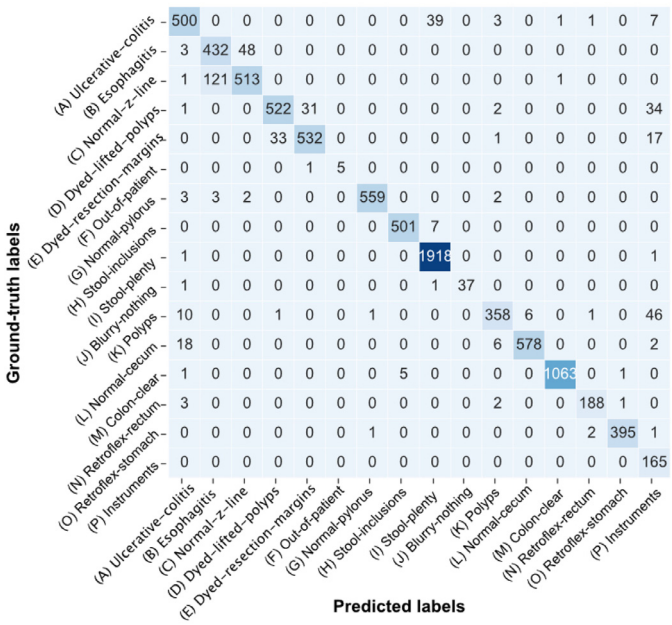


Fig. 8. Confusion matrix plot of Team S@M (Thambawita et al., 2020). A-P represents class labels.

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**
2019	**CIISR**	**0.8542**	**0.8542**	**98.9**	**1**	**1**	**12**	**3**
	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 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.

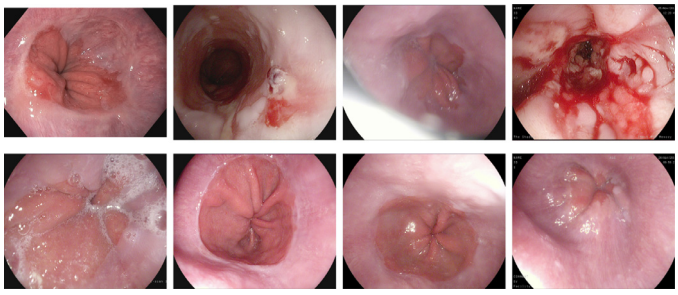


Fig. 9. Example frames from 'Esophagitis' and 'Normal-zline' class.

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 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., Mohrekesh, 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 Clini. 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. Surgi. 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.
Chheda, 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 biomedia 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 Inter. 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 biomedia 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 Biophysi. Acta (BBA)-Prot. Struct. 405 (2), 442–451.
Meng, W., Zhang, S., Yao, X., Yang, X., Xu, C., Huang, X., 2019. Biomedia 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).
Petscharnig, 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., Næss, 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. Artifi. 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. Healt. Servic. 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.

A.11 Paper XI : Exploring Deep Learning Methods for Real-Time Surgical Instrument Segmentation in Laparoscopy

Authors: D. Jha, S. Ali, N. K. Tomar, M. A. Riegler, D. Johansen, H. D. Johansen, P. Halvorsen

Abstract: Minimally Invasive Surgery (MIS) is a surgical intervention used to examine the organs inside the abdomen and has been widely used due to its effectiveness over open surgery. Due to the hardware improvements such as high definition cameras, this procedure has significantly improved and new software methods have open potential for computer-assisted procedures. However, there exists challenges and requirement to improve detection and tracking of the position of the instruments during these surgical procedures. To this end, we evaluate and compare some popular deep learning methods that can potentially be explored for the automated segmentation of surgical instruments in laparoscopy, an important step towards tool tracking. Our experimental results demonstrate that the Dual encoder attention network (DDANet) produces a superior result compared to other recent deep learning methods. DDANet produces a dice coefficient of 0.8739 and mean intersection over union of 0.8183 for the Robust Medical Instrument Segmentation (ROBUST-MIS) Challenge 2019 dataset, at a real-time speed of 101.36 frames per second which is critical for such procedures.

Published: Proceedings of the IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), 2021.

Candidate contributions: D. Jha conceptualized this work and performed all the experiments in the paper. He prepared the manuscript and performed all the evaluation and analysis. Additionally, he made a subsequent revision to the manuscript with the input from all of the co-authors and presented it at the conference.

Thesis objectives: Objective III

A.12 Paper XII : HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy

Authors: H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. Dang Nguyen, D. Johansen, C. Griwodz, H. K Stensland, E. Garcia-Ceja, P. T. Schmidt, H. L Hammer, M. A Riegler, P Halvorsen, T. d. Lange

Abstract: Artificial intelligence is currently a hot topic in medicine. However, medical data is often sparse and hard to obtain due to legal restrictions and lack of medical personnel for the cumbersome and tedious process to manually label training data. These constraints make it difficult to develop systems for automatic analysis, like detecting disease or other lesions. In this respect, this article presents HyperKvasir, the largest image and video dataset of the gastrointestinal tract available today. The data is collected during real gastro- and colonoscopy examinations at Bærum Hospital in Norway and partly labeled by experienced gastrointestinal endoscopists. The dataset contains 110,079 images and 374 videos, and represents anatomical landmarks as well as pathological and normal findings. The total number of images and video frames together is around 1 million. Initial experiments demonstrate the potential benefits of artificial intelligence-based computer-assisted diagnosis systems. The HyperKvasir dataset can play a valuable role in developing better algorithms and computer-assisted examination systems not only for gastro- and colonoscopy, but also for other fields in medicine.

Published: Scientific Data

Candidate contributions: D. Jha contributed to the conception and design of the manuscript. He contributed to dataset annotation, cleaning, and organizing the dataset. He also participated in drafting and revision of the manuscript.

Thesis objectives: Objective I

SCIENTIFIC DATA

Check for updates

OPEN

DATA DESCRIPTOR

HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy

Hanna Borgli[1,3,15], Vajira Thambawita[1,2,15], Pia H. Smedsrud[1,3,6,15], Steven Hicks[1,2,15], Debesh Jha[1,7,15], Sigrun L. Eskeland[4], Kristin Ranheim Randel[3,10], Konstantin Pogorelov[8], Mathias Lux[11], Duc Tien Dang Nguyen[5], Dag Johansen[7], Carsten Griwodz[3], Håkon K. Stensland[3,8], Enrique Garcia-Ceja[13], Peter T. Schmidt[9,14], Hugo L. Hammer[1,2,15], Michael A. Riegler[1,15,16], Pål Halvorsen[1,2,15,16] ✉ & Thomas de Lange[4,6,12,15,16]

Artificial intelligence is currently a hot topic in medicine. However, medical data is often sparse and hard to obtain due to legal restrictions and lack of medical personnel for the cumbersome and tedious process to manually label training data. These constraints make it difficult to develop systems for automatic analysis, like detecting disease or other lesions. In this respect, this article presents *HyperKvasir*, the largest image and video dataset of the gastrointestinal tract available today. The data is collected during real gastro- and colonoscopy examinations at Bærum Hospital in Norway and partly labeled by experienced gastrointestinal endoscopists. The dataset contains 110,079 images and 374 videos, and represents anatomical landmarks as well as pathological and normal findings. The total number of images and video frames together is around 1 million. Initial experiments demonstrate the potential benefits of artificial intelligence-based computer-assisted diagnosis systems. The *HyperKvasir* dataset can play a valuable role in developing better algorithms and computer-assisted examination systems not only for gastro- and colonoscopy, but also for other fields in medicine.

Background & Summary

The human gastrointestinal (GI) tract is subject to numerous different abnormal mucosal findings ranging from minor annoyances to highly lethal diseases. For example, according to the International Agency for Research on Cancer (https://gco.iarc.fr/today/fact-sheets-cancers), the specialized cancer agency of the World Health Organization (WHO), GI cancer globally accounts for about 3.5 million new cases each year. These cancer types usually have combined mortality of about 63% and 2.2 million deaths per year[1–3].

Endoscopy is currently the gold-standard procedure for examining the GI tract, but its effectiveness is considerably limited by the variation in operator performance[4–6]. This causes, for example, an average 20% polyp miss-rate in the colon[7]. Thus, improved endoscopic performances, high-quality clinical examinations, and systematic screening are significant factors to prevent GI disease-related morbidity and deaths. The recent rise of artificial intelligence (AI)-enabled support systems has shown promise in giving healthcare professionals the tools needed to provide quality care at a large scale[8,9]. The core of an efficient AI-based system is the combination of quality data and algorithms which teach a model to solve real-world problems like detecting pre-cancerous lesions or cancers in images. Today's AI-based systems are predominantly using a subfield of AI called machine learning (ML), which usually requires training on thousands of data samples to perform well on any given task.

[1]SimulaMet, Oslo, Norway. [2]Oslo Metropolitan University, Oslo, Norway. [3]University of Oslo, Oslo, Norway. [4]Department of Medical Research, Bærum Hospital, Bærum, Norway. [5]University of Bergen, Bergen, Norway. [6]Augere Medical AS, Oslo, Norway. [7]UIT The Arctic University of Norway, Tromsø, Norway. [8]Simula Research Laboratory, Oslo, Norway. [9]Department of Medicine (Solna), Karolinska Institutet, Stockholm, Sweden. [10]Cancer Registry of Norway, Oslo, Norway. [11]Klagenfurt University, Klagenfurt, Austria. [12]Medical Department, Sahlgrenska University Hospital-Mölndal, Mölndal, Sweden. [13]SINTEF Digital, Oslo, Norway. [14]Department of Medicine, Ersta hospital, Stockholm, Sweden. [15]These authors contributed equally: Hanna Borgli, Vajira Thambawita, Pia H. Smedsrud, Steven Hicks, Debesh Jha, Hugo L. Hammer, Michael A. Riegler, Pål Halvorsen, Thomas de Lange. [16]These authors jointly supervised this work: Michael A. Riegler, Pål Halvorsen, Thomas de Lange. ✉e-mail: paalh@simula.no

Dataset	Findings	Size	Availability
CVC-356[18]	Polyps	356 images†	by request●
CVC-ClinicDB[19] (also named CVC-612)	Polyps	612 images†	open academic
CVC-VideoClinicDB[18] (also named CVC-12k)	Polyps	11954 images†	by request●
CVC-ColonDB[62]	Polyps	380 images†ψ	by request●
Endoscopy Artifact detection 2019[63]	Endoscopic Artifacts	5,138 images	open academic
ASU-Mayo polyp database[20]	Polyps	18,781 images†	by request●
ETIS-Larib Polyp DB[64]	Polyps	196 images†	open academic
KID[65]◊	Angiectasia, bleeding, inflammations, polyps	2371 images and 47 videos	open academic●
GIANA 2017[66]◊	Polyps & Angiodysplasia	3462 images and 38 videos	by request
GIANA 2018[67,68]◊	Polyps & Small bowel lesions	8262 images and 38 videos	by request
GASTROLAB[69]	GI lesions	Some 100s of images and few videos	open academic♣
WEO Clinical Endoscopy Atlas[70]	GI lesions	152 images	by request♣
GI Lesions in Regular Colonoscopy Data Set[71]	GI lesions	76 images†	by request
Atlas of Gastrointestinal Endoscope[72]	GI lesions	1295 images	unknown●
El salvador atlas of gastrointestinal video endoscopy[73]	GI lesions	5071 video clips	open academic♣
Kvasir[22]	Polyps, esophagitis, ulcerative colitis, Z-line, pylorus, cecum, dyed polyp, dyed resection margins, stool	8000 images	open academic
Kvasir-SEG[49]	Polyps	1000 images†	open academic
Nerthus[74]	Stool - categorization of bowel cleanliness	21 videos	open academic

Table 1. An overview of existing GI datasets. †Including ground truth segmentation masks. ◊Video capsule endoscopy. ●Not available anymore. ψContour. ♣Not really a dataset usable for machine learning. It is more a medical atlas or database for education with a several low-quality samples of various findings in the GI tract.

However, health data is often sparse and hard to obtain due to legal constraints and structural problems in data collection. Nevertheless, an increasing number of promising AI solutions aimed for diagnostics in endoscopy[10–17] are being developed. The datasets used for these systems, like CVC[18,19] and the ASU-Mayo polyp database[20], are rather small in the context of ML research. In other non-medical ML areas, datasets such as ImageNet[21] consists of 14 million images. Table 1 gives an overview of all, to the best of our knowledge, existing datasets of images and videos from the human GI tract. As can be observed, they are rather small, and often limited to polyps. Several of these have also lately become unavailable.

The images and videos in *HyperKvasir* were collected prospectively from routine clinical examinations performed at a Norwegian hospital from 2008 to 2016. We retrieved the images from the Picsara image documentation database (CSAM, Norway), a plug-in to the electronic medical record system, in 2016. As a first step, 4,000 of these images were labeled into eight different classes by medical experts and published as the Kvasir dataset[22]. The dataset was later extended to 8,000 images. Using Kvasir, researchers all over the world have started developing different ML models and AI systems for GI endoscopy[23–38]. Moreover, the Kvasir dataset has been used to organize international competitions, i.e., the Medico Task at MediaEval in 2017[39] and 2018[40] and the ACM Multimedia 2019 BioMedia Grand Challenge[41].

Based on the lessons learned from publishing the Kvasir dataset and organizing competitions, it became clear that one of the biggest challenges in medical AI is still data availability. Data is hard to retrieve from the health care systems, approvals from medical committees are hard to get, medical experts have limited time, and there are no efficient tools to label such data. Therefore, with *HyperKvasir*, we significantly increase both the amount of labeled medical data for supervised learning and also release a large amount of unlabeled data. The new dataset contains 110,079 images and 374 videos from various GI examinations, resulting in 1 million images and frames in total. Regarding the value of unlabeled data, recent work in the ML community has shown remarkable improvements to tackle the challenge of lack of data. Instead of learning from a large set of annotated data, algorithms can now learn from sparsely labeled and unlabeled data. This technique is known as semi-supervised learning and has lately been successfully applied in different medical image analyses[42]. Examples of semi-supervised learning are self-learning[43,44] and neural graph learning[45], which both make use of unlabeled data in addition to a small number of labeled data to extract additional information[43,44,46]. We believe these new algorithms might be the development needed to make AI even more useful for medical applications. The unlabeled data of *HyperKvasir* is intended to be used in medical and technical communities to explore semi-supervised and unsupervised methods, and users of the data might even consider employing their own local experts to provide labels . Subsequently, in addition to the data description, we provide a baseline analysis using the labeled classes of the dataset and feasible future research directions for researchers interested in using the dataset.

Methods

The image and video data were collected using standard endoscopy equipment from Olympus (Olympus Europe, Germany) and Pentax (Pentax Medical Europe, Germany) at the Department of Gastroenterology, Bærum Hospital, Vestre Viken Hospital Trust, Norway. Vestre Viken provides health care services to 490,000 people, of which 189,000 are covered by Bærum hospital. Parts of the collected data were annotated with class labels and segmentation masks. The annotations were done by at least one experienced gastroenterologist from Bærum

hospital, the Cancer Registry of Norway or Karolinska University Hospital in Sweden, and one or more experienced persons working in the medical field such as a junior doctor or Ph.D. student. Though several physicians have assessed each labeled data record of the dataset, there is a chance that some of the assessments might be biased by the well-known observer variation, particularly regarding subtle changes like low-grade reflux esophagitis and ulcerative colitis. Such discrepancies have been demonstrated in previous studies[47,48]. To tackle this further, we decided to combine some of the findings that are prone to bias into one class (details about the classes and combinations can be found in the data records descriptions). Finally, a large number of unlabeled images are provided.

The study was approved by Norwegian Privacy Data Protection Authority and exempted from patient consent because the data were fully anonymous. All metadata was removed, and all files renamed to randomly generated file names before the internal IT department at Bærum hospital exported the files from a central server. The study was exempted from approval from the Regional Committee for Medical and Health Research Ethics - South East Norway since the collection of the data did not interfere with the care given to the patient. Since the data is anonymous, the dataset is publicly shareable based on Norwegian and General Data Protection Regulation (GDPR) laws. Apart from this, the data has not been pre-processed or augmented in any way.

Class labels per image. The method for labeling images can be split into three distinct steps. First, experienced gastroenterologists involved in the project decided which classes should be included in the labeling process, based on medical relevance and the data collected. The selected classes were described in detail by medical experts. Second, two junior doctors or Ph.D. students working in the field annotated a subset of the images to the provided classes. Once this pre-labeling step was done, the medical experts checked the labels and adjusted when necessary. Cases where no consent could be found were discarded and replaced with new images from the dataset. The first dataset we created consisted of 4,000 images from eight classes[22]. This was later extended to 8,000 images for the same eight classes. For *HyperKvasir*, the dataset is further extended to 10,662 images and 23 classes. In total, *HyperKvasir* contains 110,079 images (10,662 labeled and 99,417 unlabeled images) from the GI tract.

Segmentation masks per image. *HyperKvasir* includes images with corresponding segmentation masks and bounding boxes for 1,000 images from the polyp class. To create the segmentation masks, we uploaded 1,000 polyp images to the Labelbox platform (https://www.labelbox.com/). Labelbox is a tool that allows pixel-accurate labeling of image regions. First, a junior doctor and a Ph.D. student pre-segmented the 1,000 images. A gastroenterologist subsequently went trough all images verifying and correcting the pre-labeled segmentation masks. A detailed description of the annotation process and an example use-case of the dataset can be found in[49,50].

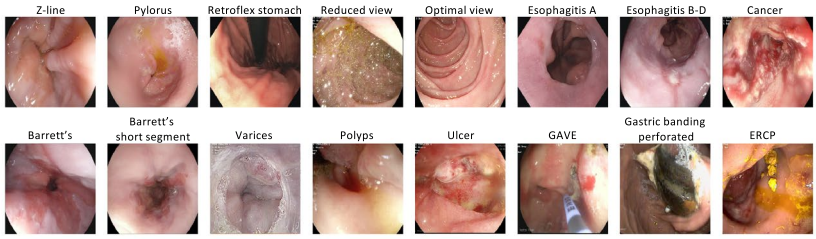
Descriptions per video. To get the labels per video, we uploaded the video data to a video annotation platform provided by Augere Medical AS (Oslo, Norway). Each video was analyzed and labeled by an experienced gastroenterologist. The class labels selected by the gastroenterologist were representing the main finding in the video as accurately as possible. For example, if the video contained footage of a polyp, the label for that video would be polyp. Additionally, there are examples of multiple findings in the same video. If so, these and more detailed descriptions are included in the *video-labeling.csv* file.

Data Records

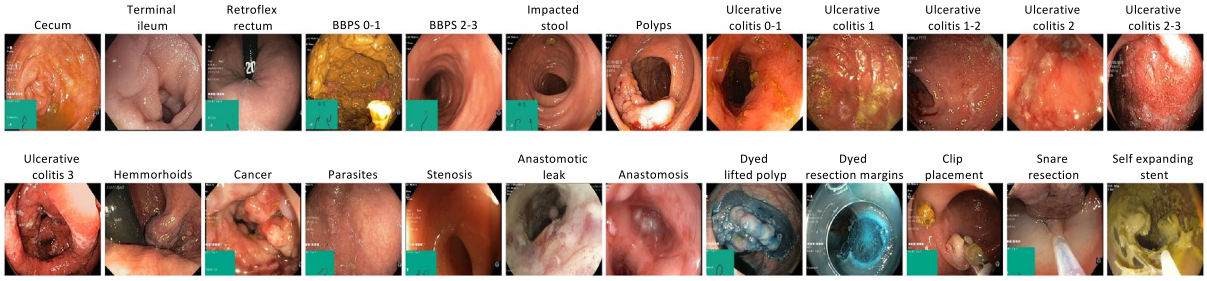
The full *HyperKvasir*[51] dataset, with all its images, videos and metadata, is available from the Open Science Framework (OSF) via the link https://doi.org/10.17605/OSF.IO/MH9SJ. The dataset is also available at https://datasets.simula.no/hyper-kvasir. *HyperKvasir* is open access and licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0). In total, the dataset consists of four main data records. The records are labeled images, segmented images, unlabeled images, and labeled videos. All the various labeled classes are shown in Fig. 1, i.e., 16 classes from the upper GI tract (Fig. 1a) and 24 classes from the lower GI tract (Fig. 1b). The dataset has a size of circa 66.4GB (not including metadata files and segmentation masks), 32.5GB for videos and 33.9GB for images. An overview of all data records in the dataset is given in Table 2. Some of the images and videos contain a picture in picture (green thumbnail in the lower left corner) which represents the Olympus ScopeGuide™ (Olympus Europe, Germany), used by the endoscopist to get a topographic view of the colon. Details about image and video resolutions and video frame rates can be found in the Figs. 2 and 3. The following subsections provide additional details for each data record.

Labeled images. In total, the dataset contains 10,662 labeled images stored using the JPEG format, where Fig. 4 shows the 23 different classes representing the labeled images and the number of images in each class. A CSV file is provided (image-labels.csv) giving the mapping between the image (file name) and the labeling for each image. These classes are structured according to location in the GI tract and the type of finding as shown in Fig. 5. We defined four main categories of findings where the first and the third are found both in the upper an lower GI tract:

- **Anatomical landmarks:** Anatomical landmarks are characteristics of the GI tract used for orientation during endoscopic procedures. Furthermore, they are used to confirm a complete extent of the examination. Landmarks exist both in the upper GI tract (esophagus, stomach and duodenum) and in the lower GI tract (terminal ileum, colon and rectum). However, in the small bowel, there are no specific landmarks to be used for topographical localization of a lesion.
- **Quality of mucosal views:** Complete visualization of all the mucosa is crucial not to overlook pathological findings. In the colon, there exist a classification for the quality of the mucosal vizualisation, the Boston Bowel Preparation Scale (BBPS)[52].



(a) Upper GI tract.



(b) Lower GI tract.

Fig. 1 Image examples of the various labeled classes for images and/or videos.

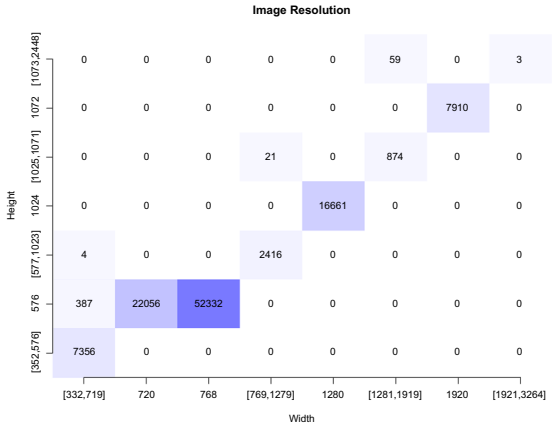


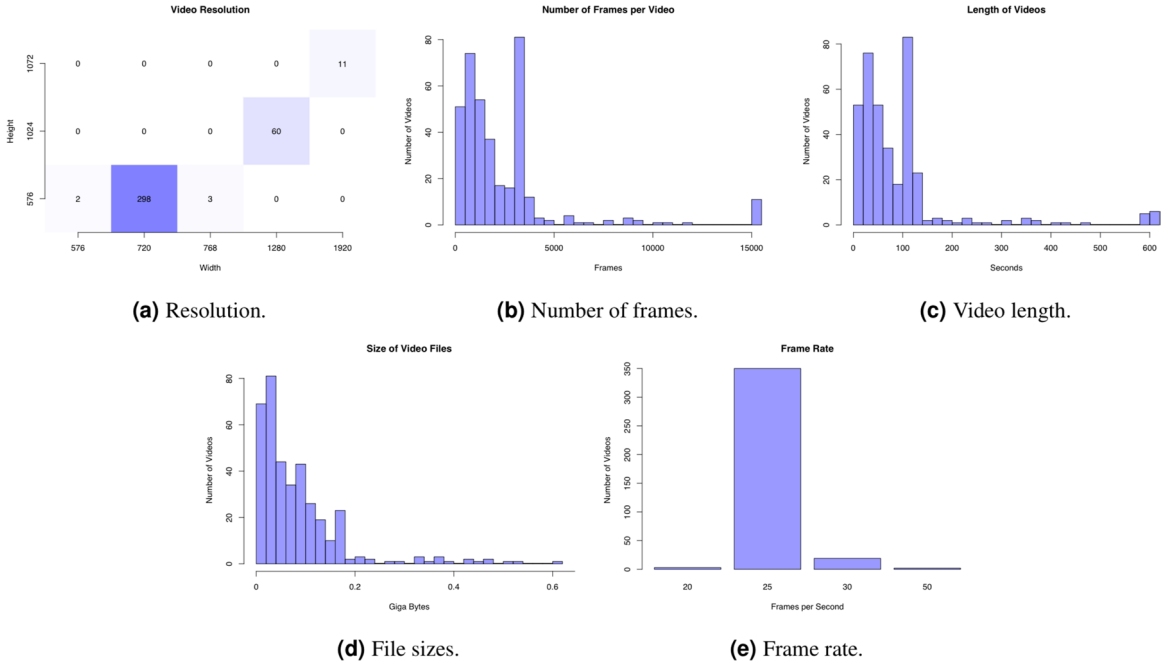
Fig. 2 Resolution of the 110,079 images in *HyperKvasir*.

Data Record	# Files	Description	Size (MB)
Labeled images	10,662 images	23 classes of findings	3,960
Segmented Images	1,000 images	Segmentation masks for polyp findings	57
Unlabeled Images	99,417 images	Unlabeled	29,940
Videos	374 videos	30 different classes	32,539

Table 2. Overview of the data records in the *HyperKvasir* dataset.

- **Pathological findings:** All parts of the gastrointestinal tract can be affected by abnormalities or findings due to disease. Most pathological findings can be seen as more or less obvious changes in the intestinal wall mucosa. These findings are classified according to the Minimal Standard Terminology, defined by the World Endoscopy Organization[53].
- **Therapeutic interventions:** When a lesion or pathological finding is detected, a therapeutic intervention is frequently required to treat the condition, e.g., lifting and resecting a polyp, dilation of a stenosis or injection of a bleeding ulcer.

Each class and the images belonging to it is stored in the corresponding folder of the category it belongs to. For example, the 'polyp' folder in the category pathological findings in the lower GI tract contains all polyp images,



(a) Resolution. **(b)** Number of frames. **(c)** Video length.

(d) File sizes. **(e)** Frame rate.

Fig. 3 Statistics of the 374 videos in *HyperKvasir*.

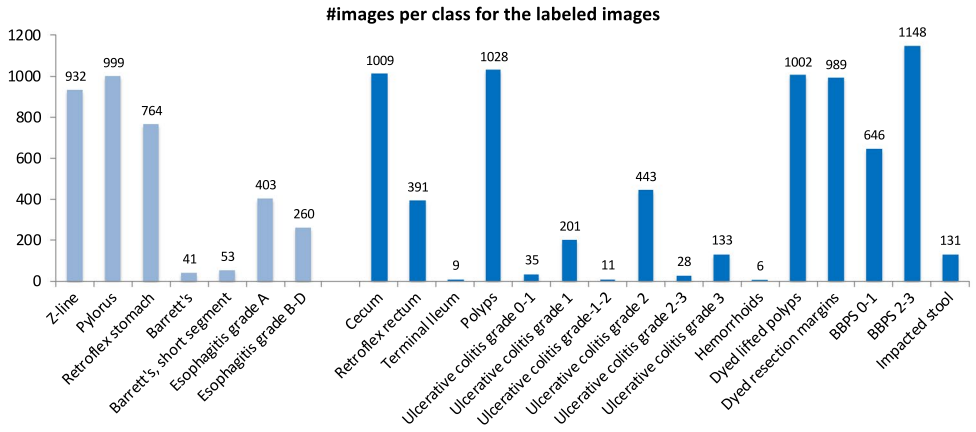


Fig. 4 The number of images in the various *HyperKvasir* labeled image classes according to the file folders.

the 'barrett's' folder in the category pathological findings in the upper GI tract contains all images of Barrett's esophagus, etc. As observed in Fig. 2, the number of images per class are not balanced, which is a general challenge in the medical field due to the fact that some findings occur more often than others. This adds an additional challenge for researchers, since methods applied to the data should also be able to learn from a small amount of training data. Below, we detail each class further.

Upper Gastrointestinal tract. The upper GI tract examined by endoscopy includes the esophagus, stomach, and duodenum. Below, we give a description of the various classes of findings found here.

As seen in Fig. 5, we have labeled three classes of *anatomical landmarks* in the upper GI tract. The normal **Z-line** is the anatomical junction between the squamous epithelium of the esophagus and columnar epithelium of the stomach. A normal Z-line is located at the same level as the gastroesophageal junction. **Retroflex stomach** means that the endoscope is retroflexed, looking back to visualize the cardia and fundus in the upper parts of the stomach. The **pylorus** is the anatomical junction between the stomach and duodenal bulb, and a sphincter regulating the emptying process of the stomach into the duodenum.

All the following classes are defined as *pathological findings* in the upper GI tract. Reflux esophagitis is an inflammation mostly affecting the lower third of the esophagus, near the Z-line. Reflux esophagitis can be graded according to the Los Angeles (LA) classification[54]. The esophagitis LA classification is defined into four classes as (A) mucosal breaks shorter than 5mm, without continuity across mucosal folds where subtle changes can be difficult to differentiate from a normal Z-line; (B) mucosal breaks longer than 5mm that does not extend between the

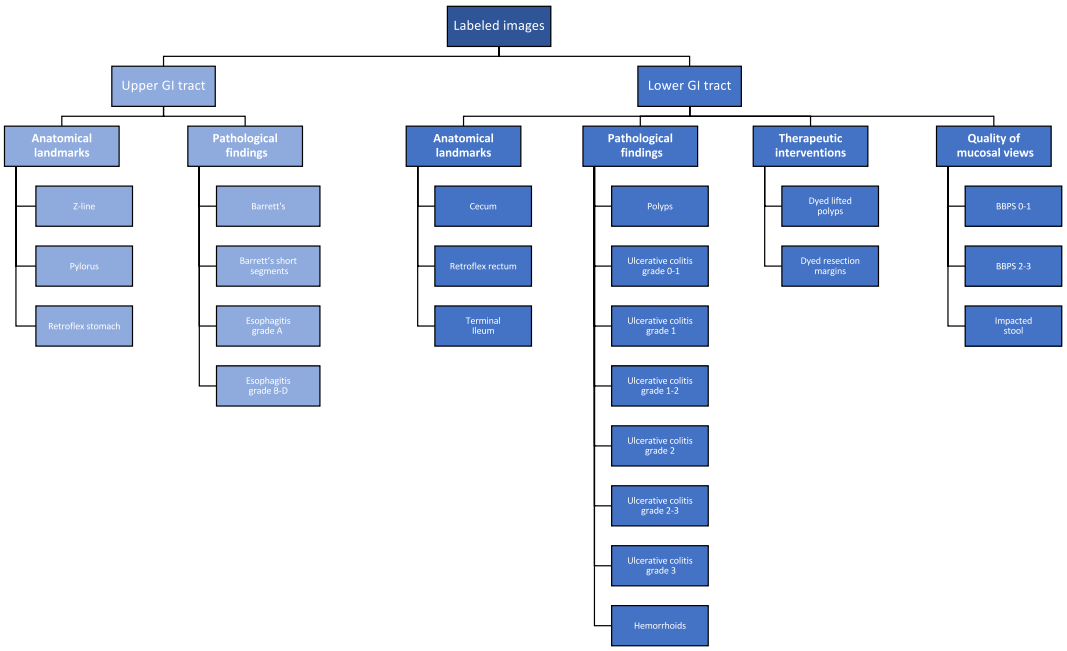


Fig. 5 The various image classes structured under position and type, also the structure of the stored images.

tops of two mucosal folds; (C) one (or more) mucosal break that is continuous between the tops of two or more mucosal folds, but which involves less than 75% of the circumference; and (D) one (or more) mucosal break that is continuous between the tops of two or more mucosal folds and involves more than 75% of the circumference. We have split esophagitis into two classes because there exists an important observer variation in the assessment of low grade esophagitis[47]. The two classes are **esophagitis A** and **esophagitis B-D**. This binary classification of the images makes it possible to assess whether mis-classification between normality and esophagitis only concern grade A. Barrett's esophagus represents a metaplastic transformation of the squamous epithelium of the esophagus into a gastric like columnar epithelium. Barrett's esophagus is considered a premalignant condition, meaning it might develop into cancer. Biopsies showing the presence of specialized intestinal metaplasia confirms the diagnosis. Barrett's esophagus can be graded according to the Prague classification, describing the circumferential and longitudinal extension of the disease[55]. We have split the images of Barrett's esophagus into two classes. **Barrett's** long-segment and **Barrett's, short-segment** esophagus where a short segment is characterized by a longitudinal extension of less than 3 cm[55].

Lower gastrointestinal tract. The lower GI tract examined by colonoscopy includes the terminal ileum (last part of the small bowel), the colon and the rectum (the large bowel). Below, we describe the classes of the lower GI tract in the dataset.

We have labeled three classes of *anatomical landmarks* in the lower GI tract. The ileum is the distal 2/3 of the small bowel, recognized by visible intestinal villi. Endoscopically, the ileum cannot be distinguished from other parts of the small bowel. During colonoscopy, the distal 5–20 cm of the ileum, named **terminal ileum**, can be reached and examined. The visualization of the terminal ileum confirms complete colonoscopy. **Cecum** is the proximal end of the large bowel and is characterized by the visualization of the appendiceal orifice and the ileo-cecal valve. Complete examination of the whole colon can only be confirmed if the medial wall of the cecum has been visualized, that is the area between the appendiceal orifice and the ileo-cecal valve. The most distal part of the rectum is one of the blind zones of the colon. Therefore, the endoscope is retroflexed in the rectum to visualize the dentate line and the circumference of the proximal orifice of the anal canal, which is called **retroflex rectum**.

The *quality of the mucosal views* is a key quality indicator and should always be evaluated because a clean bowel is essential to detect pathological findings. In this respect, the degree of bowel cleansing during a colonoscopy is described by the Boston Bowel Preparation Scale (BBPS)[56]. BBPS consists of four different degrees which are: (BBPS 0) unprepared colon segment with no mucosa seen due to solid stool that cannot be cleared; (BBPS 1) portions of the mucosa of the colon segment seen, but other areas of the colon segment not well seen due to staining, residual stool and/or opaque liquid; (BBPS 2) minor amount of small fragments of stool and/or opaque liquid, but mucosa of colon segment seen well; and (BBPS 3) entire mucosa of colon segment seen well with no residual fragments of stool or opaque liquid. The bowel cleansing is deemed adequate if the BBPS score is 2 or 3 in all three segments of the colon after flushing. Therefore, we have labeled our images into the two **BBPS 0-1** and **BBPS 2-3** classes where class 0–1 represents inadequate bowel preparations, and the class 2–3 represents adequate bowel preparation. Moreover, a frequent finding in persons above the age of 50 years are pockets in the colon wall called diverticula and if numerous called diverticulosis. Sometimes stool is impacted in these diverticula and may increase the risk of diverticulitis. In the dataset, this is presented in the **impacted stool** class.

The following classes are defined as *pathological findings* in the lower GI tract. Ulcerative colitis is a chronic inflammatory bowel disease often debuting in the twenties. The degree and extent of the disease is determined by colonoscopy and can be classified according to the Mayo Score[57]. The Mayo Score for ulcerative colitis is defined: (Score 0) inactive, where the mucosa only has normal vascular patterns; (Score 1) mild with erythema, decreased vascular pattern, mild friability; (Score 2) moderate with erythema, absent vascular pattern, mild friability, erosions; and (Score 3) severe with spontaneous bleeding and ulcerations. For ulcerative colitis, we provide six different labeled classes, both the Mayo Score classes (**Ulcerative colitis 1/2/3**) and some classes in-between where it is difficult to determine the exact class and because previous studies have shown important observer variation in the assessment of the degree of inflammation (**Ulcerative colitis 0-1/1-2/2-3**)[48]. **Polyps** are most frequently neoplastic lesions of the large bowel. They have mainly three different shapes; protruding in the lumen, flat or excavated according to the Paris Classification[58]. Their size vary from 1 mm to several cm. The prevalence increases with age. The most common types of polyps are premalignant and can transform into cancer. Thus, it is important to discover polyps and remove the suspicious polyps during endoscopy. **Hemorrhoids** are pathologically swollen veins in the anus or lower rectum. When present in the rectum, they are called internal hemorrhoids, and when found in the anus, they are called external hemorrhoids.

Finally, *therapeutic interventions* show treatments of detected pathological findings. It includes for example lifting and removal of neoplastic tissue (polyps) and injection therapy of bleeding ulcer. The **dyed lifted polyps** class contains images of polyps lifted with submucosal injection using a solution containing indigo carmine. This is done prior to polyp resection for better diagnosis and easier resection. The dye is recognized by the blue color underneath the polyp. After resection of dyed polyps with a snare, the resection margins and site appears blue due to the indigo carmine solution. Images of these type of resection margin are presented in the **dyed resection margins** class.

Segmented images. For the set of segmented images, we provide the original image, a segmentation mask and a bounding box for 1,000 images from the polyp class. In the mask, the pixels depicting polyp tissue, the region of interest, are represented by the foreground (white mask), while the background (in black) does not contain polyp pixels. The bounding box is defined as the outermost pixels of the found polyp. For this segmentation set, we have two folders, one for images and one for masks, each containing 1,000 JPEG-compressed images. The bounding boxes for the corresponding images are stored in a JavaScript Object Notation (JSON) file. The image and its corresponding mask have the same filename. The images and files are stored in the segmented images folder. It is important to point out that the images of polyp class from the Kvasir dataset had duplicates in the images folder. These duplicates were replaced by high-quality polyp images from the colon and segmented.

Unlabeled images. In total, the dataset contains 99,417 unlabeled images. The unlabeled images can be found in the unlabeled folder which is a subfolder in the image folder, together with the other labeled image folders. In addition to the unlabeled image files, we also provide the extracted global features and possible unsupervised clustering assignments in the *HyperKvasir* Github repository as Attribute-Relation File Format (ARFF) files. ARFF files can be opened and processed using, for example, the WEKA machine learning library, or they can easily be converted into Comma-Separated Values (CSV) files.

Labeled videos. The labeled videos are recorded for clinical purposes and thus represent daily practice. In total, 374 videos are provided in the dataset, which correspond to 9.78 hours of videos and 889,372 video frames that can be converted to images if needed. In total, an experienced gastroenterologist have identified 30 classes of findings, and Fig. 6 shows how many videos we have identified for each class. The class describes the video as a whole using the main finding, but additionally, many videos include more than one category and several classes where, for example, a single video can contain polyps, dyed lifted polyps and dyed resection margins. The video file format is Audio Video Interleave (AVI), and they are stored in the folder called labeled videos. As seen in Fig. 7, the videos are further organized and stored according to either upper or lower GI tract and then the four main categories as for the labeled images described above. In addition to the video files, a CSV file is provided (video-labels.csv) containing the videos' *videoID* and *labeling*. Here, the VideoID contains the corresponding video file name, and the labeling includes the upper or lower location, the category and the class with some detailed descriptions of the video. Below, we describe the new classes per category for the in total 60 videos from the upper GI tract and the 60 videos from the lower GI tract.

Upper Gastrointestinal tract. As seen in Fig. 7, we have many of the same classes for videos and for images, but since we have labeled all our videos, more classes are added for both the upper and lower GI-tract. In the upper GI tract, the three classes of *anatomical landmarks* (**Z-line**, **Pylorus** and **Retroflex stomach**) are described in the section for labeled images above. In the category of *pathological findings*, both **Barrett's** esophagus and **esophagitis** are also described above, but here we also added some new classes. The first is **polyps** where the description above of polyps in the colon is also valid for the upper GI-tract. In addition, five new classes not previously described are included. Mucosal **ulcers** are quite common in the upper GI tract. Ulcers are nearly always caused by Helicobacter pylori infection, ulcerogenic medication, or cancer. Ulcers are characterized according to the Forrest classification to predict the risk of bleeding[59]. Forrest I represents ongoing bleeding, Forrest II presents some signs of previous bleeding; and Forrest III does not show any sign of bleeding. The second class **Gastric antral vascular ectasia (GAVE)** represents dilated small superficial vessels in the mucosa of the gastric antrum. These lesions may cause chronic bleeding and subsequent anemia and are frequently treated by argon plasma coagulation (APC) to prevent further bleeding. **Varices** (dilated veins) in both the esofagus and the fundus of the stomach are most frequently caused by chronic liver diseases complicated with liver cirrhosis. The varices represent a major risk for severe bleeding. **Cancer** of the esophagus and the stomach are common findings in the upper

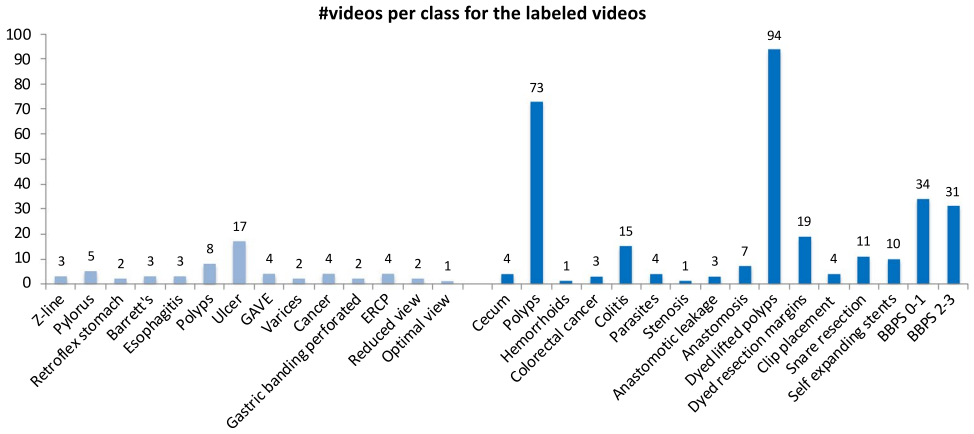


Fig. 6 The number of videos in the various *HyperKvasir* labeled video classes according to the file folders.

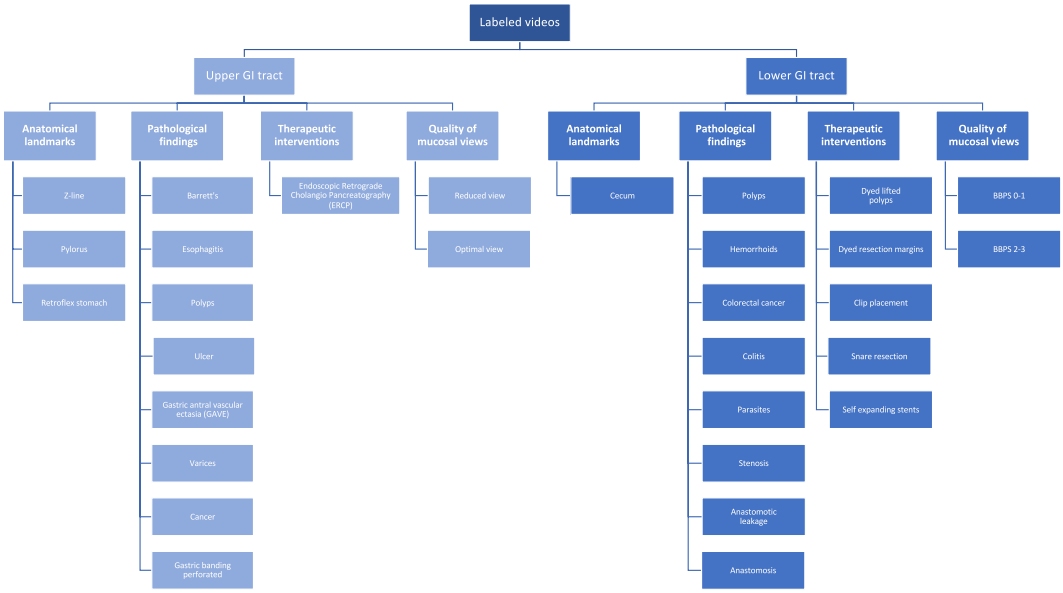


Fig. 7 The various video classes structured under position and type, which is also the structure of the video folders.

GI-tract. The last class **gastric banding perforated** shows a rare finding, which is the complication of previous gastric banding operation where the band perforates the wall of the stomach. The category of *therapeutic interventions* are introduced for the Upper GI-tract especially because they are nearly always best illustrated by videos and can also serve important educational purposes. Since most of the *therapeutic interventions* are presented as secondary to a pathological finding we only include **Endoscopic Retrograde Choleangio-Pancreatografi (ERCP)** a procedure to treat gall-duct abnormalities as an independent class. However, other common therapeutic interventions such as the two thermal methods; APC and heatherprobe as well as injection therapy with adrenaline and application of hemospray to stop bleeding can be found under second findings in the csv file. In the category *quality of mucosal view*, we also added a footage showing **reduced view** due to opaque liquid in the stomach or air bubbles in the duodenum. Reduced view increases the risk of missing lesions. In opposite, **optimal view** demonstrates excellent visualization of the duodenum.

Lower Gastrointestinal tract. The videos from the lower GI tract illustrate mainly the same categories and classes as the labeled images. Nevertheless, they increase the diversity of the dataset. The category *anatomical landmarks* differs from the labeled images as it only contains the **cecum** class and does not include the classes of terminal ileum and retroflex rectum, only defined as second findings. The two categories pathological findings and therapeutic intervention also are a bit different compared to the labeled images. In the category *pathological findings*, we still have the above described **polyps** and **hemorrhoids** classes. However, all classes of ulcerative colitis are merged to **colitis** and also includes ischemic colitis and infectious colitis. The new class **colorectal cancer**, the second most deadly cancer worldwide[60], was added. Colorectal cancer may present itself in different ways in the colon, from tiny lesions with a diameter of 1 cm to larger tumors obstructing the entire lumen of the bowel and

cover bowel segments of several centimeters. Moreover, **parasites**, a common finding of small worms moving around in the colon, are more often encountered in tropical areas. **Stenosis** is characterized by a narrow obstruction of the bowel caused either by inflammation or malignant diseases. Large neoplastic lesions like cancers are surgically resected and subsequently an **anastomosis** is made to restore normal bowel function. The anastomosis can be visualized during follow-up colonoscopies. A feared complication after large bowel surgery is **anastomotic leakage**, potentially causing smaller or larger cavities of anastomotic leak especially in the rectum. The last decade mini-invasive endoscopic *therapeutic interventions* has to some extent replaced traditional and laparoscopic surgery regarding the treatment of large polyps and stenosis of the colon. The classes **dyed lifted polyp** and **dyed resection margin** are described under labeled images but videos improve the illustration of the technique. Three new classes are presented showing removal of polyps by simple **snare resection** or endoscopic mucosal resection (EMR). To prevent or stop bleeding after these resections, **clip placement** of metallic clips are illustrated. **Self expanding stents** are used to open and dilate either benign or malignant stenosis. Finally, in the *quality of mucosal views* category, we have removed the impacted stool class we have for images, and include only the above described **BBPS 0-1** and **BBPS 2-3** classes. Here, it is also worth noting that many of the videos in BBPS 2-3 are perfectly clean (BBPS 3), i.e., as then described in the csv-file, these contain videos of normal mucosa (also marked as finding 2) which can be extracted in normal images or videos when needed.

Technical Validation

To demonstrate the technical quality of the dataset, we performed multiple experiments to provide some baseline metrics and to give some insight into the dataset's statistical qualities. If the reader wants information about classification and segmentation approaches and experiments comparing state of the art methods using parts of this dataset, the reader is referred to other studies[49].

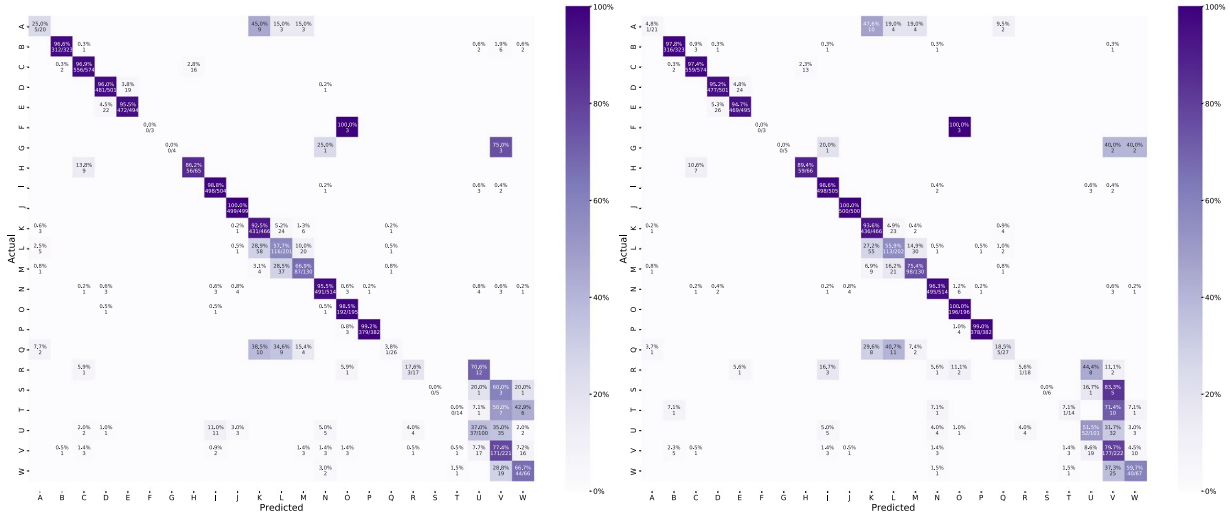
Baseline for supervised image classification. The presented dataset is suited for a variety of different tasks, one of which is image classification. As a preliminary step to evaluate how state-of-the-art methods perform on the labeled part of *HyperKvasir*, we performed a series of experiments based on methods that have previously achieved good results on GI tract image classification. The purpose of these experiments is merely to give example baseline results to be used by future researches to compare and measure their results. In total, we ran five experiments using different methods. The methods were primarily selected from the best performing methods presented at the MediaEval Medico Task[39,40]. Each method is based on deep convolutional neural networks, which is currently state-of-the-art within image classification. Common for all experiments is that the images were resized to 224×224 before being fed into the networks. All networks are based on common architectures, slightly modified to accommodate our task of classifying 23 different classes of images. The specifics of each method is further explained below:

- *Pre-Trained ResNet-50* is a TensorFlow implementation of the ResNet-50 architecture using ImageNet initialized weights. The network was trained in two steps. First, an initial training over 7 epochs, and then a fine-tuning step over 3 epochs which only trained the layers after the 100th index. Images were loaded using a batch size of 32, and the weights were optimized using Adam with a learning rate of 0.001.
- *Pre-Trained ResNet-152* is a PyTorch implementation of the ResNet-152 architecture using ImageNet initialized weights. The network was trained over 50 epochs using a batch size of 32, and optimized using Stochastic gradient descent (SGD) with a learning rate of 0.001. No fine-tuning was used for this method.
- *Pre-Trained DenseNet-161* is a PyTorch implementation of the standard DenseNet-161 architecture using ImageNet initialized weights. The network was trained over 50 epochs using a batch size of 32, and optimized using SGD with a learning rate of 0.001. No fine-tuning was used for this method.
- *Averaged ResNet-152 + DenseNet-161*[38,61] is an approach that combines the ResNet-152 and DenseNet-161 approach by averaging the output of both models as the final prediction. Both models were trained simultaneously by backpropagating the averaged loss through both models. Overall, the networks were trained for 50 epochs using a batch size of 32. SGD was used to optimize the weights with a learning rate of 0.001. Both the ResNet-152 and DenseNet-161 models were initialized using the best weights of the above Pre-Trained ResNet-152 and Pre-Trained DenseNet-161 implementations.
- *ResNet-152 + DenseNet-161 + MLP*[38,61] is similar to the previous method using both ResNet-152 and DenseNet-161 to generate a prediction. However, instead of averaging the output of each model, this method uses a simple multilayer perceptron (MLP) to estimate the best way to average the output of each model. All networks were trained simultaneously over 50 epochs using a batch size of 32. The weights were optimized using SGD with a learning rate of 0.001. Both the ResNet-152 and DenseNet-161 models were initialized using the best weights of the above two implementations of Pre-Trained ResNet-152 and Pre-Trained DenseNet-161.

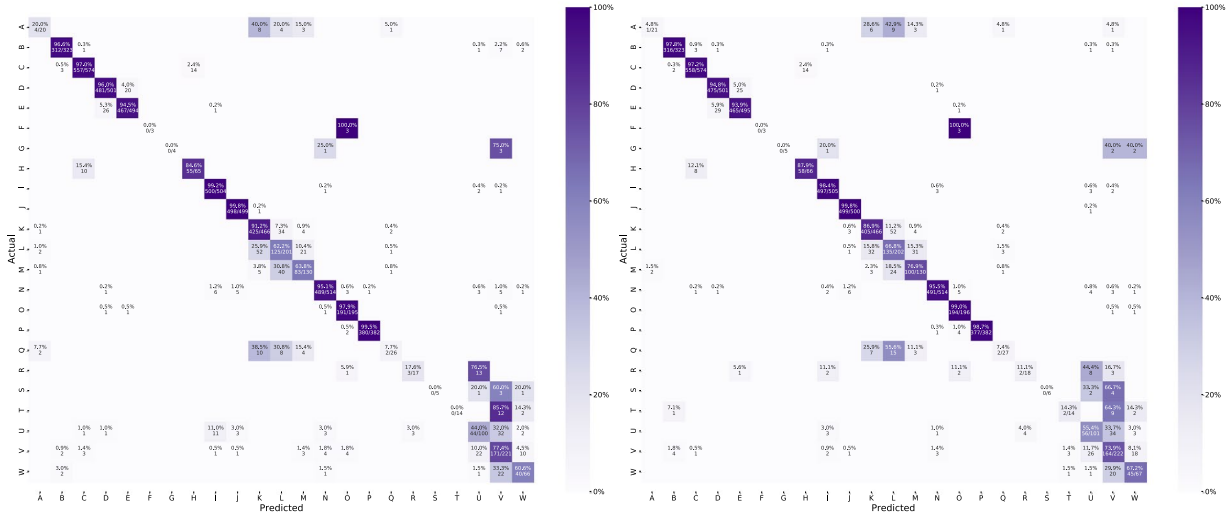
Each method was evaluated using standard classification metrics including the macro-averaged and micro-averaged F1-score, precision, and recall. Additionally, we calculated the Matthews correlation coefficient (MCC) for each experiment using the multi-class generalization which is also known as the R_K. The results in Table 3 show that each method beats the random and majority class baseline by a large margin. However, the presented numbers also indicate that there is room for improvement. Looking at the confusion matrices in Fig. 8, we see that some classes are harder to identify than others. For example, there is a lot of confusion surrounding the difference between the grades of ulcerative colitis and esophagitis. Furthermore, there is also some confusion between specific classes such as dyed lifted polyps and dyed resection margins, and distinguishing Barrett's from esophagitis or a normal Z-line. At least the confusion between classes of Z-line, esophagitis and Barrett's esophagus is similar to the human variation in the assessment of these lesions. Thus, it is challenging to create a ground truth.

Method	Macro Average			Micro Average			
	Precision	Recall	F1-score	Precision	Recall	F1-score	MCC (R_K)
Pre-Trained ResNet-50	0.589	0.536	0.530	0.839	0.839	0.839	0.826
Pre-Trained ResNet-152	0.639	0.605	0.606	0.906	0.906	0.906	0.898
Pre-Trained DensNet-161	0.640	0.616	0.619	0.907	0.907	0.907	0.899
Averaged ResNet-152 + DenseNet-161	**0.633**	**0.615**	**0.617**	**0.910**	**0.910**	**0.910**	**0.902**
ResNet-152 + DenseNet-161 + MLP	0.612	0.606	0.605	0.909	0.909	0.909	0.902
Random Guessing	0.044	0.038	0.034	0.044	0.044	0.044	0.000
Majority Class	0.004	0.043	0.008	0.108	0.108	0.108	N/A

Table 3. Average results for the five tested classification approaches, i.e., average of the results for the two splits.



(a) *Averaged ResNet-152 + DenseNet-161* evaluated on split 0. **(b)** *Averaged ResNet-152 + DenseNet-161* evaluated on split 1.



(c) *Pre-Trained DenseNet-161* evaluated on split 0. **(d)** *Pre-Trained DenseNet-161* evaluated on split 1.

Fig. 8 Confusion matrices for Averaged ResNet-152 + DenseNet-161 and Pre-Trained DenseNet-161 including both splits. These confusion matrices were selected based on their performance. Averaged ResNet-152 + DenseNet-161 achieved the best micro-averaged results while the Pre-Trained DenseNet-161 achieved the best macro-averaged result. The color codes represent the percentages of the total number of images within each class. The labeling of the classes is as follows: (A) Barrett's; (B) bbps-0-1; (C) bbps-2-3; (D) dyed lifted polyps; (E) dyed resection margins; (F) hemorrhoids; (G) ileum; (H) impacted stool; (I) normal cecum; (J) normal pylorus; (K) normal Z-line; (L) oesophagitis-a; (M) oesophagitis-b-d; (N) polyp; (O) retroflex rectum; (P) retroflex stomach; (Q) short segment Barrett's; (R) ulcerative colitis grade 0-1; (S) ulcerative colitis grade 1-2; (T) ulcerative colitis grade 2-3; (U) ulcerative colitis grade 1; (V) ulcerative colitis grade 2; (W) ulcerative colitis grade 3.

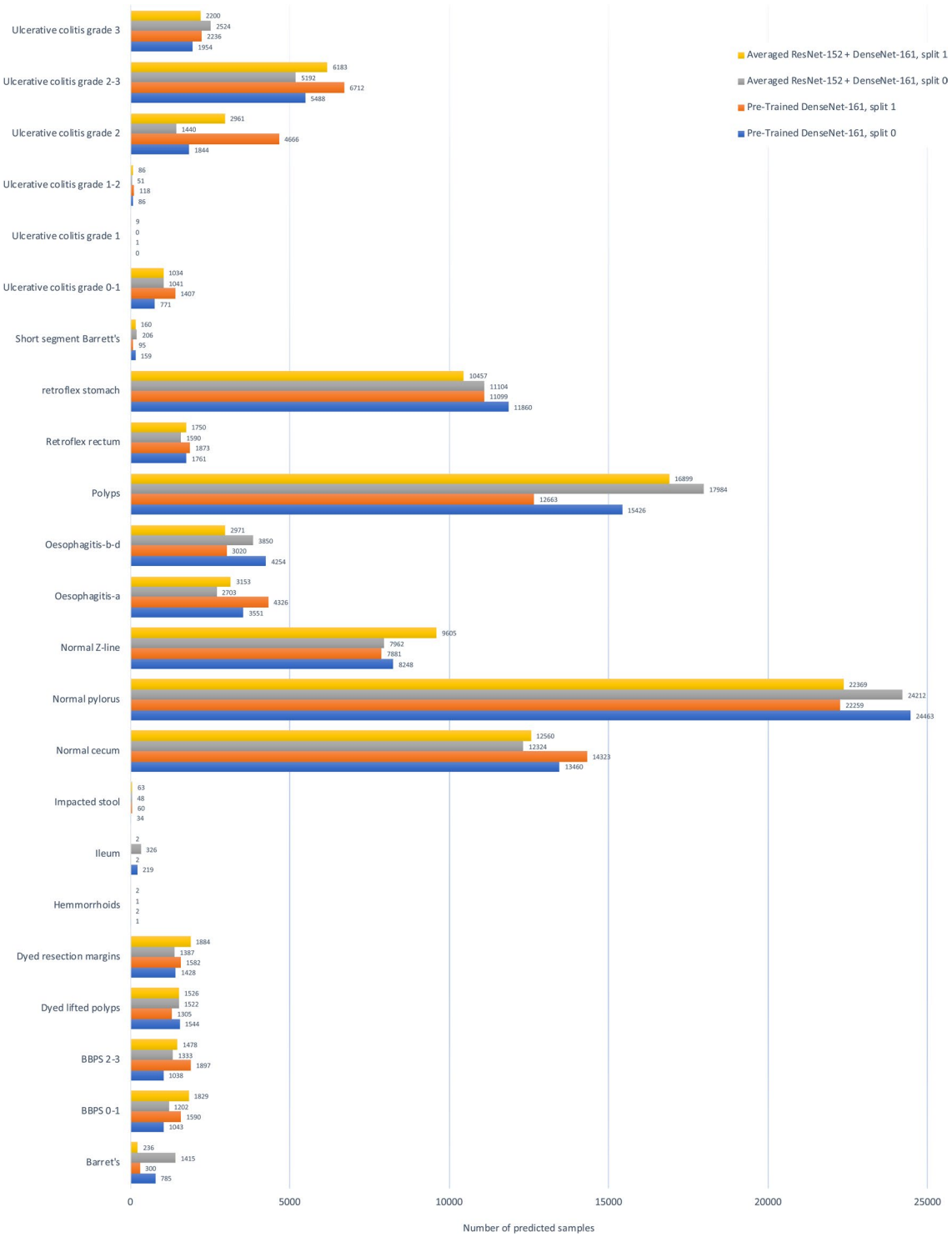


Fig. 9 Unlabeled image data predictions for *Averaged ResNet-152 + DenseNet-161* and *Pre-Trained DenseNet-161*.

Composition of unlabeled data. In order to show the approximate composition of the unlabelled data, we present some initial experiments to analyze the provided data which do not have annotated labels from medical experts We used our pre-trained classification model to simply classify the unlabeled data to indicate how many of the labeled classes are in the unlabeled data and to get an overall idea about data distribution of the 99,417 images. In particular, we used the best two classification models from the previous experiments, i.e., Pre-Trained DenseNet-161 and Averaged ResNet-152 + DenseNet-161 using split_0 and split_1 from the previous experiment. The result are shown in Fig. 9. In the results, we can observe that a large number of predictions are assigned to the class normal pylorus, while a smaller number of predictions are assigned to the classes hemorrhoids and ulcerative colitis grade 1-2. However, these predictions are similar to that of the class-level accuracies of the ML

model on the labeled data. Therefore, we can assume that the classes which achieved a high number of correct predictions on the labeled images are also more accurate on the unlabeled data. In contrast, it is hard to make any conclusions on the labels which had a low number of predictions as the models are not accurate enough. For future work, researchers could go trough the classifications of the unlabeled data and, for example, create a larger labeled dataset or perform failure analysis to find out why classes were confused or miss-classified. The class labels created during this experiments are available in the GitHub repository.

Validation Summary

In the technical validation section, we provided baseline metrics and gave insight into the dataset's statistical qualities to demonstrate it's technical quality. With the large number of images available in *HyperKvasir*, we encourage other researchers to investigate and develop new and improved methods for the medical domain. This also includes an improved methodology for creating the ground truth in classes where there is a substantial inter-observer variation in the assessment, which might be used by other researchers to increase the number of labels and segmentations for the dataset.

Usage Notes

In our research on detecting, classifying, and segmenting normal and abnormal findings in the GI tract, we have collected, to the best of our knowledge, the largest and most diverse dataset. These data are made available as a resource to the research community enabling researchers not only to have the ability to research the detection or classification of various GI findings but also differentiate between severity of the findings.

In short, we have used the labeled data to research the classification and segmentation of GI findings using both computer vision and ML approaches to potentially be used in live and post-analysis of patient examinations. Areas of potential utilization are analysis, classification, segmentation, and retrieval of images and videos with particular findings or particular properties from the computer science area. The labeled data can also be used for teaching and training in medical education. Having expert gastroenterologists providing the ground truths over various findings, *HyperKvasir* provides a unique and diverse learning set for future clinicians. Moreover, the unlabeled data is well suited for semi-supervised and unsupervised methods, and, if even more ground truth data is needed, the users of the data can use their own local medical experts to provide the needed labels. Finally, the videos can in addition be used to simulate live endoscopies feeding the video into the system like it is captured directly from the endoscopes enable developers to do image classification.

The dataset includes a series of scripts and text files that aim to help researchers quickly get started using the dataset for standard ML tasks such as classification. These are available at the GitHub repository for the dataset: http://www.github.com/simula/hyper-kvasir. Moreover, we provide three official splits of the dataset that can be used for cross-validation experiments. Keeping splits consistent between methods helps maintain a fair comparison of results. The scripts used to generate the plots, split data into different folds, and generate annotation files are included for reproducibility and transparency. These files may also be used to further experiment with the dataset. Finally, we include the files used to create our preliminary experiments.

There is currently a lot of research being performed in the field of GI image and video analysis, and we welcome and encourage future contributions in this area. This is not limited to using the dataset for comparisons and reproducibility of experiments, but also publishing and sharing new data in the future.

Code availability

In addition to releasing the data, we also make available the code used in the experiments. All code and additional data required for the experiments are available on GitHub at http://www.github.com/simula/hyper-kvasir.

Received: 31 December 2019; Accepted: 21 July 2020;
Published online: 28 August 2020

References

1. Brenner, H., Kloor, M. & Pox, C. P. Colorectal cancer. *The Lancet* **383**, 1490–502, https://doi.org/10.1016/S0140-6736(13)61649-9 (2014).
2. Torre, L. A. *et al*. Global cancer statistics, 2012. *CA: A Cancer J. for Clin.* **65**, 87–108, https://doi.org/10.1056/NEJMoa0907667 (2015).
3. World Health Organization - International Agency for Research on Cancer. Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012 (2012).
4. Hewett, D. G., Kahi, C. J. & Rex, D. K. Efficacyandeffectivenessofcolonoscopy: howdowebridgethegap? *Gastrointest. Endosc. Clin.* **20**, 673–684, https://doi.org/10.1016/j.giec.2010.07.011 (2010).
5. Lee, S. H. *et al*. Endoscopic experience improves interobserver agreement in the grading of esophagitis by los angeles classification: conventional endoscopy and optimal band image system. *Gut liver* **8**, 154, https://doi.org/10.5009/gnl.2014.8.2.154 (2014).
6. Van Doorn, S. C. *et al*. Polyp morphology: an interobserver evaluation for the paris classification among international experts. *The Am. J. Gastroenterol.* **110**, 180–187, https://doi.org/10.1038/ajg.2014.326 (2015).
7. Kaminski, M. F. *et al*. Quality indicators for colonoscopy and the risk of interval cancer. *New Engl. J. Medicine* **362**, 1795–1803, https://doi.org/10.1056/NEJMoa0907667 (2010).
8. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Medicine* **25**, 44–56, https://doi.org/10.1038/s41591-018-0300-7 (2019).
9. Riegler, M. *et al*. Multimedia and medicine: Teammates for better disease detection and survival. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 968–977, https://doi.org/10.1145/2964284.2976760 (2016).
10. Riegler, M. *et al*. EIR - efficient computer aided diagnosis framework for gastrointestinal endoscopies. In *Proceedings of the IEEE International Workshop on Content-Based Multimedia Indexing (CBMI)*, 1–6, https://doi.org/10.1109/CBMI.2016.7500257 (2016).
11. Alammari, A. *et al*. Classification of ulcerative colitis severity in colonoscopy videos using cnn. In *Proceedings of the ACM International Conference on Information Management and Engineering (ACM ICIME)*, 139–144, https://doi.org/10.1145/3149572.3149613 (2017).
12. Wang, Y., Tavanapong, W., Wong, J., Oh, J. H. & De Groen, P. C. Polyp-alert: Nearreal-timefeedbackduringcolonoscopy. *Comput. Methods Programs Biomed.* **120**, 164–179, https://doi.org/10.1016/j.cmpb.2015.04.002 (2015).
13. Hirasawa, T., Aoyama, K., Fujisaki, J. & Tada, T. 113 application of artificial intelligence using convolutional neural network for detecting gastric cancer in endoscopic images. *Gastrointest. Endosc.* **87**, AB51, https://doi.org/10.1016/j.gie.2018.04.025 (2018).

14. Wang, L., Xie, C. & Hu, Y. Iddf2018-abs-0260 deep learning for polyp segmentation. *Gut* **67**, A84–A85, https://doi.org/10.1136/gutjnl-2018-IDDFabstracts.181 (2018).
15. Mori, Y. *et al*. Real-Time Use of Artificial Intelligence in Identification of Diminutive Polyps During Colonoscopy: A Prospective Study. *Annals Intern. Medicine* **169**, 357–366, https://doi.org/10.7326/M18-0249 (2018).
16. Bychkov, D. *et al*. Deep learning based tissue analysis predicts outcome in colorectal cancer. *Sci. Reports* **8**, 3395, https://doi.org/10.1038/s41598-018-21758-3 (2018).
17. Min, M. *et al*. Computer-aided diagnosis of colorectal polyps using linked color imaging colonoscopy to predict histology. *Sci. reports* **9**, 2881, https://doi.org/10.1038/s41598-019-39416-7 (2019).
18. Bernal, J. & Aymeric, H. Miccai endoscopic vision challenge polyp detection and segmentation. https://endovissub2017-giana.grand-challenge.org/home/, Accessed: 2017-12-11 (2017).
19. Bernal, J. *et al*. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput. Med. Imaging Graph.* **43**, 99–111, https://doi.org/10.1016/j.compmedimag.2015.02.007 (2015).
20. Tajbakhsh, N., Gurudu, S. R. & Liang, J. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Transactions on Med. Imaging* **35**, 630–644, https://doi.org/10.1109/TMI.2015.2487997 (2016).
21. Deng, J. *et al*. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255, https://doi.org/10.1109/CVPR.2009.5206848 (2009).
22. Pogorelov, K. *et al*. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the ACM Multimedia Systems Conference (ACM MMSYS)*, 164–169, https://doi.org/10.1145/3083187.3083212 (2017).
23. Pogorelov, K. *et al*. Deep learning and hand-crafted feature based approaches for polyp detection in medical videos. In *Proceedings of the IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, 381–386, https://doi.org/10.1109/CBMS.2018.00073 (2018).
24. Berstad, T. J. D. *et al*. Tradeoffsusingbinaryandmulticlassneuralnetworkclassificationformedicalmultidiseasedetection. In *Proceedings of the IEEE International Symposium on Multimedia (ISM)*, 1–8, https://doi.org/10.1109/ISM.2018.00009 (2018).
25. de Lange, T., Halvorsen, P. & Riegler, M. Methodology to develop machine learning algorithms to improve performance in gastrointestinal endoscopy. *World J. Gastroenterol.* **24**, 5057–5062, https://doi.org/10.3748/wjg.v24.i45.5057 (2018).
26. Hicks, S. *et al*. 383 deep learning for automatic generation of endoscopy reports. *Gastrointest. Endosc.* **89**, AB77, https://doi.org/10.1016/j.gie.2019.04.053 (2019).
27. Ahmad, J., Muhammad, K., Lee, M. Y. & Baik, S. W. Endoscopic image classification and retrieval using clustered convolutional features. *J. Med. Syst.* **41**, 196, https://doi.org/10.1007/s10916-017-0836-y (2017).
28. Owais, M., Arsalan, M., Choi, J., Mahmood, T. & Park, K. R. Artificial intelligence-based classification of multiple gastrointestinal diseases using endoscopy videos for clinical diagnosis. *J. Clin. Medicine* **8**, 986, https://doi.org/10.3390/jcm8070986 (2019).
29. Ahmad, J., Muhammad, K. & Baik, S. W. Medical image retrieval with compact binary codes generated in frequency domain using highly reactive convolutional features. *J. Med. Syst.* **42**, 24, https://doi.org/10.1007/s10916-017-0875-4 (2017).
30. Harzig, P., Einfalt, M. & Lienhart, R. Automatic disease detection and report generation for gastrointestinal tract examination. *Proceedings of the ACM International Conference on Multimedia (ACM MM)* **5**, 2573–2577, https://doi.org/10.1145/3343031.3356066 (2019).
31. Kasban, H. & Salama, D. H. A robust medical image retrieval system based on wavelet optimization and adaptive block truncation coding. *Multimed. Tools Appl.* **78**, 35211–35236, https://doi.org/10.1007/s11042-019-08100-3 (2019).
32. Ghatwary, N., Zolgharni, M. & Ye, X. Gfd faster r-cnn: Gabor fractal densenet faster r-cnn for automatic detection of esophageal abnormalities in endoscopic images. *International Workshop on Machine Learning in Medical Imaging (MLMI)* **11861**, 89–97, https://doi.org/10.1007/978-3-030-32692-0_11 (2019).
33. Ghatwary, N. M., Ye, X. & Zolgharni, M. Esophageal abnormality detection using densenet based faster r-cnn with gabor features. *IEEE Access* **7**, 84374–84385, https://doi.org/10.1109/ACCESS.2019.2925585 (2019).
34. Hicks, S. A. *et al*. Mimir: an automatic reporting and reasoning system for deep learning based analysis in the medical domain. In *Proceedings of the ACM Multimedia Systems Conference (ACM MMSYS)*, 369–374, https://doi.org/10.1145/3204949.3208129 (2018).
35. Hicks, S. *et al*. Dissecting deep neural networks for better medical image classification and classification understanding. In *Proceedings of the IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, 363–368, https://doi.org/10.1109/CBMS.2018.00070 (2018).
36. Hicks, S. A. *et al*. Comprehensible reasoning and automated reporting of medical examinations based on deep learning analysis. In *Proceedings of the ACM Multimedia Systems Conference (ACM MMSYS)*, 490–493, https://doi.org/10.1145/3204949.3208113 (2018).
37. Pogorelov, K. *et al*. Opensea: open search based classification tool. In *Proceedings of the ACM Multimedia Systems Conference (ACM MMSYS)*, 363–368, https://doi.org/10.1145/3204949.3208128 (2018).
38. Thambawita, V. L. *et al*. Anextensivestudyoncross-datasetbiasandevaluationmetricsinterpretationformachinelearning applied to gastrointestinal tract abnormality classification. *ACM Transactions on Comput. for Healthc.* (2020).
39. Riegler, M. *et al*. Multimedia for medicine: the medico task at mediaeval 2017. In *Proceeding of the MediaEval Benchmarking Initiative for Multimedia Evaluation Workshop (MediaEval)* (2017).
40. Pogorelov, K. *et al*. Medico multimedia task at mediaeval 2018. In *Proceeding of the MediaEval Benchmarking Initiative for Multimedia Evaluation Workshop (MediaEval)* (2018).
41. Hicks, S. *et al*. Acm multimedia biomedia 2019 grand challenge overview. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2563–2567, https://doi.org/10.1145/3343031.3356058 (2019).
42. Cheplygina, V., de Bruijne, M. & Pluim, J. P. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med. Image Analysis* **54**, 280–296, https://doi.org/10.1016/j.media.2019.03.009 (2019).
43. Hénaff, O. J., Razavi, A., Doersch, C., Eslami, S. & Oord, A. V. D. Data-efficient image recognition with contrastive predictive coding. arXiv preprint arXiv:*1905.09272* (2019).
44. Misra, I. & van der Maaten, L. Self-supervised learning of pretext-invariant representations. *arXiv preprint arXiv:1912.01991* (2019).
45. Bui, T. D., Ravi, S. & Ramavajjala, V. Neuralgraphlearning: Trainingneuralnetworksusinggraphs. In *Proceedings of the ACM International Conferenceon Web Searchand Data Mining (WSDM)*, 64–71,https://doi.org/10.1145/3159652.3159731 (2018).
46. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722* (2019).
47. Amano, Y. *et al*. Interobserver agreement on classifying endoscopic diagnoses of nonerosive esophagitis. *Endoscopy* **38**, 1032–1035, https://doi.org/10.1055/s-2006-944778 (2006).
48. De Lange, T., Larsen, S. & Aabakken, L. Inter-observer agreement in the assessment of endoscopic findings in ulcerative colitis. *BMC gastroenterology* **4**, 9, https://doi.org/10.1186/1471-230X-4-9 (2004).
49. Jha, D. *et al*. Kvasir-seg: A segmented polyp dataset. In *Proceeding of International Conference on Multimedia Modeling (MMM)*, vol. 11962, 451–462, https://doi.org/10.1007/978-3-030-37734-2_37 (2020).
50. Jha, D. *et al*. Resunet++: An advanced architecture for medical image segmentation. In *Proceedings of International Symposium on Multimedia (ISM)*, 225–230, https://doi.org/10.1109/ISM46123.2019.00049 (2019).
51. Borgli, H. *et al*. The HyperKvasir Dataset. *Open Science Framework*, https://doi.org/10.17605/OSF.IO/MH9SJ (2020).
52. Calderwood, A. H. & Jacobson, B. C. Comprehensive validation of the boston bowel preparation scale. *Gastrointest. endoscopy* **72**, 686–692, https://doi.org/10.1016/j.gie.2010.06.068 (2010).
53. Aabakken, L. *et al*. Standardized endoscopic reporting. *J. Gastroenterol. Hepatol.* **29**, 234–240, https://doi.org/10.1111/jgh.12489 (2014).

54. Lundell, L. R. *et al.* Endoscopic assessment of oesophagitis: clinical and functional correlates and further validation of the los angeles classification. *Gut* **45**, 172–180, https://doi.org/10.1136/gut.45.2.172 (1999).
55. Sharma, P. *et al.* The development and validation of an endoscopic grading system for barrett's esophagus: The prague c & m criteria. *Gastroenterology* **131**, 1392–1399, https://doi.org/10.1053/j.gastro.2006.08.032 (2006).
56. Lai, E. J., Calderwood, A. H., Doros, G., Fix, O. K. & Jacobson, B. C. The boston bowel preparation scale: a valid and reliable instrument for colonoscopy-oriented research. *Gastrointest. Endosc.* **69**, 620–625, https://doi.org/10.1016/j.gie.2008.05.057 (2009).
57. Schroeder, K. W., Tremaine, W. J. & Ilstrup, D. M. Coated oral 5-aminosalicylic acid therapy for mildly to moderately active ulcerative colitis. *The New Engl. J. Medicine* **317**, 1625–1629, https://doi.org/10.1056/NEJM198712243172603 (1987).
58. Lambert, R. Theparisendoscopicclassificationofsuperficialneoplasticlesions: esophagus,stomach,andcolon: November 30 to december 1, 2002. *Gastrointest Endosc* **58**, S3–S43, https://doi.org/10.1016/S0016-5107(03)02159-X (2003).
59. Forrest, J. H., Finlayson, N. & Shearman, D. Endoscopy in gastrointestinal bleeding. *The Lancet* **304**, 394–397, https://doi.org/10.1016/s0140-6736(74)91770-x (1974).
60. Bray, F. *et al.* Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* **68**, 394–424, https://doi.org/10.3322/caac.21492 (2018).
61. Thambawita, V. *et al.* The medico-task 2018: Disease detection in the gastrointestinal tract using global features and deep learning. In *Proceeding of the MediaEval Benchmarking Initiative for Multimedia Evaluation Workshop (MediaEval)* (2018).
62. Bernal, J., Sánchez, J. & Vilarino, F. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognit.* **45**, 3166–3182, https://doi.org/10.1016/j.patcog.2012.03.002 (2012).
63. Ali, S. *et al.* Endoscopy artifact detection (ead 2019) challenge dataset. *arXiv preprint arXiv:1905.03209* (2019).
64. Silva, J., Histace, A., Romain, O., Dray, X. & Granado, B. Toward embedded detection of polyps in wce images for early diagnosisofcolorectalcancer. *Int. J. Comput. Assist. Radiol. Surg.* **9**, 283–293, https://doi.org/10.1007/s11548-013-0926-3 (2014).
65. Koulaouzidis, A. *et al.* Kid project: an internet-based digital video atlas of capsule endoscopy for research purposes. *Endosc. international open* **5**, E477–E483, https://doi.org/10.1055/s-0043-105488 (2017).
66. Bernal, J. & Aymeric, H. Gastrointestinal Image ANAlysis (GIANA) Angiodysplasia D&L challenge. https://endovissub2017-giana.grand-challenge.org/home/, Accessed: 2017-11-20 (2017).
67. Angermann, Q. *et al.* Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis. *Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures (CARE CLIP)* **10550**, 29–41, https://doi.org/10.1007/978-3-319-67543-5_3 (2017).
68. Bernal, J. *et al.* Polyp detection benchmark in colonoscopy videos using gtcreator: A novel fully configurable tool for easy and fast annotation of image databases. In Proceedings of Computer Assisted Radiology and Surgery (CARS), https://hal.archives-ouvertes.fr/hal-01846141 (2018).
69. Gastrolab - the gastrointestinal site, http://www.gastrolab.net/index.htm. Accessed: 2019-12-12.
70. Weo clinical endoscopy atlas, http://www.endoatlas.org/index.php. Accessed: 2019-12-12.
71. Gastrointestinal lesions in regular colonoscopy dataset, http://www.depeca.uah.es/colonoscopy_dataset/, Accessed: 2019-12-12.
72. The atlas of gastrointestinal endoscope, http://www.endoatlas.com/atlas_1.html. Accessed: 2019-12-12.
73. El salvador atlas of gastrointestinal video endoscopy, http://www.gastrointestinalatlas.com/index.html. Accessed: 2019-1216.
74. Pogorelov, K. *et al.* Nerthus: A bowel preparation quality video dataset. In *Proceedings of the ACM Multimedia Systems Conference (ACM MMSYS)*, 170–174, https://doi.org/10.1145/3083187.3083216 (2017).

Acknowledgements

We would like to acknowledge various people at Bærum Hospital for making the data available. Moreover, the work is partially funded in part by the Research Council of Norway, project numbers 263248 (Privaton) and 282315 (AutoCap).

Author contributions

S.A.H., V.T., P.H., H.L.H., M.A.R. and T.d.L. conceived the experiment(s), S.A.H., V.T., H.L.H. and M.A.R. conducted the experiment(s), H.B., S.A.H., M.A.R., P.H. and T.d.L. prepared and cleaned the data for publication, and all authors analyzed the results and reviewed the manuscript.

Competing interests

Authors P.H.S., D.J., C.G., M.A.R., P.H. and T.d.L. all own shares in the Augere Medical AS company developing AI solutions for colonoscopies. The Augere video annotation system was used to label the videos. There is no commercial interest from Augere regarding this publication and dataset. Otherwise, the authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to P.H.

A.13 Paper XIII: Kvasir-Capsule, a video capsule endoscopy dataset

Authors: P. H Smedsrud, H. L Gjestang, O. O Nedrejord, E. Næss, V. Thambawita, S. Hicks, H. Borgli, D. Jha, T. J. Derek Berstad, S. L Eskeland, M. Lux, H. Espeland, A. Petlund, D. T. Dang Nguyen, D. Johansen, P. T. Schmidt, H. L. Hammer, T. d. Lange, M. Riegler, and P. Halvorsen

Abstract: Artificial intelligence (AI) is predicted to have profound effects on the future of video capsule endoscopy (VCE) technology. The potential lies in improving anomaly detection while reducing manual labour. However, medical data is often sparse and unavailable to the research community, and qualified medical personnel rarely have time for the tedious labelling work. In this respect, we present Kvasir-Capsule, a large VCE dataset collected from examinations at Hospitals in Norway. Kvasir-Capsule consists of 118 videos which can be used to extract a total of 4,820,739 image frames. We have labelled and medically verified 44, 228 frames with a bounding box around detected anomalies from 13 different classes of findings. In addition to these labelled images, there are 4, 776,479 unlabelled frames included in the dataset. Initial work demonstrates the potential benefits of AI-based computer-assisted diagnosis systems for VCE. However, they also show that there is great potential for improvements, and the Kvasir-Capsule dataset can play a valuable role in developing better algorithms in order for VCE technology to reach its true potential.

Published: Scientific Data

Candidate contributions: D. Jha contributed to the drafting of the manuscript. Additionally, he contributed to the revision of the manuscript.

Thesis objectives: Objective I

scientific data

OPEN

DATA DESCRIPTOR

Kvasir-Capsule, a video capsule endoscopy dataset

Pia H. Smedsrud[1,3,6,15]✉, Vajira Thambawita[1,2,15], Steven A. Hicks[1,2,15], Henrik Gjestang[1,3], Oda Olsen Nedrejord[1,3], Espen Næss[1,3], Hanna Borgli[1,3], Debesh Jha[1,7,15], Tor Jan Derek Berstad[6], Sigrun L. Eskeland[4], Mathias Lux[10], Håvard Espeland[6], Andreas Petlund[6], Duc Tien Dang Nguyen[5], Enrique Garcia-Ceja[13], Dag Johansen[7], Peter T. Schmidt[8,9], Ervin Toth[14], Hugo L. Hammer[1,2], Thomas de Lange[4,6,11,12,15,16], Michael A. Riegler[1,15,16] & Pål Halvorsen[1,2,15,16]

Artificial intelligence (AI) is predicted to have profound effects on the future of video capsule endoscopy (VCE) technology. The potential lies in improving anomaly detection while reducing manual labour. Existing work demonstrates the promising benefits of AI-based computer-assisted diagnosis systems for VCE. They also show great potential for improvements to achieve even better results. Also, medical data is often sparse and unavailable to the research community, and qualified medical personnel rarely have time for the tedious labelling work. We present *Kvasir-Capsule*, a large VCE dataset collected from examinations at a Norwegian Hospital. *Kvasir-Capsule* consists of 117 videos which can be used to extract a total of 4,741,504 image frames. We have labelled and medically verified 47,238 frames with a bounding box around findings from 14 different classes. In addition to these labelled images, there are 4,694,266 unlabelled frames included in the dataset. The *Kvasir-Capsule* dataset can play a valuable role in developing better algorithms in order to reach true potential of VCE technology.

Background & Summary

The small bowel constitutes the gastrointestinal (GI) tract's mid-part, situated between the stomach and the large bowel. It is three to four meters long and has a surface of about 30 m^2, including the villi's surface. As part of the digestive system, it plays a crucial role in absorbing nutrients[1]. Therefore, disorders in the small bowel may cause severe growth retardation in children and nutrient deficiencies in children and adults[1]. This organ may be affected by chronic diseases, like Crohn's disease, coeliac disease, and angiectasias, or malignant diseases like lymphoma and adenocarcinoma[2,3]. These diseases may represent a substantial health challenge for both the patients and the society, and a thorough examination of the lumen is frequently necessary to diagnose and treat them[4]. However, due to its anatomical location, the small bowel is less accessible for inspection by flexible endoscopes commonly used for the upper GI tract and the large bowel. Since early 2000, video capsule endoscopy (VCE)[5] has been used, usually as a complementary test for patients with GI bleeding[4]. A VCE consists of a small capsule containing a wide-angle camera, light sources, batteries, and other electronics. The patient swallows the capsule capturing a video as it moves passively throughout the GI tract. A recorder, carried by the patient or included in the capsule, stores the video before a medical expert examines it after the procedure.

VCE devices exist in various versions and brands such as Given Imaging (Medtronic), Ankon Technologies, Chongqing Science, IntroMedic, CapsoVision, and Olympus. The frame rate typically varies between 1 and 30 frames per second, capturing in total between 50 and 100 thousand frames, with pixel-resolutions in the range of

[1]SimulaMet, Oslo, Norway. [2]Oslo Metropolitan University, Oslo, Norway. [3]University of Oslo, Oslo, Norway. [4]Department of Medical Research, Bærum Hospital, Gjettum, Norway. [5]University of Bergen, Bergen, Norway. [6]Augere Medical AS, Oslo, Norway. [7]UIT The Arctic University of Norway, Tromsø, Norway. [8]Karolinska Institutet, Department of Medicine, Solna, Sweden. [9]Ersta Hospital, Department of Medicine, Stockholm, Sweden. [10]Klagenfurt University, Wörthersee, Austria. [11]Medical Department, Sahlgrenska University Hospital-Mölndal Hospital, Göteborg, Sweden. [12]Department of Molecular and Clinical Medicine, Sahlgrenska Academy, University of Gothenburg, Göteborg, Sweden. [13]SINTEF Digital, Oslo, Norway. [14]Department of Gastroenterology, Skåne University Hospital, Malmö Lund University, Malmö, Sweden. [15]These authors contributed equally: Pia H. Smedsrud, Vajira Thambawita, Steven A. Hicks, Debesh Jha, Thomas de Lange, Michael A. Riegler, Pål Halvorsen. [16]These authors jointly supervised: Thomas de Lange, Michael A. Riegler, Pål Halvorsen. ✉e-mail: pia@simula.no

Dataset	Findings	Size	Availability
KID[54]	Angiectasia, bleeding, inflammations, polyps	2,371 images + 47 videos	open academic•
GIANA 2017[55]	Angiectasia†	600 images	by request
GIANA2018[56,57]	Polyps and small bowel lesions†	8262 images + 38 videos	by request
CAD-CAP[58,59]	Normal frames, fresh blood, vascular lesion, ulcerative and inflammatory lesions	25,000 images	by request◇
Gastrolab[60]	Crohns diseases, small bowel (video)+ GI lesions	Few hundred images and videos	open academic•

Table 1. An overview of existing VCE datasets from the GI tract. †Including ground truth segmentation masks. •Not available anymore. ◇The Computer-Assisted Diagnosis for CAPsule endoscopy (CAD-CAP) Database - used for the angiectasia detection.

256 × 256 to 512 × 512. Some of the vendors have software to remove duplicated frames due to slow movement. However, a large number of frames need to be analysed by a medical expert, resulting in a tedious and error-prone operation. In the related area of colonoscopy, operator variation and detection performance are reported problems[6–8] resulting in high miss rates[9]. In VCE analysis, essential findings are missed due to lack of concentration, insufficient experience and knowledge[10–12]. Furthermore, physicians may have trouble handling the associated technology, and infrequent VCE use leads to lack of confidence[13], resulting in inter-observer and intra-observer variations in the assessments[12].

The technical developments for automated image and video analysis have sky-rocketed, and multimedia solutions in medicine show tremendous potential[14,15]. An increasing number of promising machine learning solutions are being developed for automated diagnosis of colonoscopies[16–23] using open datasets[24–27]. Regarding automated VCE data analyses, machine learning approaches also produce promising results regarding detection and classification rates[28–35]. Machine learning, or artificial intelligence (AI) in general, is likely to have profound effects on the VCE technology's future, not only for improving variation and detection rates but also for estimating the capsule's localisation[13,36].

Regardless of promising initial results, there is room for improvements in detection rate, reduced manual labour, and AI explainability. Large amounts of data are needed[37,38], particularly annotated data[35], and access to these data are often scarce[39]. As shown in Table 1, very few, small VCE datasets are made publicly available, and several have become unavailable. We have previously published the HyperKvasir dataset[27]. Nevertheless, this and similar datasets containing images from *colonoscopies* and *esophagogastroscopies* are not applicable because they do not depict the small bowel, characterised by the intestinal villi displaying a different surface than the rest of the bowel. Also, the image resolution and the frame rate of VCEs are much lower. The small bowel is not air inflated during a VCE examination, as is the case with traditional colonoscopies. Different optics are also used, and the movement of the capsule is uncontrolled in contrast to flexible endoscopes used during manual examinations.

Therefore, we present a large VCE dataset, called *Kvasir-Capsule*, consisting of 117 videos with 4,741,504 frames and 14 classes of findings. The dataset contains labelled images and their corresponding full videos, and also unlabelled videos. Recent work in the machine learning community has shown significant improvements regarding sparsely labelled and unlabelled data value. Semi-supervised learning algorithms are successfully applied in different medical image analyses[40,41] using self-learning[42,43] and neural graph learning[44]. Finally, we provide a baseline analysis and outline possible future research directions using *Kvasir-Capsule*.

Methods

The VCE videos were collected from consecutive clinical examinations performed at the Department of Medicine, Bærum Hospital, Vestre Viken Hospital Trust in Norway, which provides health care services to 490,000 people, of which about 200,000 are covered by Bærum Hospital. The examinations were conducted between February 2016 and January 2018 using the Olympus Endocapsule 10 System[45] including the Olympus EC-S10 endocapsule (Fig. 1a) and the Olympus RE-10 endocapsule recorder (Fig. 1b). Originally, the videos were captured at a rate of 2 frames per second, in a resolution of 336 × 336, and encoded using H.264 (MPEG-4 AVC, part 10). The videos were exported in AVI format using the Olympus system's export tool packaged and encapsulated in the same H.264 format, i.e., the frame formats are the same, but the frame rate specification is changed to 30 fps by the export tool.

Initially, a trained clinician analysed all videos using the Olympus software, selecting thumbnails from lesions and normal findings as part of their clinical work. In spring 2019, all the 117 anonymous videos and thumbnails were exported from a stand-alone workstation using the Olympus software. The Olympus video capsule system has user-friendly functionalities like Omni-selected Mode, skipping images that overlap with previous ones.

All metadata were removed and files renamed with randomly generated file names, before exporting the videos and thumbnails that were shared. Thus, data in the dataset are fully anonymized, as approved by Privacy Data Protection Authority and in accordance with relevant guidelines and regulations of the Regional Committee for Medical and Health Research Ethics - South East Norway. The data has not been pre-processed or augmented in any way apart from this. Subsequently, for clinical analyses of the videos, a central expert reader selected and categorized thumbnails with pathological findings. These thumbnails were traced to their corresponding video segments and the videos were uploaded to a video annotation platform (provided by Augere Medical AS, Norway) for efficient viewing and labelling. Next, three master students labelled and marked the findings with bounding boxes for each frame. The bounding boxes were designed to include the entire lesion and as little as possible of the surrounding mucosa. If the students were unsure about the labelling, the expert reader verified the frames. All labels regarding anatomical structures and normal clean mucosa were then confirmed by one junior

(a) Olympus EC-S10 endocapsule **(b)** Olympus RE-10 endocapsule recorder

Fig. 1 VCE equipment used for data collection.

Data Record	# Files
Labelled images	47,238
Labelled videos	43
Unlabelled images	4,694,266
Unlabelled videos	74

Table 2. Overview of the data records in the *Kvasir-Capsule* dataset.

medical doctor and the expert reader. Finally, all the annotations were once more verified by the expert reader and subsequently validated by a second expert reader. If the second reviewer disagreed with the annotations, the first reviewer reassessed the images to see whether he then agreed with the second reviewer to get an agreement. After the validation process by the second reviewer there was a disagreement on twenty-six findings in seven examinations; nineteen concerning erroneous terminology of the class lymphoid hyperplasia which was changed to lymphangiectasia. The other seven were related to the interpretation of the finding. After reviewing these findings, the first reviewer agreed with the second one to finally reach a perfect agreement. After this procedure, the video frames were exported as images. Hence, a total of four medical persons have selected, analysed and verified the data, and a total of 47,238 frames are labelled.

The Norwegian Privacy Data Protection Authority approved the export of anonymous images for the creation of the database, without consent from participants. It was exempted from approval from the Regional Committee for Medical and Health Research Ethics - South East Norway. Since the data is anonymised and all metadata removed, the dataset is publicly shareable based on Norwegian and General Data Protection Regulation (GDPR) laws.

Data Records

The *Kvasir-Capsule* dataset is available from the Open Science Framework (OSF)[46]. Table 2 gives an overview of all data records in the dataset. In total, the dataset consists of 4,741,621 main data records, i.e., 47,238 images with labels and bounding box masks, the 43 corresponding labelled videos (the videos from which the images are extracted), and 74 unlabelled videos (from which labelled images have not been extracted). 4,694,266 unlabelled images can further be extracted from all the videos combined. All the various labelled classes are shown in Fig. 2. The dataset has a total size of circa 89 GB. Note that the unlabelled images are not extracted and included in the uploaded data due to unnecessary duplication of data, but can easily be extracted from the videos.

The dataset is stored according to the data records listed above, and described in more detail below. We have a "labelled images" catalogue which contains archive files of each labelled class of images. We have a "labelled videos" catalogue which contains all the videos where we have annotated findings from, and an "unlabelled videos" catalogue containing the videos that are not annotated.

Labelled images. In total, the dataset contains 47,238 labelled images stored using the PNG format, where Fig. 3 shows the 14 different classes representing the labelled images and the number of images in each class. The provided *metadata.csv* comma-separated value (CSV) file gives the mapping between file name, the labelling for the image, the corresponding video, and the video frame number. Moreover, the CSV file gives information about the bounding box outlining the finding. Some samples are given in Fig. 4 where the first line gives the type of each element in the lines below. This means that the file *filename* of the labelled image which is the frame *frame_number* extracted from the *video_id* video. Moreover, the finding is from the category *finding_category* and class *finding_class*. Finally, the four x_i, y_i pairs are the four pixel coordinates for the bounding box, e.g., in the first three lines they are empty, meaning that there is no finding with a bounding box in this labelled image. There is one line in the file per each labelled image.

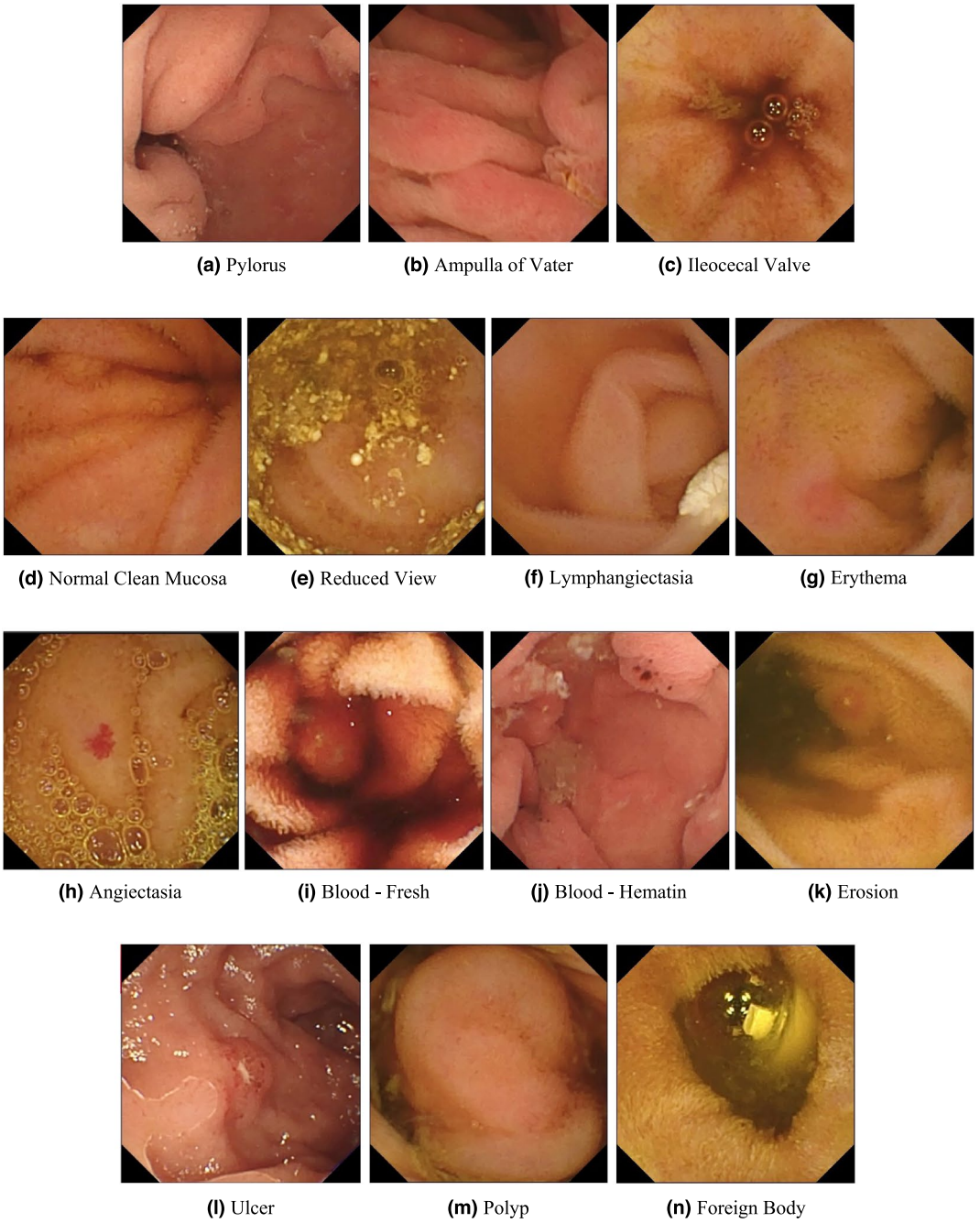


Fig. 2 Image examples of the various labelled classes for images. Images (**a**) to (**c**) are from the *Anatomy* category, and images (**d**) to (**n**) are from the *Luminal findings category*.

We defined two main categories of findings, namely anatomy and luminal findings. Each category, their classes and belonging images are stored in their corresponding folder. As observed in Fig. 3, the number of images per class is not balanced. This is a global challenge in the medical field because some findings are more common than others, which adds a challenge for researchers since methods applied to the data should also be able to learn from a small amount of training data.

Categories of findings. We have organised the dataset in two main categories with their corresponding classes according to the World Endoscopy Association Minimal Standard Terminology version 3.0 (MST 3.0), though we have not included the subcategories or intermediate level to simplify the dataset[47].

Anatomy. The category of *Anatomy* contains anatomical landmarks characterising the GI tract. These landmarks may be used for orientation during endoscopic procedures. However, for small bowel VCE their role is to verify the passage of the capsule trough the entire small bowel to confirm a complete examination. We have labelled three anatomical landmarks, the first two delineate the upper (proximal) and lower (distal) end of the

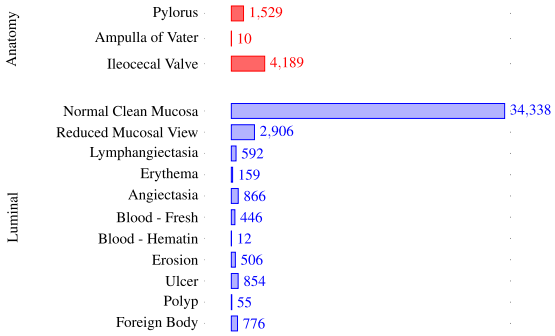


Fig. 3 The number of images in the various Kvasir-Capsule labelled image classes.

```
filename;video_id;frame_number;finding_category;finding_class;x1;y1;x2;y2;x3;y3;x4;y4
...
0728084c8da942d9_22805.jpg;0728084c8da942d9;22805;Luminal;Normal clean mucosa;;;;;;;
0728084c8da942d9_22806.jpg;0728084c8da942d9;22806;Luminal;Normal clean mucosa;;;;;;;
0728084c8da942d9_22807.jpg;0728084c8da942d9;22807;Luminal;Normal clean mucosa;;;;;;;
...
0728084c8da942d9_28789.jpg;0728084c8da942d9;28789;Luminal;Erosion;195;226;244;226;244;265;195;265
0728084c8da942d9_28798.jpg;0728084c8da942d9;28798;Luminal;Erosion;183;212;213;212;213;265;183;265
0728084c8da942d9_28799.jpg;0728084c8da942d9;28799;Luminal;Erosion;197;213;229;213;229;267;197;267
...
```

Fig. 4 Samples from the *metadata.csv* CSV file.

small bowel, respectively. The **pylorus** is the anatomical junction between the stomach and small bowel and is a sphincter (circular muscle) regulating the emptying of the stomach into the duodenum. The **ileocecal valve** marks the transition from the small bowel to the large bowel and is a valve preventing reflux of colonic contents, stool, back into the small bowel. The third one, the **ampulla of Vater**, is the junction between the duodenum and the gall duct.

Luminal findings. Endoscopic examinations may detect various *luminal findings*, this include the subcategories content of the bowel lumen, the aspect of the mucosa and mucosal lesions (pathological findings) that could be either flat, elevated or excavated. These subcategories are not shown in the dataset. Normally, the small bowel contains only a certain amount of yellow or brown liquid considered as clean mucosa. However, larger amounts of content may preclude a complete visualisation of the mucosa crucial to verify normal mucosa and detection of all pathological(abnormal) findings. For the lumen content assessment, we have labelled five classes. **Normal clean mucosa** depicts clean small bowel with no or small amount of fluid and mucosa with healthy villi and no pathological findings. This class can also double as a "normal" class versus the pathological luminal finding class (see below). The class **reduced mucosal view** shows small bowel content reducing the view of the mucosa, like stool or bubbles. However, lesions in the upper GI tract or small bowel may bleed, causing the appearance of **blood - fresh** colouring the liquid red. In cases with minimal bleeding, one may observe small black stripes called **blood - hematin** on the mucosal surface. The **foreign body** class include tablet residue or retained capsules which can also be observed in the lumen.

Abnormalities, called lesions or pathological findings, in the small bowel can be seen as changes to the mucosal surface. Typical mucosal changes sometimes cover larger segments, such as a reddish appearance called erythematous mucosa, is labelled as **erythema**. The mucosal wall can also have different focal lesions. The classes of lesions represented in the *Kvasir-Capsule* dataset are **angiectasia**s; small superficial dilated vessels causing chronic bleeding and subsequently anaemia. It mostly occurs in people with chronic heart and lung diseases[48]. Excavated lesions erode to different extents the surface of the mucosa. Most common are **erosion**s, covered by a tiny fibrin layer, while larger erosions are called **ulcer**s. As an example, Crohn's disease is a chronic inflammation of the small bowel characterised by ulcers and erosions of the mucosa. It may cause strictures of the lumen, making the absorption and passage of nutrients difficult[49]. **Lymphangiectasia**, which represents dilated lymphoid vessels in the mucosal wall, and **polyp**s, which may be precancerous lesions, are visible as protruding from the mucosal wall.

Labelled videos. Labelled videos are the full 43 videos from which we extracted the above mentioned labelled image classes. In total, these videos correspond to approximately 19 hours of video and 47,238 labelled video frames. Several segments of each video was labelled, and these segments are what was exported as the labelled images. As previously mentioned, one can find the frame number and video of origin of each extracted image in the CSV-file. Even though we already have extracted the most interesting frames (images) found by the clinicians from these videos, they do contain 1,932,047 non-labelled frames that could be interesting in future research. One could also extract the video sequences around the various findings.

Method		Macro average			Micro average			
		Precision	Recall	F1-score	Precision	Recall	F1-score	MCC
Normal CEL	DensNet-161 (fold 0)	0.2165	0.2341	0.1923	0.7375	0.7375	0.7375	0.3707
	DensNet-161 (fold 1)	0.3493	0.3158	0.2996	0.7327	0.7327	0.7327	0.4604
	Avereage	0.2829	0.2749	0.2459	0.7351	0.7351	0.7351	0.4156
	ResNet-152 (fold 0)	0.3302	0.2401	0.1970	0.7203	0.7203	0.7203	0.3520
	ResNet-152 (fold 1)	0.3431	0.2805	0.2789	0.7481	0.7481	0.7481	0.4718
	Average	**0.3367**	0.2603	0.2379	0.7342	0.7342	0.7342	0.4119
Weighted CEL	DensNet-161 (fold 0)	0.2933	0.2939	0.2523	0.7195	0.7195	0.7195	0.3998
	DensNet-161 (fold 1)	0.3163	0.2914	0.2581	0.6991	0.6991	0.6991	0.4054
	Average	0.3048	**0.2927**	0.2552	0.7093	0.7093	0.7093	0.4026
	ResNet-152 (fold 0)	0.2136	0.2872	0.2186	0.6568	0.6568	0.6568	0.3588
	ResNet-152 (fold 1)	0.3033	0.2799	0.2478	0.6890	0.6890	0.6890	0.3966
	Average	0.2585	0.2836	0.2332	0.6729	0.6729	0.6729	0.3777
Weighted sampling	DensNet-161 (fold 0)	0.2525	0.2794	0.2315	0.7332	0.7332	0.7332	0.4111
	DensNet-161 (fold 1)	0.3463	0.2830	0.2806	0.7400	0.7400	0.7400	0.4547
	Average	0.2994	0.2812	**0.2560**	**0.7366**	**0.7366**	**0.7366**	**0.4329**
	ResNet-152 (fold 0)	0.2637	0.2930	0.2334	0.7324	0.7324	0.7324	0.4088
	ResNet-152 (fold 1)	0.3088	0.2619	0.2417	0.7316	0.7316	0.7316	0.4520
	Average	0.2862	0.2774	0.2375	0.7320	0.7320	0.7320	0.4304

Table 3. Results for all classification experiments. Experiments were done with and without weighted cross-entropy loss (CEL) and using a weighted sampling technique. Bold numbers represent the best average value of that column.

Unlabelled videos. We also provide 74 videos, which contain approximately 25 hours of video and 2,762,219 video frames, without any labels. As previously mentioned, unlabelled data can still have great value. Sparsely labelled or unlabelled data can be important for recently emerging semi-supervised learning algorithms. These videos are of the same format and quality as the labelled videos, except we do not provide any annotations. This means that users of the dataset can either use medical experts to provide further labels, or use the data in unsupervised or semi-supervised learning approaches.

Technical Validation

To evaluate the technical quality of *Kvasir-Capsule*, we performed a series of classification experiments. We trained two CNN-based classifiers to classify the labelled data. Both architectures have previously shown excellent performance on classifying GI-related imagery from traditional colonoscopies[50,51], and should be a good benchmark for VCE-related data. The two algorithms are based on standard CNN architectures, namely DenseNet-161[52] and ResNet-152[53]. All experiments were performed over two-fold cross-validation using categorical cross-entropy loss with and without class weighting. We also used weighted sampling, which balances the dataset by removing and adding images for each class based on a given set of weights. To ensure a fair and robust evaluation, no video is shared between splits. Thus, the frames used for training were independent from the frames used for validation. This also means that there are frames depicting the same finding in each split which then are related to each other, but no related frames distributed across the splits. The effect should therefore be similar to traditional data augmentation techniques used by many researchers today such as multiple rotations, angles and crops.

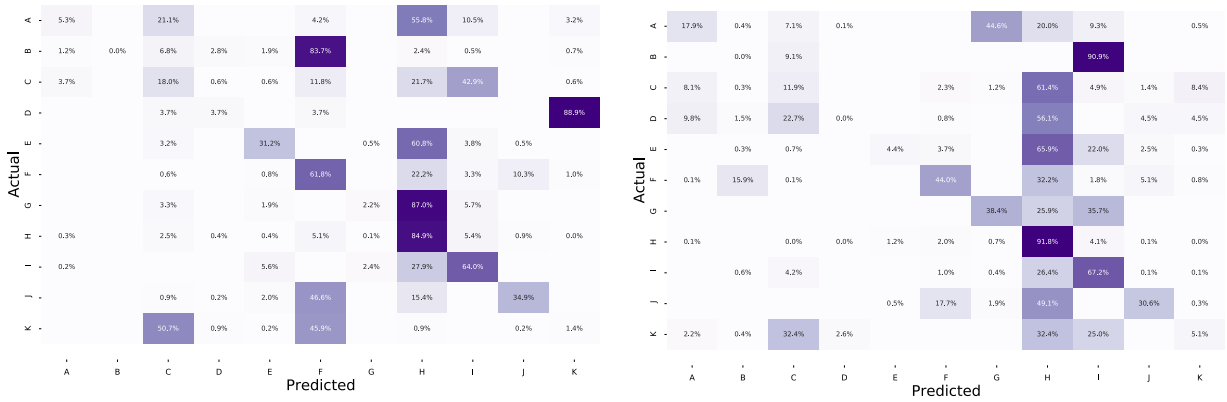
The purpose of these experiments is two-fold. First, we create a baseline for future researchers using the *Kvasir-Capsule* dataset. Second, by using an algorithm that has previously shown good results on classifying GI images, we evaluate how challenging the task of categorizing VCE-related data is. Note that for the classification experiments, we removed the blood - hematin, ampulla of Vater, and polyp classes due to the small number of findings. The results for the two classification algorithms are shown in Table 3 and confusion matrices for the best average MCC value in Fig. 5. We estimated micro-averaged and macro-averaged values for precision, recall and F1-score for each method. The Matthews correlation coefficient (MCC) was calculated using the multi-class generalization, also called the R_K. In short, if TP, TN, FP, and FN are the true positives, true negatives, false positives, and false negatives, respectively, these metrics are defined as follows[26]:

Precision. This metric is also frequently called the *positive predictive value*, and shows the ratio of samples that are correctly identified as positive among the returned samples (the fraction of retrieved samples that are relevant):

$$precision = \frac{TP}{\#\ of\ all\ returned\ samples} = \frac{TP}{TP + FP}$$

Recall. This metric is also frequently called *sensitivity*, *probability of detection* and *true positive rate*, and it is the ratio of samples that are correctly identified as positive among all existing positive samples:

$$recall = \frac{TP}{\#\ of\ all\ positives} = \frac{TP}{TP + FN}$$



(a) Confusion matrix for model evaluated on split 0.

(b) Confusion matrix for model evaluated on split 0.

Fig. 5 Confusion matrices for the best average MCC value which is from the weighted sampling technique. The labeling of the classes is as follows: (A) Angiectasia; (B) Blood - fresh; (C) Erosion; (D) Erythema; (E) Foreign Body; (F) Ileocecal valve; (G) Lymphangiectasia; (H) Normal clean mucosa; (I) Pylorus; (J) Reduced Mucosal View; (K) Ulcer.

F1 score (F1). A measure of a test's accuracy by calculating the harmonic mean of the precision and recall:

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall} = \frac{2TP}{2TP + FP + FN}$$

Matthews correlation coefficient (MCC). MCC takes into account true and false positives and negatives, and is a balanced measure even if the classes are of very different sizes. For the multiclass classification generalization, it is often called the R_k statistic. In following equation, t_k is the number of times class k actually occurred, p_k is the number of times class k was predicted, c is the total number of samples correctly predicted, and s is the total number of samples:

$$MCC = \frac{c \times s - \sum_k^K p_k \times t_k}{\sqrt{(s^2 - \sum_k^K p_k^2) \times (s^2 - \sum_k^K t_k^2)}}$$

The micro and macro averages are different ways to average metrics calculated over multiple classes. The macro average is the arithmetic mean of all the scores of different classes, i.e., calculates the metric per class and then calculates the average of these over the number of classes. For example, it is defined for precision as the sum of precision scores for all classes ($precicion_1 + \ldots + precicion_n$) divided by the number of classes (n). The micro average is not counting class wise first, but looking at the total number of true and false findings. For example, for precision, it is defined as sum of true positives ($TP_1 + \ldots + TP_n$) for all the n classes divided by the all returned positive predictions ($TP_1 + FP_1 + \ldots + TP_n + FP_n$).

Considering the results, we experience that classifying VCE data is quite a challenging task. For example, several of the classes are erroneously predicted as **Normal clean mucosa**. On the other hand, the class with the most accurate predictions is also **Normal clean mucosa**, reaching 85% in fold one and 91% in fold two. This is expected as the class comprise approximately 73% of the labelled images. This points out the challenges of making reliable systems as there are multiple aspects to consider, e.g., the resolution of VCE frames are lower compared to gastro- or colonoscopies, and many of the findings are subtle where even clinicians have difficulties differentiating between the classes. As noticed when comparing the images in Fig. 2, several findings are hard to see and easily mixed. For example, erosions can often be mistaken as small residues, and it can be difficult to differentiate normal mucosa from slight erythema. Thus, these results show the potential of AI-based analysis, but also further motivates the need to publish this dataset for more investigations and research into better specific algorithms for VCE data. The code used to conduct all experiments, produce all plots, and the images contained in each split are available on GitHub (https://github.com/simula/kvasir-capsule), i.e., to increase reproducibility and facilitate researches to perform comparable experiments on the *Kvasir-Capsule* dataset.

Usage Notes

To the best of our knowledge, we have collected the largest and most diverse public available VCE dataset. *Kvasir-Capsule* is made available to enable researchers to develop detection or classification methods of various GI findings using for example computer vision and machine learning approaches. As the labelled findings also include bounding boxes, areas of potential use are analysis, classification, segmentation, and retrieval of images and videos of particular findings or properties. Moreover, the ground truths of various findings by the expert gastroenterologists provide a unique and diverse learning set for future clinicians, i.e., the labelled data can be used for teaching and training in medical education.

The unlabelled data is well suited for semi-supervised and unsupervised machine learning methods, and, if even more ground truth data is needed, the users of the data can have medical experts provide the needed labels. In this respect, recent work has shown remarkable improvements in the area of semi-supervised learning, also successfully applied in medical image analyses[40]. Instead of learning from a large set of annotated data, algorithms learn from sparsely labelled and unlabelled data. Self-learning[42,43] and neural graph learning[44] are both examples using unlabelled data in addition to a small amount of labelled data to extract additional information[41–43]. In an area with scarce data, these new algorithms might be the technology needed to make AI truly useful for medical applications.

An important note in general for this type of AI-based detection systems is that one should be careful about how the dataset is split into for example training and test sets in order to avoid having related frames in several of the sets. This will give an unfair effect on the results. Thus, the splits should be completely different, probably even at the level of patients. As described below, an example of such a split in found in our GitHub repository (see below in the Code Availability section).

Currently, there is substantial research in GI image and video analysis. We welcome future contributions such as using the dataset for comparisons and reproducibility of experiments and further encourage publishing and sharing of new data.

Code availability

In addition to releasing the data, we also publish code used for the baseline experiments. All code and additional data required for the experiments, including our splits into training and test datasets, are available on GitHub via http://www.github.com/simula/kvasir-capsule.

Received: 13 August 2020; Accepted: 15 April 2021;
Published online: 27 May 2021

References

1. Greenwood-Van Meerveld, B., Johnson, A. C. & Grundy, D. Gastrointestinal physiology and function. In *Gastrointestinal Pharmacology*, 1–16 (Springer, 2017).
2. McLaughlin, P. D. & Maher, M. M. Primary malignant diseases of the small intestine. *American Journal of Roentgenology* **201**, W9–W14 (2013).
3. Thomson, A. *et al.* Small bowel review: diseases of the small intestine. *Digestive diseases and sciences* **46**, 2555–2566 (2001).
4. Enns, R. A. *et al.* Clinical practice guidelines for the use of video capsule endoscopy. *Gastroenterology* **152**, 497–514 (2017).
5. Costamagna, G. *et al.* A prospective trial comparing small bowel radiographs and video capsule endoscopy for suspected small bowel disease. *Gastroenterology* **123**, 999–1005 (2002).
6. Hewett, D. G., Kahi, C. J. & Rex, D. K. Efficacy and effectiveness of colonoscopy: how do we bridge the gap? *Gastrointestinal Endoscopy Clinics* **20**, 673–684 (2010).
7. Lee, S. H. *et al.* Endoscopic experience improves interobserver agreement in the grading of esophagitis by los angeles classification: conventional endoscopy and optimal band image system. *Gut and liver* **8**, 154 (2014).
8. Van Doorn, S. C. *et al.* Polyp morphology: an interobserver evaluation for the paris classification among international experts. *The American Journal of Gastroenterology* **110**, 180–187 (2015).
9. Kaminski, M. F. *et al.* Quality indicators for colonoscopy and the risk of interval cancer. *New England Journal of Medicine* **362**, 1795–1803 (2010).
10. Zheng, Y., Hawkins, L., Wolff, J., Goloubeva, O. & Goldberg, E. Detection of lesions during capsule endoscopy: physician performance is disappointing. *American Journal of Gastroenterology* **107**, 554–560 (2012).
11. Chetcuti, S. Z. & Sidhu, R. Capsule endoscopy-recent developments and future directions. *Expert review of gastroenterology & hepatology* **15**, 127–137 (2021).
12. Rondonotti, E. *et al.* Can we improve the detection rate and interobserver agreement in capsule endoscopy? *Digestive and Liver Disease* **44**, 1006–1011 (2012).
13. Cave, D. R., Hakimian, S. & Patel, K. Current controversies concerning capsule endoscopy. *Digestive Diseases and Sciences* **64**, 3040–3047 (2019).
14. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine* **25**, 44 (2019).
15. Riegler, M. *et al.* Multimedia and medicine: Teammates for better disease detection and survival. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 968–977 (2016).
16. Riegler, M. *et al.* EIR - efficient computer aided diagnosis framework for gastrointestinal endoscopies. In *Proceedings of the IEEE International Workshop on Content-Based Multimedia Indexing (CBMI)*, 1–6 (2016).
17. Alammari, A. *et al.* Classification of ulcerative colitis severity in colonoscopy videos using cnn. In *Proceedings of the ACM International Conference on Information Management and Engineering (ICIME)*, 139–144 (2017).
18. Wang, Y., Tavanapong, W., Wong, J., Oh, J. H. & De Groen, P. C. Polyp-alert: Near real-time feedback during colonoscopy. *Computer Methods and Programs in Biomedicine* **120**, 164–179 (2015).
19. Hirasawa, T., Aoyama, K., Fujisaki, J. & Tada, T. 113 application of artificial intelligence using convolutional neural network for detecting gastric cancer in endoscopic images. *Gastrointestinal Endoscopy* **87**, AB51 (2018).
20. Wang, L., Xie, C. & Hu, Y. Iddf2018-abs-0260 deep learning for polyp segmentation. *BMJ Publishing Group* (2018).
21. Mori, Y. *et al.* Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy: a prospective study. *Annals of internal medicine* **169**, 357–366 (2018).
22. Bychkov, D. *et al.* Deep learning based tissue analysis predicts outcome in colorectal cancer. *Scientific Reports* **8**, 1–11 (2018).
23. Min, M. *et al.* Computer-aided diagnosis of colorectal polyps using linked color imaging colonoscopy to predict histology. *Scientific reports* **9**, 2881 (2019).
24. Bernal, J. & Aymeric, H. Miccai endoscopic vision challenge polyp detection and segmentation. *Web-page of the 2017 Endoscopic Vision Challenge*, https://endovissub2017-giana.grand-challenge.org/home/ (2017).
25. Tajbakhsh, N., Gurudu, S. R. & Liang, J. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Transactions on Medical Imaging* **35**, 630–644 (2016).
26. Pogorelov, K. *et al.* Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the ACM on Multimedia Systems Conference (MMSYS)*, 164–169 (2017).
27. Borgli, H. *et al.* Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data* **7**, 1–14 (2020).

28. Yuan, Y. & Meng, M. Q.-H. A novel feature for polyp detection in wireless capsule endoscopy images. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 5010–5015 (2014).
29. Yuan, Y. & Meng, M. Q.-H. Deep learning for polyp recognition in wireless capsule endoscopy images. *Medical Physics* **44**, 1379–1389 (2017).
30. Karargyris, A. & Bourbakis, N. G. Detection of small bowel polyps and ulcers in wireless capsule endoscopy videos. *IEEE Transactions on Biomedical Engineering* **58**, 2777–2786 (2011).
31. Leenhardt, R. *et al*. A neural network algorithm for detection of gi angiectasia during small-bowel capsule endoscopy. *Gastrointestinal endoscopy* **89 1**, 189–194 (2019).
32. Pogorelov, K. *et al*. Deep learning and handcrafted feature based approaches for automatic detection of angiectasia. In *Proceedings of IEEE Conference on Biomedical and Health Informatics (BHI)*, 365–368 (2018).
33. Pogorelov, K. *et al*. Bleeding detection in wireless capsule endoscopy videos—color versus texture features. *Journal of applied clinical medical physics* **20** (2019).
34. Rahim, T., Usman, M. A. & Shin, S. Y. A survey on contemporary computer-aided tumor, polyp, and ulcer detection methods in wireless capsule endoscopy imaging. *Computerized Medical Imaging and Graphics* **85**, 101767 (2020).
35. Soffer, S. *et al*. Deep learning for wireless capsule endoscopy: a systematic review and meta-analysis. *Gastrointestinal Endoscopy* (2020).
36. Yang, Y. J. The future of capsule endoscopy: The role of artificial intelligence and other technical advancements. *Clinical Endoscopy* **53**, 387 (2020).
37. Park, J. *et al*. Recent development of computer vision technology to improve capsule endoscopy. *Clinical endoscopy* **52**, 328 (2019).
38. Iakovidis, D. K. & Koulaouzidis, A. Software for enhanced video capsule endoscopy: challenges for essential progress. *Nature Reviews Gastroenterology & Hepatology* **12**, 172–186 (2015).
39. Jani, K. K. & Srivastava, R. A survey on medical image analysis in capsule endoscopy. *Current Medical Imaging* **15**, 622–636 (2019).
40. Cheplygina, V., de Bruijne, M. & Pluim, J. P. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis* **54**, 280–296 (2019).
41. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738 (2020).
42. Henaff, O. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, 4182–4192 (PMLR, 2020).
43. Misra, I. & Maaten, L. V. D. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6707–6717 (2020).
44. Bui, T. D., Ravi, S. & Ramavajjala, V. Neural graph learning: Training neural networks using graphs. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, 64–71 (2018).
45. Olympus. The endocapsule 10 system. *Olympus homepage*, https://www.olympus-europa.com/medical/en/Products-and-Solutions/Products/Product/ENDOCAPSULE-10-System.html (2013).
46. Thambawita, V. *et al*. The kvasir-capsule dataset. *Open Science Framework* https://doi.org/10.17605/OSF.IO/DV2AG (2020).
47. Aabakken, L. *et al*. Standardized endoscopic reporting. *Journal of Gastroenterology and Hepatology* **29**, 234–240 (2014).
48. Chetcuti Zammit, S. *et al*. Overview of small bowel angioectasias: clinical presentation and treatment options. *Expert review of gastroenterology & hepatology* **12**, 125–139 (2018).
49. Gomollón, F. *et al*. 3rd european evidence-based consensus on the diagnosis and management of crohn's disease 2016: part 1: diagnosis and medical management. *Journal of Crohn's and Colitis* **11**, 3–25 (2017).
50. Thambawita, V. *et al*. An extensive study on cross-dataset bias and evaluation metrics interpretation for machine learning applied to gastrointestinal tract abnormality classification. *ACM Transactions on Computing for Healthcare* **1**, 1–29 (2020).
51. Thambawita, V. *et al*. The medico-task 2018: Disease detection in the gastrointestinal tract using global features and deep learning. In *Proceedings of the MediaEval 2018 Workshop* (2018).
52. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269 (2017).
53. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778 (2016).
54. Koulaouzidis, A. *et al*. Kid project: an internet-based digital video atlas of capsule endoscopy for research purposes. *Endoscopy international open* **5**, E477–E483 (2017).
55. Bernal, J. & Aymeric, H. Gastrointestinal Image ANAlysis (GIANA) Angiodysplasia D&L challenge. *Web-page of the 2017 Endoscopic Vision Challenge*, https://endovissub2017-giana.grand-challenge.org/home/ (2017).
56. Angermann, Q. *et al*. Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis. In *Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures*, 29–41 (Springer, 2017).
57. Bernal, J. *et al*. Polyp detection benchmark in colonoscopy videos using gtcreator: A novel fully configurable tool for easy and fast annotation of image databases. In *Proceedings of 32nd CARS conference* (2018).
58. Computer-assisted diagnosis for capsule endoscopy (cad-cap) database. *The 2019 GIANA Grand Challenge web-page*, https://giana.grand-challenge.org/WCE/ (2019).
59. Leenhardt, R. *et al*. Cad-cap: a 25,000-image database serving the development of artificial intelligence for capsule endoscopy. *Endoscopy international open* **8**, E415 (2020).
60. Gastrolab. *The Gastrointestinal Site*, http://www.gastrolab.net/index.htm (1996).

Acknowledgements

We would like to acknowledge various people at Bærum Hospital for making the data available. Moreover, the work is partially funded by the Research Council of Norway (RCN), project number 282315 (AutoCap), and our experiments have been performed on the Experimental Infrastructure for Exploration of Exascale Computing (eX3) also supported by RCN, contract 270053.

Author contributions

S.A.H., V.T., P.H., M.A.R., P.H.S. and T.d.L. conceived the experiment(s), S.A.H. and V.T. conducted the experiment(s), P.H.S., H.G., O.O.N., E.N., V.T., S.A.H., M.A.R., P.H. and T.d.L. prepared and cleaned the data for publication, and all authors analysed the results and reviewed the manuscript.

Competing interests

Authors P.H.S., T.J.D.B., H.E., A.P., D.J., T.d.L., M.A.R., and P.H. all own shares in the Augere Medical AS company developing AI solutions for colonoscopies. The Augere video annotation system was used to label the data. There is no commercial interest from Augere regarding this publication and dataset. Otherwise, the authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to P.H.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A.14 Paper XIV: An extensive study on cross-dataset bias and evaluation metrics interpretation for machine learning applied to gastrointestinal tract abnormality classification

Authors: V. Thambawita, D. Jha, H. L. Hammer, H. D. Johansen, D. Johansen, P. Halvorsen, and M. A Riegler

Abstract: Precise and efficient automated identification of gastrointestinal (GI) tract diseases can help doctors treat more patients and improve the rate of disease detection and identification. Currently, automatic analysis of diseases in the GI tract is a hot topic in both computer science and medical-related journals. Nevertheless, the evaluation of such an automatic analysis is often incomplete or simply wrong. Algorithms are often only tested on small and biased datasets, and cross-dataset evaluations are rarely performed. A clear understanding of evaluation metrics and machine learning models with cross datasets is crucial to bring research in the field to a new quality level. Toward this goal, we present comprehensive evaluations of five distinct machine learning models using global features and deep neural networks that can classify 16 different key types of GI tract conditions, including pathological findings, anatomical landmarks, polyp removal conditions, and normal findings from images captured by common GI tract examination instruments. In our evaluation, we introduce performance hexagons using six performance metrics, such as recall, precision, specificity, accuracy, F1-score, and the Matthews correlation coefficient to demonstrate how to determine the real capabilities of models rather than evaluating them shallowly. Furthermore, we perform cross-dataset evaluations using different datasets for training and testing. With these cross-dataset evaluations, we demonstrate the challenge of actually building a generalizable model that could be used across different hospitals. Our experiments clearly show that more sophisticated performance metrics and evaluation methods need to be applied to get reliable models rather than depending on evaluations of the splits of the same dataset—that is, the performance metrics should always be interpreted together rather than relying on a single metric.

Published: ACM Transactions on Computing for Healthcare

A.14. Paper XIV: An extensive study on cross-dataset bias and evaluation metrics interpretation for machine learning applied to gastrointestinal tract abnormality classification

Candidate contributions: D. Jha contributed to the conception and design of the classic machine learning experiments used in the paper. He performed all the experiments for the classical machine learning approaches used in the manuscript. Additionally, D. Jha prepared and revised the manuscript subsequently together with V. Thambawita and all other co-authors.

Thesis objectives: Objective II

An Extensive Study on Cross-Dataset Bias and Evaluation Metrics Interpretation for Machine Learning Applied to Gastrointestinal Tract Abnormality Classification

VAJIRA THAMBAWITA, SimulaMet and Oslo Metropolitan University
DEBESH JHA, SimulaMet and UiT—The Arctic University of Norway
HUGO LEWI HAMMER, Oslo Metropolitan University and SimulaMet
HÅVARD D. JOHANSEN and DAG JOHANSEN, UiT—The Arctic University of Norway
PÅL HALVORSEN, SimulaMet and Oslo Metropolitan University
MICHAEL A. RIEGLER, SimulaMet

Precise and efficient automated identification of gastrointestinal (GI) tract diseases can help doctors treat more patients and improve the rate of disease detection and identification. Currently, automatic analysis of diseases in the GI tract is a hot topic in both computer science and medical-related journals. Nevertheless, the evaluation of such an automatic analysis is often incomplete or simply wrong. Algorithms are often only tested on small and biased datasets, and cross-dataset evaluations are rarely performed. A clear understanding of evaluation metrics and machine learning models with cross datasets is crucial to bring research in the field to a new quality level. Toward this goal, we present comprehensive evaluations of five distinct machine learning models using global features and deep neural networks that can classify 16 different key types of GI tract conditions, including pathological findings, anatomical landmarks, polyp removal conditions, and normal findings from images captured by common GI tract examination instruments. In our evaluation, we introduce performance hexagons using six performance metrics, such as recall, precision, specificity, accuracy, F1-score, and the Matthews correlation coefficient to demonstrate how to determine the real capabilities of models rather than evaluating them shallowly. Furthermore, we perform cross-dataset evaluations using different datasets for training and testing. With these cross-dataset evaluations, we demonstrate the challenge of actually building a generalizable model that could be used across different hospitals. Our experiments clearly show that more sophisticated performance metrics and evaluation methods need to be applied to get reliable models rather than depending on evaluations of the splits of the same dataset—that is, the performance metrics should always be interpreted together rather than relying on a single metric.

CCS Concepts: • **Computing methodologies** → **Cross-validation**; **Supervised learning by classification**; **Machine learning approaches**; • **Applied computing** → **Life and medical sciences**;

This work was funded in part by the Research Council of Norway under project number 263248 (Privaton).
Authors' addresses: V. Thambawita, SimulaMet, Oslo, Norway, and Oslo Metropolitan University, Oslo, Norway; email: vajira@simula.no; D. Jha, SimulaMet, Oslo, Norway, and UiT—The Arctic University of Norway, Tromsø, Norway; email: debesh@simula.no; H. L. Hammer, Oslo Metropolitan University, Norway, and SimulaMet, Oslo, Norway; email: hugoh@oslomet.no; H. D. Johansen and D. Johansen, UiT—The Arctic University of Norway, Tromsø, Norway; emails: {havard.johansen, dag.johansen}@uit.no; P. Halvorsen, SimulaMet, Oslo, Norway, and Oslo Metropolitan University, Oslo, Norway; email: paalh@simula.no; M. A. Riegler, SimulaMet, Oslo, Norway; email: michael@simula.no.

2637-8051/2020/06-ART17 $15.00
https://doi.org/10.1145/3386295

Additional Key Words and Phrases: Medical, computer-aided diagnosis, global features, deep learning, multi-class classification, gastrointestinal tract diseases, polyp classification, Kvasir, Nerthus, CVC-356, CVC-612, CVC-12K, cross-dataset evaluations

ACM Reference format:
Vajira Thambawita, Debesh Jha, Hugo Lewi Hammer, Håvard D. Johansen, Dag Johansen, Pål Halvorsen, and Michael A. Riegler. 2020. An Extensive Study on Cross-Dataset Bias and Evaluation Metrics Interpretation for Machine Learning Applied to Gastrointestinal Tract Abnormality Classification. *ACM Trans. Comput. Healthcare* 1, 3, Article 17 (June 2020), 29 pages.
https://doi.org/10.1145/3386295

1 INTRODUCTION

Cancer is one of the leading causes of death worldwide and a significant barrier to life expectancy [12]. In particular, the gastrointestinal (GI) tract can be affected by a variety of diseases and abnormalities [52]. Using data from the Global Cancer Observatory,[1] Bray et al. [12] estimated that, for 2018, there would be around 5 million new luminal GI cancer incidences and about 3.6 million deaths due to GI cancer.[2] The most frequently diagnosed GI cancers in 2018 for new cases were colorectal cancer (CRC) (6.1%), stomach cancer (5.7%), liver cancer (4.7%), rectum cancer (3.9%), and esophageal cancer (3.2%) out of 36 types of cancers [12].

Gastroscopy and colonoscopy are the most successful medical procedures for GI endoscopy examinations. Among both, colonoscopy has been proven to be an effective preventative method by improving declination in the occurrence of Colorectal Cancer (CRC) by 30% [41]. During a colonoscopic procedure, an endoscopist inserts a colonoscope carefully through the anus to examine the rectum and colon. A tiny wide-angle video camera mounted at the end of the colonoscope captures a live video signal of the internal mucosa of the patient's colon. The endoscopist uses the video signal for real-time diagnosis of the patient, where one of the primary goals is to identify and remove abnormalities such as polyps [77].

The current EU guidelines [74] recommend GI tract screening for all people older than 50 years. Such regular screenings can be of great significance for early detection and prevention of cancer inside the GI tract, but they are challenging due to many factors. Moreover, a colonoscopy examination is entirely an operator-dependent screening procedure [63]. The detection rate of GI tract lesions mostly relies on the clinical experience of the gastroenterologist. The shortage of experienced gastroenterologists, and the clinicians' tiredness and lack of concentration during the colonoscopic examination, can lead to missing polyps that otherwise would be detected [68]. The estimated miss rate for the subject undergoing a colonoscopy examination is 25% [39].

Although considerable work has been done to develop and improve systems for automatic polyp detection, the performance of existing solutions is still behind that of an expert endoscopist [7, 16, 44, 75, 76]. Most of the published works in the field use non-public datasets or develop models from too-small training, validation, and test sets [7, 75, 76]. The performance metrics used to measure the performance of methods are also not sufficient (e.g., see the first part of Table 1). Thus, it is difficult for researchers to compare and reproduce some of the present related works. Moreover, the state-of-the-art research in this field does not present the generalizability of their solutions using cross-dataset evaluations. As a result, it creates a distrust for applying these machine learning (ML) solutions in practice.

An automatic and efficient computer-aided diagnosis (CAD) system in a clinic could assist medical experts during the endoscopic and colonoscopy procedure to improve the detection rate by finding unrecognized lesions and act as a second observer by providing better insights to the gastroenterologist concerning the presence and types of lesions. With this inspiration, we conducted five experiments to classify 16 classes of GI tract conditions

[1] https://gco.iarc.fr.
[2] We have considered the statistic of esophagus, stomach, colon, rectum, anus, gallbladder, and pancreas.

Table 1. Overview of the Related Work

Reference	Year	REC	PREC	SPEC	ACC	MCC	F1	Rk	FPS
Hwang et al. [27]	2007	0.9600	0.8300	—	—	—	—	—	15
Li & Meng [40]	2012	0.8860	—	0.9620	0.9240	—	—	—	—
Zhou et al. [83]	2014	0.7500	—	0.9592	0.9077	—	—	—	—
Wang et al. [76]	2014	0.8140	—	—	—	—	—	—	0.14
Mamonov et al. [43]	2014	0.4700	—	0.9000	—	—	—	—	—
Wang et al. [77]	2015	0.9770	—	—	0.9570	—	—	—	10
Riegler et al. [57]	2016	0.9850	0.9388	0.7250	0.8770	—	—	—	~300
Shin & Balasingham [63]	2017	0.9082	0.9271	0.9176	0.9126	—	—	—	—
Riegler et al. [58]	2017	0.9850	0.9390	0.7250	0.8770	—	—	—	~75
Yu et al. [78]	2017	0.5005	0.4917	—	0.9471	—	0.4830	0.5357	—
Pogorelov et al. [54]	2017	0.8260	0.8290	0.9750	0.9570	—	0.8260	0.8020	46
Agrawal et al. [1]	2017	—	—	—	0.9610	0.8260	0.8470	—	—
Naqvi et al. [45]	2017	—	0.7665	0.9660	0.9420	0.7360	0.7670	—	—
Petscharnig et al. [48]	2017	0.7550	0.7550	0.9650	0.9390	0.7200	0.7550	0.7240	—
Pogorelov et al. [52]	2017	0.9060	0.9060	0.9810	0.9690	—	—	—	30
Yuan et al. [79]	2018	0.8180	0.7232	—	—	—	0.7431	—	—
Wang et al. [75]	2018	0.9438	—	0.9592		—	—	—	—
Mori & Kudo [44]	2018	>0.9000	—	>0.9000	—	—	—	—	—
MediaEval 2018 Medico Task [53] (The following experiments were done using the 2018 Medico dataset.)									
Hoang et al. [25]	2018	0.9281	**0.9426**	**0.9963**	**0.9932**	**0.9312**	**0.9342**	**0.9398**	23
Hicks et al. [24]	2018	0.9218	0.9378	0.9959	0.9924	0.9228	0.9236	0.9325	624
Borgli et al. [10]	2018	0.8572	0.8708	0.9956	0.9918	0.8555	0.8555	0.9280	—
Kirkerød et al. [36]	2018	0.8433	0.8514	0.9944	0.9896	0.8366	0.8367	0.9082	—
Dias & Dias [18]	2018	0.8205	0.8414	0.9938	0.9885	0.8146	0.8114	0.8983	8.61
Taschwer et al. [70]	2018	0.8673	0.8826	0.9933	0.9876	0.8641	0.8662	0.8897	—
Ostroukhova et al. [46]	2018	0.8236	0.8281	0.9911	0.9835	0.8115	0.8145	0.8539	1E-100
Khan & Tahir [33]	2018	0.6203	0.7173	0.9767	0.957	0.6025	0.5868	0.6302	43329
Steiner et al. [64]	2018	0.4219	0.5146	0.9717	0.9469	0.3901	0.3913	0.5368	—
Ko et al. [37]	2018	0.5005	0.4916	0.9715	0.9471	0.4608	0.4829	0.5357	0.5357
Thambawita et al. (Ours) [71]	2018	**0.9361**	0.9319	**0.9963**	**0.9932**	0.9283	0.9297	0.9397	—

REC, recall (sensitivity); ACC, accuracy; MCC, Matthews correlation coefficient; F1, F1-score; Rk, Rk correlation coefficient; FPS, frames per second.
The results of the Medico Task may slightly vary compared to the proceeding note papers because of different ways of calculating the multi-class performance metrics by the organizers. The highest score for the MediaEval 2018 Medico Task is marked in bold.

for the Medico Multimedia Task at MediaEval 2018 [53]. One example for each of the 16 classes is depicted in Figure 1.

In this work, we focus on identifying the limitations of generalizing ML models across different datasets and how to interpret the evaluation metrics in that context. For this, we are using global feature (GF)-based and deep learning (DL)-based methods that performed well at the 2018 Medico Task [53], where one specific dataset was used. In addition, here we explore the different performance metrics of both methods (GF and deep learning (DL)) to identify the limitations of each. We show that combined complex deep neural network (DNN) models outperform other methods. Finally, we explore how multi-class models perform on polyp and non-polyp detection with and without retraining the model for the two specific classes. The effects of retraining for classifying the sub-categories of the same dataset and using them in other datasets are analyzed in detail to identify

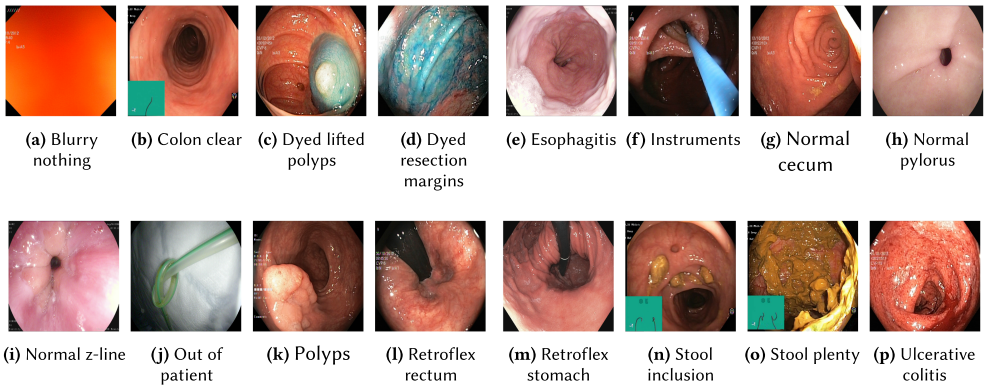


Fig. 1. Sample images of GI findings. Each image represents one of the 16 classes from the dataset used for the Medico 2018 Challenge [50, 51].

the cross-dataset generalization capabilities of our models. We emphasize that a large number of performance measures do not show the real performance of ML models. We also highlight the necessity of having cross-dataset evaluations to determine the real capabilities of ML models before using them in clinical settings.

To study cross-dataset bias and metrics interpretation, our contributions are as follows:

(1) We present five ML classification models to classify multi-class findings (anatomical landmark, pathological findings, polyp removal conditions, and normal findings) of the GI tract. Using a limited imbalanced dataset, we experiment with approaches ranging from Global Feature (GF) approaches to simple Deep Neural Network (DNN) and complex DNN approaches with transfer learning. Moreover, we present a detailed evaluation using six performance metrics to show the real classification performance of ML models. In addition, we analyze and present detailed evaluation results of using multi-class classification ML models for classifying binary classes (sub-classes of the multi-class categories) with and without retraining to evaluate the generalizability of our models. We emphasize the difficulties of using well-performing ML methods in cross-datasets as a result of the reluctance of ML models to cross-dataset generalization. We present this negative impact with the aid of another evaluation using the receiver operating characteristic (ROC) curve and the precision-recall (PR) curve of the best model. We also demonstrate when a Receiver Operating Characteristic (ROC) curve is good to use and when it is better to use a PR curve.
(2) With the preceding point, we emphasize the requirement of detailed cross-dataset evaluations to identify generalizability of ML models before using them as universal models in live applications. Because good performance measures with a single dataset do not necessarily imply good real-world performance, we argue that researchers should present cross-dataset evaluations for building a generalizable model rather than presenting performance values for the test datasets, which is separated from the same training data source.

Moreover, with respect to the 2018 Medico Task [53], our best DNN method achieved the highest recall, specificity, and accuracy for multi-class classification of the GI tract findings. We achieved a Matthews correlation coefficient (MCC) (0.0029 less) and an Rk correlation coefficient[3] (0.0001 less) nearly equal to the winning team. With this achievement, we demonstrate all of the steps, from designing to training and testing, for reaching such performance using this model and its expandability using different pre-trained networks.

[3]The Rk correlation coefficient and the MCC were the most important considered metrics for winning the 2018 Medico Task.

In Section 2, we present related work and the performance of relevant existing solutions. Section 3 discusses the methodology used for our GF-based approaches and the theoretical foundation for our work. The DNN-based approaches are similarly described in Section 4. Our experimental results are presented and analyzed in Section 5, followed by a discussion in Section 6 on how our results can be helpful to other researchers. In Section 7, we conclude our findings.

2 RELATED WORK

Many methods and algorithms have been proposed for GI tract disease detection/classification using videos and images from colonoscopy and gastroscopy as input. The problem of polyp detection has by far received the most attention by researchers. Images and videos of polyps and other abnormalities inside the GI tract are usually collected using a specific-purpose camera and imaging system, like ScopeGuide from Olympus. The information gathered from these types of devices may be of great significance for later examination and must be handled with great care. Polyps generally have characteristics different from the normal surrounding healthy tissue and are often easy for clinicians to detect. There are several good datasets available for training and testing on polyps (the details about the available polyp dataset can be found in other works [16, 30]), and binary classification methods are relatively straightforward to implement.

The other active research efforts include developing an automatic and real-time detection system for GI bleeding, ulcerative lesion, blood-based abnormality, tumor, and angiectasia, and for multi-class data of the GI tract that comprise anatomical landmarks (e.g., z-line, pylorus, and cecum), pathological findings (e.g., esophagitis and ulcerative colitis), and normality and regular findings (e.g., normal colon mucosa and stool). Suitable datasets for research in these areas are less developed and lack adequate content. Similarly, presented performance measures in these areas are not adequate because of not presenting enough performance metrics or not presenting cross-dataset evaluations.

Table 1 presents an overview of important works related to GI disease detection/classification and the 2018 Medico Task [53] using Computer-Aided Diagnosis (CAD), from automatic polyp detection to multi-class disease detection and classification systems. The dataset used for the experiments in the first half of the Table 1 is different. Therefore, the results cannot be directly compared; however, the results in the lower half of the table can be compared, as the algorithms are tested on the same dataset.

Most of the research in the medical field only focuses on designing an automated disease detection system for detecting or classifying specific disease or abnormality, such as polyp detection or ulcer detection. Because patients may suffer from more than one type of disease at the time, a working multi-class disease detection system will help treatment. The performance of existing multi-abnormality detection systems is, however, not satisfactory and cannot assist doctors in CAD in real time while undergoing colonoscopies. Furthermore, these research works have not evaluated all performance metrics at once to analyze the real behavior of their classification models. Yet none of the preceding methods have performed cross-dataset evaluations to prove the capabilities for using the ML models in real CAD systems.

For handcrafted (HC) feature-based methods, image descriptors like global or local image features (e.g., color, texture, and edges) are extracted, and later on, various ML classifiers (e.g., logistic model tree (LMT) [71], random forest classifier [43], or support vector machine (SVM) [76]) are employed to perform analysis using these features. HC descriptors (manually designed features) are useful for the gastroenterologist while identifying specific abnormality regions inside the GI tract. For instance, as blood has a particular range of chromaticity, we can specify a specific chromaticity range where features of bleeding abnormality seem to be concentrated [31]. Riegler et al. [58] achieved an F1-score of 0.909 with a GF-based approach and an F1-score of 0.875 with a DL-based approach with a multi-class GI tract dataset. With the ASU-Mayo polyp dataset, the GF-based approach achieves an F1-score of 0.961, whereas the DL-based approach could obtain 0.936. They further suggested that the combination of both approaches may lead to improved performance. In addition, previous work by Riegler

et al. [56] reveals that although only detecting whether a frame contains an irregularity or not, GFs can beat local features—for instance, they can at least reach the same results with regard to detection/classification and perform better than local features with regard to processing speed. In all of these works, researchers presented performance metrics using a test dataset selected from the same dataset used for the training data. Therefore, these results do not reflect the actual practical performance of the proposed methods.

A few past studies used information such as the color and texture of polyps to sketch HC descriptors [2, 3, 13, 28, 29, 32, 68]. The other category of methods for automated polyp detection used shape, intensity, edge, and spatio-temporal information. For instance, Hwang et al. [27] appropriated elliptical shape features to detect the occurrence of polyps in colonoscopy videos. Bernal et al. [7] proposed a polyp detection technique by utilizing a polyp region descriptor, which is dependent on the depth of the valley image and introduced a region growing method to detect polyps in colonoscopy images. Bernal et al. [8] additionally used valley information and enhanced their approach by improving the polyp localization results to almost 30%. Bernal et al. [6] also performed additional evaluations using valley information and demonstrated better performance, especially for smaller polyps and decreased the polyp miss rate. Park et al. [47] utilized spatio-temporal features for automatic polyp detection. The recently completed related work that uses the cross-sectional profile to detect protruding polyps automatically is the polyp detection system Polyp-Alert [77], which can provide near real-time feedback during colonoscopies. However, the system is limited to polyp detection and is slow for live examinations. Tajbakhsh et al. [68] proposed a method for automatic polyp detection from colonoscopy videos that uses context information to remove non-polyp and shape information to localize polyp reliably. Riegler et al. [58] utilized various GFs and achieved high precision and recall above 90%. Yuan et al. [79] employed a bottom-up and top-down saliency approach for automated polyp detection. Although these research works discuss improving the performance of ML models, they have not evaluated the performance of the ML models with cross-datasets.

As convolutional neural network (CNN) architectures have achieved exceptional gains in medical image and video analysis tasks, more recent work on polyp detection is mainly based on Convolutional Neural Networks (CNNs). Tajbakhsh et al. [67] proposed a 2D-CNN method for polyp detection by learning discriminative spatial and temporal features. Yu et al. [78] proposed a 3D fully convolutional network to deal with the challenges related to automatic polyp detection for colonoscopy videos. Zhang et al. [81] suggested an enhanced single-shot multi-box detector (SSD) called *SSD-GPNet* for detecting gastric polyps, which have the potential for achieving real-time detection up to 50 FPS using Nvidia Titan V. Furthermore, they use GPDNet [82] to classify three classes of pre-cancerous gastric disease.

Researchers have also compared HC and DL methods. For instance, Pogorelov et al. [52] and Riegler et al. [58] compared several (HC- and DL-based) localization methods. Pogorelov et al. [49] evaluated their approach utilizing HC and DL methods on different available datasets for real-time polyp detection. Their best model with a generative adversarial network (GAN) obtained detection specificity of 94% and accuracy of 90.9%. The preceding research works presented good performance for predicting polyps, whereas Pogorelov et al. [49] presented evaluation results of the models with cross-datasets. However, having overlapped data sources in the cross-datasets, the shown results do not reveal the real performance in cross-dataset evaluations.

The pre-trained models, along with transfer learning mechanisms, are also becoming popular because of their capability to outperform state-of-the-art algorithms even with less training data, where the limited size of the medical dataset for experiments has always been a problem to yield better results. For the detection and localization of the polyps [9, 69], the pre-trained models with a CNN mechanism also achieve promising results. A comparison of DL with GFs for GI tract disease detection has also been presented. Pogorelov et al. [54] presented 17 different methods for multi-class classification of GI tract data with the limited number of the training dataset. They used both GFs and DL approaches in their work. They achieved the best result with modified ResNet50 features using the LMT classifier. They reached an Rk value of 80.2% and an F1-score of 82.6% with 2,000 training and 2,000 test datasets.

Comparing with the polyp detection approaches, the research on multi-class disease detection/classification on a complete GI tract system is minimal. However, for multi-class disease detection/classification (including polyp detection) inside the GI tract, we note a few contributions made in this area. For example, the authors of numerous works [1, 10, 18, 24, 25, 33, 36, 37, 46, 48, 64, 70] presented their approach in classifying disease inside the GI tract utilizing the Kvasir dataset and the MediaEval Medico 2018 dataset. The latter is a combination of the Nerthus [50] and Kvasir [51] datasets.

Hicks et al. [24] show how fine tuning a CNN model using transfer learning with data from different source domains affects classification performance. In their case, extending the generic ImageNet dataset with medical images from the LapGyn4 and Cataract-101 dataset, they obtained a high Matthews Correlation Coefficient (MCC) score of 0.9228. For the 2018 Medico Task, we proposed solutions based on GFs and DL-based methods for multi-class classification of GI tract findings [71]. Our best model was a combination of two pre-trained networks, ResNet-152 and Densenet-161, along with a multi-layer perceptron (MLP). Here, we obtained an MCC of 94.21%, an F1-score of 94.58%, and an accuracy of 99.32%. This was one of the best results in the MediEval 2018 Medico Task Challenge. We discuss the model introduced by Thambawita et al. [71] in detail in this article and reproduce similar results. Based on those models, we provide and discuss the requirement of detailed evaluations using multiple performance metrics and cross-dataset evaluations.

Recent related works show promising results in terms of evaluation metrics, such as both sensitivity and specificity despite various challenges (e.g., difficulties arise due to a dataset obtained from different modalities). The limitation with most of the recent approaches is that they target only specific problems, like bleeding detection or polyp detection. Current systems are either (i) too narrow for a flexible, multi-disease detection/classification system; (ii) tested only on a limited datasets, too small to show whether the systems would work well in hospitals, (iii) provide low processing performance for a real-time system or ignore the system performance entirely; (iv) problematic with regard to overfitting of the specific dataset and lead to unreliable results; or (v) tested using datasets that are not publicly available, making it difficult to compare the approaches with others.

In some cases, GF-based approaches produce better results. For some methods, DL performs better. The CNN approaches and pre-trained network with transfer learning mechanism approaches have the best results in most of the cases. Reusing already existing DL architectures and pre-trained models leads to excellent results in, for example, the ImageNet classification tasks. For example, the HC feature-based approach works well for true negative (TN) detection/classification tasks.

To reduce the damage of the dataset bias problem, Khosla et al. [34] directed their experiments for both classification tasks and detection problems. They used different datasets from different domains in the training stage to generalize the features extracted from their ML model. However, SVM was used as the main algorithm, and the DNN dataset bias problem was not addressed.

With the goal of making researchers aware of the dataset bias problems, Torralba and Efros [72] did informative research using basic datasets and basic ML models with the classification and detection task of computer vision. Initially, they trained a simple linear Support Vector Machine (SVM) to make a simple classifier to name a given dataset from 12 different datasets, which have nearly the same categories. They were inspired by the research done by Dollár et al. [20] to detect pedestrians. The result of the experiment for dataset classification shows a clear diagonal in the confusion matrix (CM). This implies that there are clear dataset bias features, and that these datasets have the same categories. Therefore, researchers want to apply cross-dataset generalization for avoiding dataset bias behavior of ML models. Moreover, they discussed selection bias, capture bias, category or label bias, and negative bias as the main factors for the dataset bias. This directs our research to do additional experiments to identify the significant factors of the cross-domain data generalization in the medical domain, which is more critical than the general image classification.

The classification of GI diseases is more complicated than a simple real-world object classification task where one detects faces or recognizes characters. Typical GI tract datasets are heavily imbalanced—for example, the 2018 Medico Task dataset consists of 16 classes of anatomical landmarks, pathological findings, polyp removal

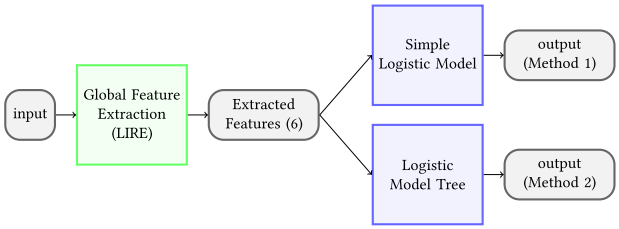


Fig. 2. Block diagram of the proposed method 1 and method 2. The pipeline starts with the input of images. GFs are extracted using the LIRE framework. These features are then used for two different classification algorithms (the SL model for method 1 and LMT for method 2).

cases, and normal and regular findings, where the polyp class has a maximum of 613 images, and the instrument class has a minimum of only 4 images. Additionally, medical datasets are captured using different endoscopic instruments, and some of the images can be noisy, blurry, over- or under-exposed, and interleaved, and can have superfluous information within the image, contain borders, and be affected by specular reflections caused by the instrument light source. Some of the images may have bleeding, whereas other images can be partially covered by stool or mucus. Moreover, the organs from mouth to anus can have multiple lesions showing different diseases, abnormalities, and internal injuries. Thus, the preceding situation leads to the necessity of distinguishing between various classes of GI tract findings. In this scenario, not only high precision and recall but also high accuracy and MCC become essential for developing an automated generalizable multi-class classification system. This implies the real requirement of measuring and analyzing all performance metrics at once. Furthermore, to prove the generalizability of models, cross-dataset evaluations are required.

3 GF-BASED APPROACHES

GFs or descriptors are features computed over the whole image or covering a regular sub-section of an image. GFs represent the overall properties of an image and are often used in image retrieval, image compression, image classification, object detection, and image collection search and distance computing [54]. Examples of GFs are shape matrix, histogram-oriented gradients (HOGs), Co-HOG, and invariant moments (Hu, Zernike). The LIRE [42] framework can be used to extract HC GFs such as texture, color distribution, and the histogram of brightness. The most commonly used GFs include joint composite descriptor (JCD), Tamura, color layout (CL), edge histogram (EH), autocolor correlogram, pyramid histogram of oriented gradients (PHOGs), color and edge directivity descriptor (CEDD, local binary patterns, and scalable color (SC). Figure 2 shows the architecture of the proposed GF-based methods (1 and 2). These methods use six selected GFs and the best ML classifiers for the provided dataset.

Feature engineering is among the most crucial and challenging parts for approaching any ML and computer vision problem. Based on the findings of Pogorelov et al. [54] and Riegler et al. [59], we choose to used JCD, Tamura, CL, EH, autocolor correlogram, and PHOG. The combinations of these features represent the overall properties of the images. We can even add more GFs, but doing so may increase the noise to the image features, which again would hurt the classification performance. Moreover, we have formulated the problem of GI tract anomaly classification as a multi-class (16-class) classification of different findings including anomalies, landmarks, and clinical markings. With the provided dataset, we computed the GFs of each image. A multi-class classification problem is a general and well-studied ML problem, and there is a variety of methods available to solve this issue with higher performance. Therefore, we sent the extracted GFs to many available ML classifiers. The whole experiment was completed with the development dataset. The 2018 Medico Task [53] shows the best classification rates with SimpleLogistic (SL) [38] and LMT [38] classifiers.

3.1 Method 1: The SL Classifier

In method 1, we combine the SL classifier from the Weka software [22] to build a linear logistic regression (LR) model with the LogitBoost [21] utility for determining attributes. The SimpleLogistic (SL) classifier can deal with binary class classification, multi-class classification, missing class, and nominal class. It can handle different types of attributes, such as binary attributes, nominal attributes, date attributes, missing values, unary attributes, and empty nominal attributes [38]. In a linear LR classifier, a simple (linear) model fits the data, and the method of model fitting is pretty stable, leading to low variances.

LogitBoost is utilized for determination of the most appropriate attributes in the data at the time of executing LR, which is done by performing a simple regression in every iteration before it converges to a solution of maximum likelihood. Therefore, LogitBoost, with a simple regression function that acts as a base learner, is utilized for fitting the logistic models. The optimum number of iterations associated with the LogitBoost algorithm to function is cross validated, which leads to the automatic selection of the attribute [65]. The SL classifier has a built-in attribute selection (if the default parameter is not changed): it stops computing simple linear regression models (i.e., performing LogitBoost iterations) when the cross-validated classification error no longer decreases. With the extracted features using LIRE, the SL classifier has not only the highest classification accuracy but also takes the lowest classification time (i.e., lowest computational complexity) when compared with other ML classification algorithms.

3.2 Method 2: The LMT

In method 2, we use the Logistic Model Tree (LMT) classifier from the Weka software. The LMT is a classification model related to a supervised training algorithm, which is a combination of LR and decision tree learning techniques [38, 62]. Thus, the LMT is considered an analogue model for solving classification problems. In the logistic variant, information gain is utilized for splitting, the LogitBoost algorithm generates an LR model at each node in the tree, and the CART algorithm [62] is utilized for pruning the tree.

The LMT uses a cross-validation (CV) technique to find several LogitBoost iterations to prevent overfitting of the training data. The LogitBoost algorithm accomplishes additive LR, which is achieved by least-square fits for every class M [19], which is shown in Equation (1):

$$L_M(x) = \sum_{i=1}^{n} \beta_i + \beta_0. \tag{1}$$

Here, β_i denotes the coefficient of the ith component of the vector x, and n denotes number of features. The LMT model uses the linear LR method to calculate the posterior probabilities of the leaf nodes [38], which is shown in Equation (2):

$$L_M(X) = -\frac{\exp\left(L_M(X)\right)}{\sum_{M=1}^{D} \exp\left(L_M(X)\right)}. \tag{2}$$

Here, D denotes the number of classes, and $\mathrm{L}_{\mathrm{M}}(\mathrm{X})$ stands for the least-square fits. The least-square fits $\mathrm{L}_{\mathrm{M}}(\mathrm{X})$ are transformed in such a way that $\sum_{M=1}^{D} \exp(L_M(X))\Big)$ is equal to zero.

4 DL APPROACHES

For our transfer learning approaches, we selected two DNNs: ResNet-152 [23] and DenseNet-161 [26] based on the top-1 error rate and top-5 error rate for the ImageNet [17, 61] classification as given in the PyTorch documentation [14]. Then, we chose ResNet-152 as the base model of the first DL approach, and this base model experiment was done under method 3 (the model is illustrated in Figure 3). This selection was made based on preliminary experiments. In the preliminary experiments, ResNet-152 showed better performance than DenseNet-161. This DenseNet-161 was in second place in the performance ranking when we compared stand-alone pre-trained DL models.

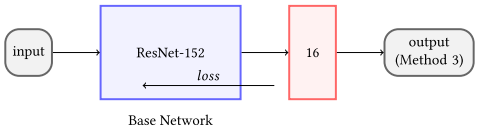


Fig. 3. Block diagram of method 3. The input is an image that is passed to a ResNet-152 neural network. A final softmax layer outputs the scores for the 16 classes.

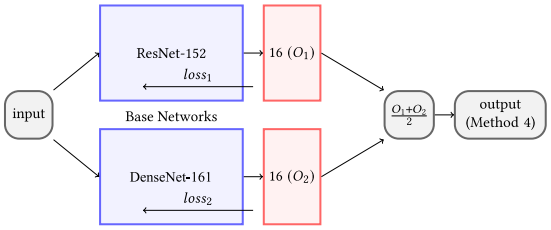


Fig. 4. Block diagram of method 4. The input image is in parallel passed to a ResNet-152 and a DenseNet-161 neural network. Two separate softmax layers calculate separate 16-class scores, which are finally combined.

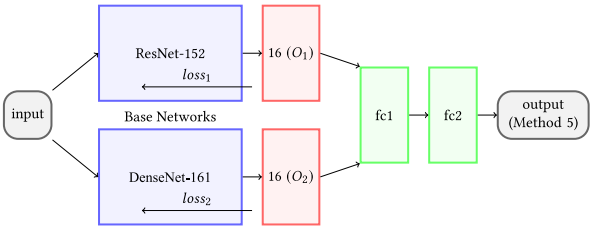


Fig. 5. Block diagram of method 5. It is similar to method 4, but instead of a single step to combine the output scores of the two neural networks, two fully connected layers are utilized.

In DL methods 4 and 5 (as illustrated in Figures 4 and 5), we used both pre-trained ResNet-152 and DenseNet-161 using the ImageNet dataset. In the following sections, we discuss data pre-processing mechanisms and training mechanisms used for all three DL methods. In later sections, we discuss these methods one by one with their fine-tuning mechanisms with more comprehensive explanations.

For the transfer learning methods, we use the data pre-processing tool of the PyTorch library to (i) resize input images, (ii) crop marginal annotations of the medical images, (iii) normalize the pixel values of input images, and (iv) apply random image transformations. Regarding image resizing, all images of the dataset were resized into 224×224 because ResNet-152 and DenseNet-161 accept images with these dimensions. By applying the central-cropping transformation of PyTorch, we minimized unnecessary effects for the final predictions of DNNs affected from annotated marks (green boxes) of the medical images as shown in Figure 1(b), (n), and (o). Center cropping did not remove important information from the images because we cropped down to 224×224 from 256×256. Our experiments show that removing the whole green box, such as those in Figure 1(b), (n), and (o), from the images by applying a larger crop size is not advisable, because for some images, too much content of the finding is lost with a large crop size. When applying the normalization function to the input images, a standard deviation (σ) of 0.5 and a mean (μ) of 0.5 were used with the normalization function in PyTorch. The mathematical equation used in this function is given in Equation (3), and c represents the three channels R, G, and B of input images. The *input* represents a tensor of pixel values of each layer. We used random transformations, random horizontal

flips, random vertical flips, and random rotations from PyTorch as data augmentation techniques.

$$input_c = \frac{input_c - \mu_c}{\sigma_c}; \quad where \quad c = [0, 1, 2] \tag{3}$$

For training all DNNs, the transfer learning mechanism was used. Then, we used cross-entropy loss [15] with weighted classes as given in Equation (4) to calculate the loss values of the DNNs:

$$loss(x, class) = weight[class] \times \left(-x[class] + \ln \left(\sum_j \exp(x[j]) \right) \right). \tag{4}$$

In this equation, the weight parameter value is calculated inversely proportional to the image count in the corresponding class. In other words, class weight values are high when the classes have fewer images. However, the inbuilt cross-entropy function given in PyTorch is used instead of implementing it from scratch. While doing preliminary experiments, we observed that there was not any effect from weighted cross-entropy loss. Then, we used the normal cross-entropy loss (Equation (5)) function for calculating the loss of the DNNs:

$$loss(x, class) = -\ln \left(\frac{\exp(x[class])}{\sum_j \exp(x[j])} \right) = -x[class] + \ln \left(\sum_j \exp\left(x[j]\right) \right). \tag{5}$$

As the optimizer of all DNNs, the stochastic gradient descent (SGD) [11] method with a momentum [66] was applied. We selected this optimizer because of its stable learning mechanism in contrast to the highly unstable learning pattern of other methods [35, 60, 80], as they show fast convergence.

During the training procedure, we changed the learning rate manually based on the progress of learning curves rather than using the inbuilt learning rate schedulers of PyTorch. Initially, we began with a high learning rate. Then, the learning rate was reduced by a factor of 10 if the training process did not show good progress in the learning curves. Finally, model weights of the best epoch based on the best validation accuracy were saved to use in the inference stage.

4.1 Method 3: DNN Approach Based on ResNet-152

Method 3 is the base method that uses only ResNet-152. A block diagram of this is illustrated in Figure 3. In this method, the last layer of ResNet-152 is modified to output 16 classes of the 2018 Medico Task from 1,000 classes of ImageNet. Usually, we freeze first layers (there is not a logical way to select the number of layers to freeze) of pre-trained networks when we do transfer learning. Then, we train the last and the new layers using the new domain data. Finally, the entire network is trained after unfreezing all parameters of the network (a method known as fine tuning).

We performed preliminary experiments to identify the influence of the preceding freezing-unfreezing technique compared to using simple fine tuning. Both techniques showed the same performance at the end of the training process, and we could not gain any performance benefit from the freezing-unfreezing method, as using the simple fine-tuning method was faster. Therefore, we decided to use the simple fine-tuning method for all experiments.

In method 3, we started the training process with a learning rate of 0.001. Then, the learning rate was decreased by a factor of 10 if we could not see any performance improvement for the validation dataset. We repeated this change of learning rate until the model came to a good stable position. In this experiment, the SGD method was used as the optimization method with a momentum of 0.9.

4.2 Method 4: DNN Approach Based on ResNet-152 and DenseNet-161

In method 4, as illustrated in Figure 4, we used two pre-trained networks on ImageNet: ResNet-152 and DenseNet-161. These networks were retrained separately into the Medico dataset using the same procedure used in

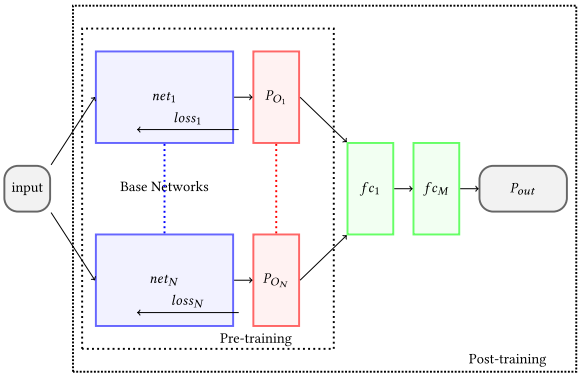


Fig. 6. Block diagram of the proposed parallel DNN merging. The training process is split into a pre-training (pre-training of individual models) and post-training step (training the whole network architecture).

method 3. Before this retraining, the networks were modified to classify the 16 classes. Then, we calculated an average probability of the two probability vectors (V_{Resnet_152} and $V_{Densenet\_161}$) output by the two separate networks: ResNet-152 and DenseNet-161. By calculating the average of these two probability vectors ($V_{answer} = \frac{1}{2}\,(V_{Resnet\_152} + V_{Densenet\_161})$), we accepted the cumulative probability decision rather than the individual decision. Using the average from these two networks, we expected to have a good decision with high confidence. For example, if the two networks return high probability values for the same class, the class probability value (confidence of classifying to that class) is high. However, when one network has a high probability and the other network has a low probability for a specific class, then the final probability value is around 0.5. This value infers that confidence about the particular class is not good enough for the final decision.

In this model, the probability of the final answer depends on the average values rather than the highest probability value returned from one of the two models. Here, the problem is that the prediction suggested from the highest probability value of one model may be the correct class compared to the selected category from the average. Finally, we trained the model using a learning rate of 0.001. In addition, we decreased the learning rate by a factor of 10 when the model did not show convergence. A momentum of 0.9 with SGD was used as in method 3.

4.3 Method 5: DNN Approach Based on ResNet-152, DenseNet-161, and MLP

Method 5 was designed to overcome the problem of method 4. The block diagram of this method is illustrated in Figure 5. The simple averaging method was not enough to make a final decision when the two networks provided two different answers. As a solution, an Multi-Layer Perceptron (MLP) was introduced instead of the simple averaging method. Then, we trained only this MLP with the pre-trained ResNet-152 and DenseNet-161 for the Medico dataset to decide the final prediction based on the probabilities that come from two networks. More details about designing this complex model are discussed in Sections 4.3.1 and 4.3.2.

4.3.1 Extendable Method 5. In this section, we show how we can improve accuracy using multiple cumulative probabilistic decisions by extending method 5 into $N \geq 2$ DNNs. In this general model, as illustrated in Figure 6, we divide the whole training process into the following four steps: (1) pre-training of individual models, (2) model selection for merging, (3) merging models with an MLP, and (4) post-training and fine tuning. Let $NETS = \{net_1, net_2, \ldots, net_N\}$ be the set of pre-trainer networks using the ImageNet dataset and P_{O_i} be the returned probability vector for model net_i.

In step 1 (pretraining), we train each DNN $net_i \in NETS$ as much as possible using the transfer learning mechanism until it gives the best predictions as described in method 3 (using different loss functions; $loss_1$ to $loss_N$).

The DNNs have their unique prediction capabilities within the given classification problem. Then, we analyze the CM of the best outcome of each DNN.

In step 2 (selection), we select networks that give different diagonals of CMs (the diagonal of a CM represents correct classifications) compared to other CMs of selected DNNs. If the diagonal of CM of network $net_i = CM_i$, then we select networks that have $CM_i \neq CM_j; j = [1, 2, \ldots, i-1, i+1, \ldots, N]$. The goal of this comparison is to identify DNN models that have different classification performances compared to each other. Equal diagonals of CMs do not imply that the networks are identical for their classifications, because there might be models that give the same diagonal numbers but lead to different classifications for a given image. If the case of equal diagonals occurs, we have to compare correctly classified images to identify the differences. The number of DNNs selected for the final training may or may not be equal to the initial number of pre-trained DNNs depending on similarities in some of the CMs.

In step 3 (merging), we use an MLP to merge all outputs of the selected DNNs. The MLP consists of M layers that take $\sum_{i=1}^{N} length\_of(P_{O_i})$ number of inputs and output P_{out} probability vector according to the given classification problem. Then, step 4 can be started by freezing all the pre-trained DNNs and training only the new MLP until it shows a good validation performance. Optionally, we can retrain the whole model without freezing any layer if we cannot achieve a performance improvement by training only the new MLP.

4.3.2 Method 5 Used by This Research Work. According to the procedure discussed in Section 4.3.1, our implementations of method 5 were designed using two parallel networks ($N = 2$): ResNet-152 and DenseNet-161. Then, we analyzed two CMs, which came from ResNet-152 and DenseNet-161. These two networks were pre-trained according to the given classification problem. Because $CM_{Resnet\_152} \neq CM_{Densenet\_161}$, we combined the two networks with an MLP. This comparison of CMs was done visually using colormaps. However, if the visual inspection of CMs is hard, mathematical operations can be used. Moreover, if the CMs are equal completely, a manual inspection of the classified images is required to identify the differences of model classifications. After combining, we froze two DNNs to proceed to the post-training step. In our experiments, the input layer of the MLP consisted of 32 input nodes. The output of the MLP was a probability vector with 16 values, which is equal to the number of classes of the Medico dataset. We used two fully connected layers, with 32 neurons and 16 neurons. In the post-training step, we started training only the MLP with a learning rate of 0.01. To do the post-training, multi-class cross-entropy loss and Stochastic Gradient Descent (SGD) were used.

5 RESULTS

In this section, we discuss the experimental setup, datasets, and results obtained from our experiments. Using these presented results, we emphasize that high scores for performance metrics do not always show the actual performance of ML methods. To show this, we present well-performing ML models that achieved good results for their performance values. Using cross-dataset testing, we present a detailed analysis of evaluation metrics to emphasize that they are not always representative to identify the real performance of models.

For all experiments, we used the same hardware platform with an Intel Core i7 eighth-generation processor with 16 GB of DDR4 RAM and an 8-GB NVIDIA GeForce 1080 GPU. However, we practiced two different software frameworks for implementing our methods. To implement the GF-based methods (1 and 2), we used the Weka framework [22]. We used the PyTorch framework for the DNN-based methods (3, 4, and 5).

5.1 Datasets

For the work performed in this article, we used the following four datasets: the 2018 Medico dataset [55], CVC-356-plus (a modified version of CVC-356 [6, 7, 73]), CVC-612-plus (a modified version of CVC-612 [6, 7, 73]), and CVC-12k [4, 5]. The training and testing datasets of the 2018 Medico Task were derived from the Kvasir dataset [51] and Nerthus dataset [50], consisting of 16 classes as shown in Table 2. These images consist of different anatomical landmarks (z-line, pylorus, cecum), pathological findings (esophagitis, polyps, ulcerative

Table 2. Summary of the 2018 Medico Dataset

Type	Images in the Development Set (#)	Images in the Test Set (#)
Blurry-nothing	176	39
Colon-clear	267	1,070
Dyed-lifted-polyps	457	590
Dyed-resection-margins	416	583
Esophagitis	444	483
Instruments	36	165
Normal-cecum	416	604
Normal-pylorus	439	569
Normal-z-line	437	636
Out-of-patient	4	6
Polyps	613	423
Retroflex-rectum	237	194
Retroflex-stomach	398	399
Stool-inclusions	130	508
Stool-plenty	366	1,920
Ulcerative-colitis	457	551

The first column shows the names of the different findings. The second and third columns show the number of images in the development and test sets.

Table 3. Overview of the Datasets Used for Our Experiments

Dataset	Training	Testing	Images (#)	Polyps (#)	Non-Polyps (#)
2018 Medico—Development	X	—	5,906	613	5,293
2018 Medico—Testing	—	X	8,740	423	8,317
CVC-356-plus	X	X	2,285	356	1,929*
CVC-612-plus	X	X	1,316	612	704
CVC-12k	—	X	11,954	10,025	1,929

*We replaced this image set with a new image set (with 1,171 images) extracted from a clear colon video collected from the Bærum Hospital, Norway, in the second stage of this research to avoid the overlap between the training data and the testing data.

In total, we have five different datasets, but the Medico dataset is split into a development part and a test part for the challenge. The training and testing columns indicate how the dataset was used in the experiments. Polyps and non-polyps indicate the number of findings. Medico and CVC-356-plus represent a bias toward non-findings. CVC-612-plus is a quite balanced dataset, and CVC-12k presents a bias toward findings. Datasets were chosen based on these distributions to represent common cases in medical imaging datasets.

colitis), endoscopic polyp removal cases (dyed and lifted polyp, dyed resection margin), and normal findings (normal colon mucosa, stool) in the GI tract. The dataset also contains images with different degrees of the Boston Bowel Preparation Scale (BBPS), ranging from 0 to 3. Some of the original images contain the endoscope position marking probe. These are seen as a small green box located in the bottom corners, showing its configuration and location of the image frame. The images used in the study were captured using an electromagnetic imaging system (Scopeguide, Olympus, Europe) [51]. In Table 3, we present a summary of the uses of the 2018 Medico dataset and other datasets for polyp and non-polyp classifications.

The Medico development dataset was used to train our ML models in the first stage. However, this dataset consists of a highly imbalanced number of images, as summarized in Table 2. Within this, the out-of-patient class had only 4 images to train our models. Therefore, only in the first stage, we used an additional 30 images

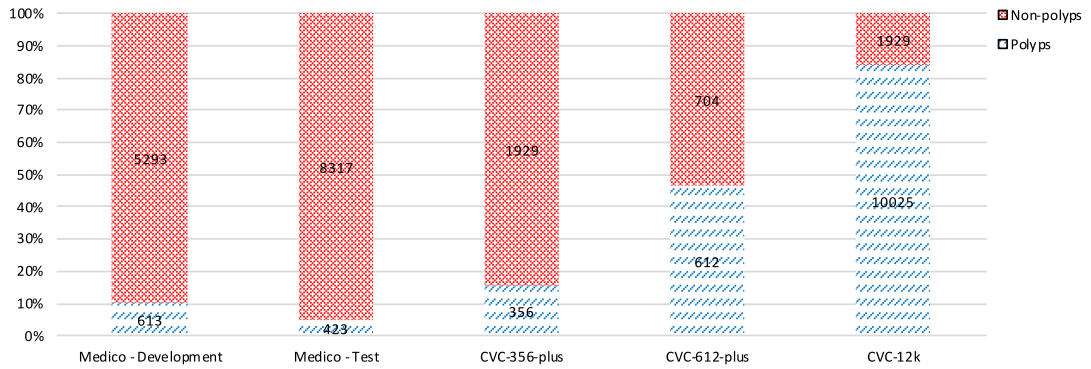


Fig. 7. Ratios of findings to non-findings in the datasets (polyp/non-polyp). The *X* axis represents the different datasets used for the binary classification. The *Y* axis represents the percentage of polyps and non-polyps. The numbers inside the bars show the actual number of polyp and non-polyp images.

that were selected randomly from the Internet to fill this class in the training dataset. These were images of flowers, vehicles, and other general stuff in our everyday life and did not have any relationship with this class. The advantage of this technique is discussed in the discussion of Section 6.

When we discussed the ML models' generalizability in the second part of the article, we used the CVC datasets to retrain and test our models. The CVC-356-plus dataset is the modified version of the CVC-356 [6, 7, 73] dataset that has only polyp images. In that modification, we added 1,929 non-polyp images from the CVC-12k [4, 5] dataset to the CVC-356 dataset and created a new dataset called *CVC-356-plus*. Similarly, the CVC-612-plus dataset was created by extending the CVC-612 dataset [6, 7, 73]. For this CVC-612-plus dataset, we added 704 non-polyp images extracted from new GI tract videos collected by the Bærum Hospital, which is part of the Vestre Viken Hospital Trust in Norway. The content of the CVC-12k dataset underwent a minor reorganization by filtering and grouping polyp and non-polyp images into two separate folders. However, the content and number of images in CVC-12k were not otherwise changed. Therefore, we refer to it by its common name.

In the second part of our research, we used the CVC-356-plus and CVC-612-plus datasets for retraining our models to classify polyps and non-polyps. In only this part of the research, we replaced 1,929 non-polyp images of the CVC-356-plus dataset with 1,171 newly extracted images from a clean and healthy colon video collected from the same hospital. We did this modification to avoid the overlap between the non-polyp images of the CVC-356-plus training dataset and the CVC-12k testing dataset.

For the dataset preparation stage, we focused on the number of polyp and non-polyp images in each dataset to analyze the correlation between the data distribution and the model performance. A bar graph of this data distribution is illustrated in Figure 7. We chose to include different proportions for the number of polyps and non-polyps to keep diversity of data percentages in each test case. In the CVC-356-plus dataset, the polyp percentage is low compared to the non-polyp percentage. In the CVC-612-plus dataset, percentages of polyps and non-polyps are around 50%. In contrast, the CVC-12k dataset has a higher polyp percentage than the non-polyp percentage. Due to this, we can study the effects of data imbalance in the training and testing datasets on the performance and interpretability of the metrics.

5.2 Analyzing Results

We discuss our results in two main sections: (i) the 16-class classification task based on the 2018 Medico Task and (ii) the polyp and non-polyp classification task to analyze generalizability of ML models.

Table 4. Evaluation Results of the 2018 Medico Task (as Provided by the Organizers of the 2018 Medico Task) [71] for the Five Methods Used in This Article

Method	REC	PREC	SPEC	ACC	MCC	F1
1	0.8457	0.8457	0.9897	0.9807	0.8353	0.8456
2	0.8457	0.8457	0.9897	0.9807	0.8350	0.8457
3	0.9376	0.9376	0.9958	0.9922	0.9335	0.9376
4	0.9400	0.9400	0.9960	0.9925	0.9360	0.9400
5	0.9458	0.9458	0.9964	0.9932	0.9421	0.9458

Based on the official results, method 5 was the best one based on the MCC score.

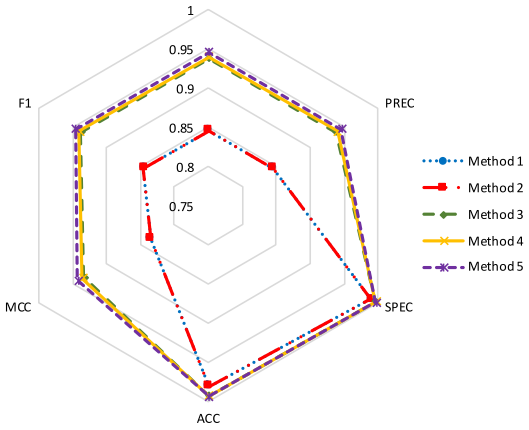


Fig. 8. Performance comparison of all five classification models for the 16 classes of the 2018 Medico test dataset. Methods 1 and 2 are similar in results but different from the other three methods (note that measurements start at 0.75).

5.2.1 16-Class Classification. In this 16-class classification task, the training dataset of the 2018 Medico Task was split into a 70% training dataset and a 30% validation dataset. Then, the test data given by the organizers was used to test the performance of five methods for classifying 16 classes of the GI tract findings.

We evaluated our five models based on the results collected by the organizers. The evaluated results of the main five models are tabulated in Table 4. With an MCC score of 0.9421, method 5 showed the best performance for classifying the 16 classes of GI tract findings. However, our GF-based approaches did not show results competitive with the DNN methods. The GF model introduced in method 1 could reach an MCC score of 0.8353. This result showed the best performance record for a GF-based method. A clear performance difference between the GF-based methods and the DNN-based methods can be seen in Figure 8. In this plot, we compared this performance difference using six performance measures: recall (REC), precision (PREC), specificity (SPEC), accuracy (ACC), MCC, and F-score (F1). According to this plot, it is clear that the areas of the hexagons covered by the GF methods are smaller than the areas covered by DNN methods. These results imply that three DL methods outperform two GF methods.

The CM of method 5 collected from the organizers of the 2018 Medico Task is tabulated in Table 5 for the in-depth investigation. According to the CM, we can identify two main bottlenecks to improve the performance of method 5. The first one is misclassification between esophagitis and normal-z-line, and the second one is misclassification between dyed-lifted-polyps and dyed-resection-margins. Therefore, images from these classes were manually examined to identify the reasons for these misclassifications. For the conflict between esophagitis

Table 5. CM of Method 5 (Our Best Model) Based on the Medico Test Dataset

		Actual Class														
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Predicted Class	Ulcerative-colitis (A)	500	—	—	—	–	—	—	—	39	—	3	—	1	1	—	7
	Esophagitis (B)	3	432	48	—	—	—	—	—	—	—	—	—	—	—	—	—
	Normal-z-line (C)	1	121	513	—	—	—	—	—	—	—	—	—	1	—	—	—
	Dyed-lifted-polyps (D)	1	—	—	522	31	—	—	—	—	—	2	—	—	—	—	34
	Dyed-resection-margins (E)	—	—	—	33	532	—	—	—	—	—	1	—	—	—	—	17
	Out-of-patient (F)	—	—	—	—	1	5	—	—	—	—	—	—	—	—	—	—
	Normal-pylorus (G)	3	3	2	—	—	—	559	—	—	—	2	—	—	—	—	—
	Stool-inclusions (H)	—	—	—	—	—	—	—	501	7	—	—	—	—	—	—	—
	Stool-plenty (I)	1	—	—	—	—	—	—	—	1,918	—	—	—	—	—	—	1
	Blurry-nothing (J)	1	—	—	—	—	—	—	—	1	37	—	—	—	—	—	—
	Polyps (K)	10	—	—	1	—	—	1	—	—	—	358	6	—	1	—	46
	Normal-cecum (L)	18	—	—	—	—	—	—	—	—	—	6	578	—	—	—	2
	Colon-clear (M)	1	—	—	—	—	—	—	5	—	—	—	—	1,063	—	1	—
	Retroflex-rectum (N)	3	—	—	—	—	—	—	—	—	—	2	—	—	188	1	—
	Retroflex-stomach (O)	—	—	—	—	—	—	1	—	—	—	—	—	—	2	395	1
	Instruments (P)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	165

The diagonal value represents true predictions (number of images) of the model. A through P are the classes corresponding to the class names in the first column. The most confusion can be observed between classes B and C, and classes D and E. Looking at the images, we can see that they are quite similar in their visual features (colors, texture, etc.).

and normal-z-line, the reason is the very close locations of these two landmarks in the GI tract. However, the confusion between dyed-lifted-polyps and dyed-restrictions is caused because of the same color patterns and the same texture structures of both types of images. With these limitations, method 5 showed the best performance with an MCC of 0.9421, which was the important measurement to win the 2018 Medico Task. Based on the MCC value, we won second place in the 2018 Medico Task. The winning team [25] relabeled the development dataset and also generated more images out of the provided instruments class by placing the instrument as a foreground over the images of dyed-lifted-polyps, dyed-resection-margins, and ulcerative colitis to balance the instrument class for improving performance. However, we developed the model by only using the images provided by the task organizers for a fair comparison of the approaches with the limited dataset. Then, our next experiments were conducted to find the reusability of these well-performed models in different datasets with polyp and non-polyp categories (sub-categories of the 16 classes of primary tasks).

5.2.2 Polyp and Non-Polyp Classification Using the Pre-Trained Models. The following analysis was performed to identify the polyp classification ability of our five models on the same test dataset and different CVC datasets. The 16-class classification results collected from the Medico Task organizers were analyzed to calculate polyp detection performance in the Medico test data. Moreover, our models were tested with the CVC-356-plus, CVC-612-plus, and CVC-12k datasets without any modifications to the five models to compare the performance of polyp detection.

According to the correct and incorrect classifications of polyps and non-polyps in the test datasets, the first large column of Table 6 was calculated to measure the polyp detection performance of five models. In this evaluation process, all 15 classes except the polyp class were considered as the non-polyp classification because the number of outputs is 16 in the first models. For comparison, the MCC values of these tests are plotted in Figure 9. This graph shows that the polyp detection performance of the same dataset (the testing dataset of the Medico Task) is higher than on the completely new datasets (CVC-356-plus, CVC-612-plus, and CVC-12k) for both the

Table 6. Polyp Classification Results with and without Retraining for All Datasets and Methods

		Without Retraining						With Retraining to 2-Class Classification					
	M	REC	PREC	SPEC	ACC	MCC	F1	REC	PREC	SPEC	ACC	MCC	F1
Test Dataset	1	0.7834	0.4899	0.9635	0.9558	0.5987	0.6028	0.9550	0.9630	0.6740	0.9553	0.5430	0.9590
	2	0.7834	0.4899	0.9635	0.9558	0.5987	0.6028	0.9540	0.9630	0.6840	0.9537	0.5400	0.9580
	3	0.9733	0.8088	0.9897	0.9890	0.8819	0.8835	0.9813	0.6577	0.9772	0.9773	0.7934	0.7876
	4	0.9599	0.8467	0.9922	0.9908	0.8969	0.8997	0.9813	0.7384	0.9845	0.9843	0.8440	0.8427
	5	0.9572	0.8463	0.9922	0.9907	0.8954	0.8984	0.9706	0.7516	0.9857	0.9850	0.8470	0.8471
CVC-356-plus	1	0.3089	0.1053	0.5158	0.4835	−0.127	0.1571	0.8450	0.7990	0.1700	0.8446	0.0750	0.7780
	2	0.3089	0.1053	0.5158	0.4835	−0.127	0.1571	0.8510	0.8420	0.2070	0.8512	0.1930	0.7930
	3	0.7865	0.3738	0.7569	0.7615	0.4198	0.5068	0.8118	0.5547	0.8797	0.8691	0.5978	0.6591
	4	0.6713	0.4003	0.8144	0.7921	0.4010	0.5016	0.6517	0.4150	0.8305	0.8026	0.4068	0.5071
	5	0.6685	0.4837	0.8683	0.8372	0.4737	0.5613	0.6713	0.6408	0.9305	0.8902	0.5906	0.6557
CVC-612-plus	1	0.7696	0.7969	0.8295	0.8016	0.6008	0.7830	0.6980	0.8070	0.6530	0.6983	0.4740	0.6590
	2	0.7696	0.7969	0.8295	0.8016	0.6008	0.7830	0.7220	0.8170	0.6800	0.7218	0.5140	0.6910
	3	0.8415	0.6242	0.5597	0.6907	0.4137	0.7168	0.8382	0.6136	0.5412	0.6793	0.3932	0.7086
	4	0.8627	0.6559	0.6065	0.7257	0.4803	0.7452	0.8578	0.6890	0.6634	0.7538	0.5265	0.7642
	5	0.8137	0.6501	0.6193	0.7097	0.4379	0.7228	0.8007	0.7061	0.7102	0.7523	0.5104	0.7504
CVC-12k	1	0.4858	0.8391	0.5158	0.4907	0.0012	0.6154	0.1650	0.7880	0.8370	0.1651	0.0130	0.0530
	2	0.4858	0.8391	0.5158	0.4907	0.0012	0.6154	0.1650	0.8210	0.8380	0.1699	0.0290	0.0630
	3	0.6112	0.9289	0.7569	0.6347	0.2722	0.7373	0.6033	0.9631	0.8797	0.6479	0.3558	0.7419
	4	0.6236	0.9458	0.8144	0.6544	0.3241	0.7517	0.6459	0.9519	0.8305	0.6757	0.3539	0.7696
	5	0.5936	0.9591	0.8683	0.6379	0.3401	0.7333	0.5576	0.9766	0.9305	0.6178	0.3595	0.7099

M, method.
For training, 2018 Medico development data was used. We can observe that for some datasets, retraining seems to improve performance.

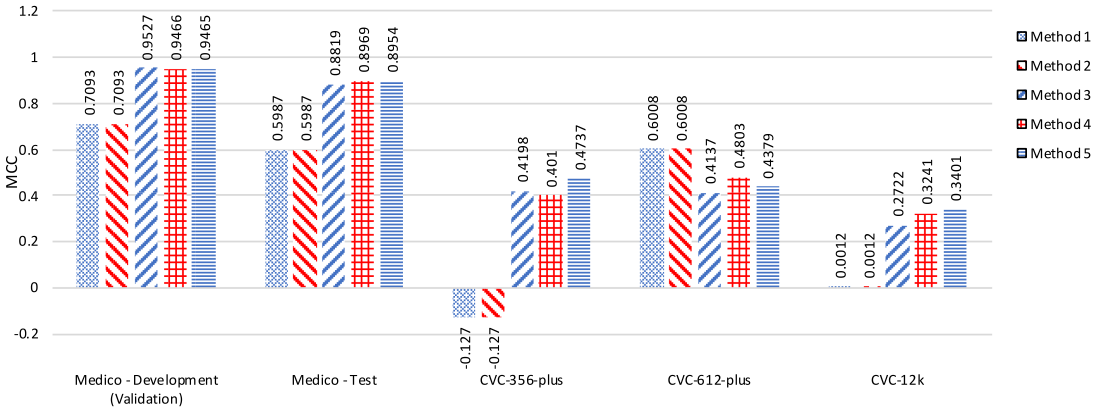


Fig. 9. Polyp and non-polyp classification capabilities (based on MCC) of all five methods that were trained using 2018 Medico development data to classify 16 classes. For most cases, methods 3 through 5 perform best. For the CVC-612-plus test data, methods 1 and 2 perform best.

GF-based approaches and the DNN approaches. This is the first analysis, and we emphasize that it shows that researchers need to do cross-dataset evaluations to prove the real capabilities of ML models.

From the first column of Table 6 and Figure 9, it is clear that the performance of the GF methods for different datasets (CVC-356plus, CVC-612-plus, and CVC-(356+612) dataset) is unpredictable because it presents huge value fluctuations in the graph with a negative MCC value. This shows the incapability of GF methods to make predictions on different datasets. The negative values of MCC in this experiment, such as −0.127 for the CVC-356-plus dataset, indicate that there is no agreement or only a non-relevant relationship between target and prediction. An MCC around zero would mean that the classifier is deciding randomly, and MCCs above zero would indicate correct classification. The closer to −1 or 1, the stronger the indication for being wrong or correct, respectively. However, the polyp detection performance of the GF-based methods in the CVC-612-plus dataset outperforms the DNN methods with an MCC value of 0.6008, whereas the best DNN method shows an MCC value of 0.4803. This prediction accuracy of the GF methods can be identified as an erroneous prediction, because the performance of this method for the other two CVC datasets shows poorer MCC scores than those of DNN-based approaches. Moreover, the DNN-based approaches show considerable steady MCC values for all new datasets, implying that the DNN methods are more generalizable than the GF methods.

Because the performance gap between the 16-class classification and polyp classification showed differences, we retrained our models to classify only the polyp and non-polyp classes. Therefore, our next experiments were performed to test how retraining our five ML models to classify polyps and non-polyps will influence performance.

For the retraining experiments, we first retrained the two GF methods with new ARFF files generated for polyp and non-polyp categories. Second, in the retraining stage of the three DNN methods, we changed only the last layer into two outputs. However, we did not change the loss function from categorical cross-entropy into binary cross-entropy because two-class categorical cross-entropy is equal to binary cross-entropy. Moreover, we retained the original optimization functions. Then, we retrained all five models using the same Medico dataset, which has only polyp and non-polyp classes. The results of these experiments are tabulated in the right columns of Table 6.

The results in Table 6 show that it can be difficult to evaluate the models and interpret the results after retraining for two-class classification. All MCC values of the five methods tested on the CVC-356-plus data show improvements. Similarly, for the CVC-612-plus test data, methods 4 and 5 show performance improvements from MCC values of 0.4803 and 0.4379 to 0.5265 and 0.5104, respectively. In contrast, methods 1, 2, and 3 show a performance drop, which is indicated by MCC values 0.6008, 0.6008, and 0.4137 reduced to 0.4740, 0.5140, and, 0.3932, respectively. Therefore, we extended our experiment by introducing additional retraining options with the CVC-356-plus and CVC-612-plus datasets. After that, the retraining process can be categorized as retraining the models to classify polyps and non-polyps using (i) only the same Medico training dataset (as tabulated in Table 6), (ii) the Medico dataset with the CVC-356-plus dataset, (iii) the Medico dataset with the CVC-612-plus dataset, and (iv) the Medico dataset with the CVC-356-plus and CVC-612-plus datasets. Then, our testing datasets are limited to two datasets: the Medico test dataset and the CVC-12k dataset. Results related to these new retraining processes can be seen in Table 7. When the models are trained using the balanced CVC-612-plus dataset in combination with the 2018 Medico development data, the DNN models show better MCC values (0.8189, 0.8555, and 0.8606) for methods 3, 4, and 5, respectively. This is true for the Medico test data and the two smaller CVC datasets. Moreover, the MCC values for the CVC-12k test data also achieve the best MCC values of 0.1421, 0.1418, and 0.1802 for methods 3, 4, and 5. An important observation from the CVC-12k dataset is also that looking at all other metrics but MCC and specificity could mislead to the assumption that the results are good—for example, scores above 0.8 for accuracy, which is often used as the only indicator for performance in similar studies.

In the first comparison, we plotted performance changes for the retraining with the different training datasets and tested them on the Medico test dataset. The changes in the Recall (REC), Precision (PREC), Specificity (SPEC), Accuracy (ACC), MCC, and F-score (F1) values can be seen as hexagon plots in Figure 10(a), (c), (e), (g), and (i),

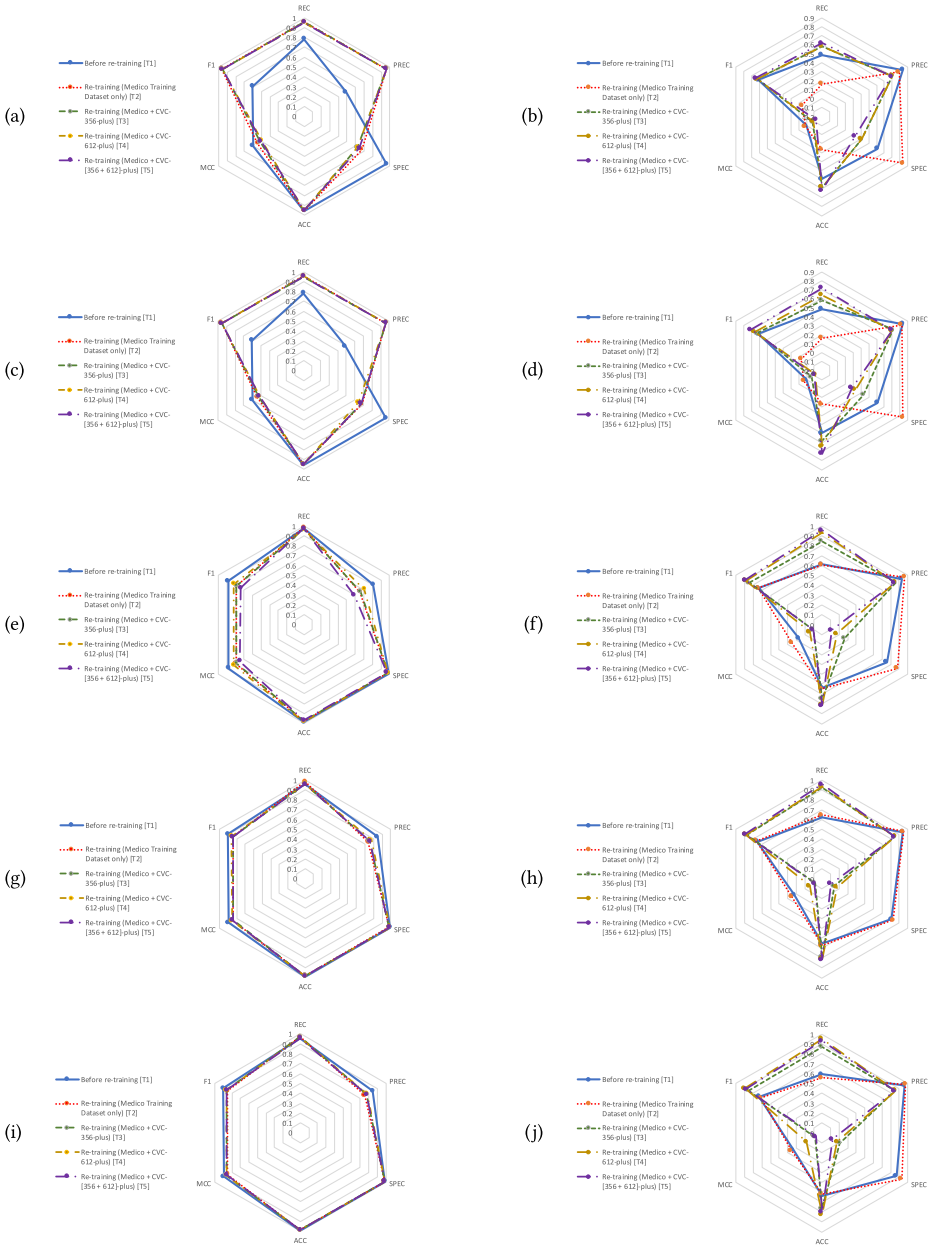


Fig. 10. Polyp and non-polyp classification using the proposed ML methods: 1, 2, 3, 4, and 5. The first column (sub-figures (a, c, e, g, i)) shows the results of the Medico test dataset, and the second column (sub-figures (b, d, f, h, j)) shows the results of the CVC-12k dataset. The methods are represented as follows: (a) and (b) for method 1, (c) and (d) for method 2, (e) and (f) for method 3, (g) and (h) for method 4, and (i) and (j) for method 5, respectively.

Table 7. Evaluation Results on Using CVC-356-plus and CVC-612-plus Combined as Training Data with Retraining to Classify Polyps and Non-Polyps

			MedicoTest Data						CVC-12k					
		M	REC	PREC	SPEC	ACC	MCC	F1	REC	PREC	SPEC	ACC	MCC	F1
Retraining Datasets with Medico Data	CVC-356-plus	1	0.9550	0.9610	0.6230	0.9549	0.5160	0.9570	0.5840	0.7040	0.3090	0.5836	−0.084	0.6320
		2	0.9520	0.9620	0.6710	0.9521	0.5260	0.9560	0.5810	0.7100	0.3360	0.5807	−0.065	0.6310
		3	0.9626	0.6630	0.9781	0.9775	0.7887	0.7852	0.8423	0.8565	0.2665	0.7494	0.1052	0.8493
		4	0.9599	0.7526	0.9859	0.9848	0.8427	0.8437	0.9192	0.8481	0.1441	0.7941	0.0810	0.8822
		5	0.9706	0.7773	0.9876	0.9868	0.8623	0.8633	0.8694	0.8507	0.2068	0.7625	0.0802	0.8599
	CVC-612-plus	1	0.9510	0.9590	0.6270	0.9508	0.4970	0.9540	0.5840	0.7030	0.3040	0.5842	−0.087	0.6320
		2	0.9530	0.9610	0.6430	0.9530	0.5160	0.9560	0.6400	0.6970	0.2240	0.6395	−0.117	0.6660
		3	0.9652	0.7092	0.9823	0.9816	0.8189	0.8177	0.9325	0.8546	0.1752	0.8103	0.1421	0.8918
		4	0.9572	0.7766	0.9877	0.9864	0.8555	0.8575	0.9336	0.8544	0.1731	0.8109	0.1418	0.8922
		5	0.9626	0.7809	0.9879	0.9868	0.8606	0.8623	0.9486	0.8571	0.1778	0.8242	0.1802	0.9005
	CVC-{356+612}	1	0.9500	0.9600	0.6480	0.9503	0.5050	0.9540	0.6180	0.6930	0.2280	0.6179	−0.129	0.6520
		2	0.9500	0.9610	0.6710	0.9503	0.5170	0.9550	0.7200	0.7010	0.1820	0.7199	−0.105	0.7100
		3	0.9733	0.5909	0.9699	0.9700	0.7458	0.7354	0.9537	0.8479	0.1109	0.8177	0.1028	0.8977
		4	0.9545	0.7596	0.9865	0.9851	0.8443	0.8460	0.9543	0.8463	0.0995	0.8164	0.0874	0.8971
		5	0.9599	0.7771	0.9877	0.9865	0.8571	0.8589	0.9278	0.8462	0.1239	0.7981	0.0699	0.8851

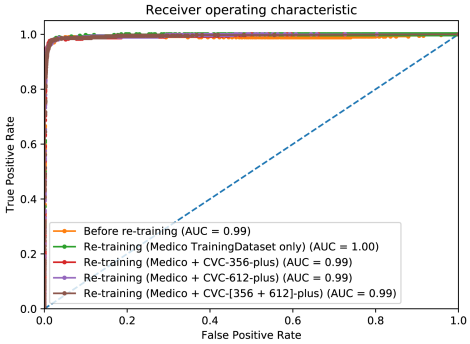
The 2018 Medico test dataset and the CVC-12k dataset are the test datasets. Using the balanced CVC-612-plus as training data, we achieve the best results. Combining CVC-356-plus and the CVC-612-plus does not improve performance. Overall, the performance is better on the Medico test dataset.

which correspond to methods 1, 2 ,3, 4, and 5, respectively. In these plots, T1 is used to present performance values before retraining the ML models into 2-class classification (binary classification). In this case, 15 classes except for the polyp class of the 16 classes were considered as the non-polyp class, and the polyp class is counted as the same polyp class. Furthermore, from T2 to T5, lines are used to present models with only two outputs. The T2 plot represents models' performance for the retraining using the Medico training dataset. Similarly, T3, T4, and T5 represent the retraining process using the Medico dataset and the CVC-356-plus dataset, the Medico dataset, and the CVC-612-plus dataset, and the Medico dataset, the CVC-356-plus dataset, and the CVC-612 dataset, respectively.

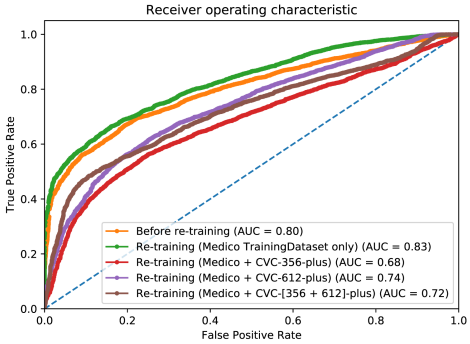
In the second series of experiments in this session, the same experiments were performed and tested on the CVC-12k dataset. The results obtained from these experiments are tabulated in Tables 6 and 7. Then, relevant results from these tables are plotted in Figure 10(b), (d), (f), (h), and (j). These plots use line notations similar to the preceding experiments.

Using the plot series in Figure 10, we can examine the reusability of ML models to classify polyps and non-polyps, which are sub-classes of the primary classes on the task. For example, if we compare plots in Figure 10(a) and (b), then we can know how method 1 performs to classify polyps and non-polyps within the test dataset the same as the training dataset and within an entirely new dataset. While investigating these plots, the proportion of the number of polyps and non-polyps is an important factor in explaining the shape of these hexagon plots.

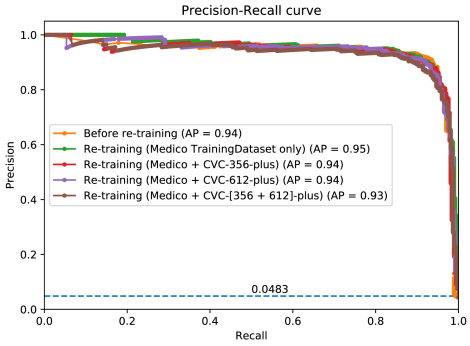
If we compare the GF methods (Figures 10(a) through (d)) and the DL methods (Figures 10(e) through (j)), it is clear that the DL methods outperform the GF methods in both the Medico Task and polyp classification task introduced in this article. This implies that the DL methods are capable of extracting deep features that cannot be extracted by manual feature extraction methods used by the GF methods. With the retraining process in the GF methods, we can see performance differences between the Medico dataset and the CVC-12k dataset. The main conclusion that we make is that GF-based methods are not able to capture the underlying patterns that would allow for efficient classification; thus, their performance is low.



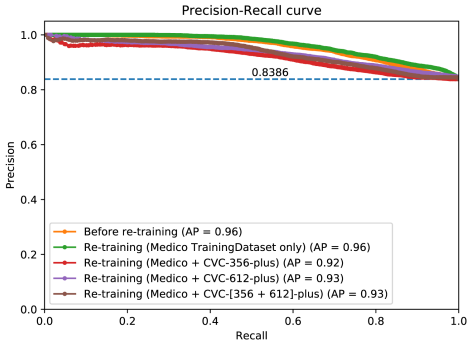
(a) ROC and PR curves for the Medico test dataset.



(b) ROC and PR curves for the CVC-12k dataset.



(c) PRC curves for the Medico test dataset.



(d) PRC curves for the CVC-12k dataset.

Fig. 11. ROC and Precision-Recall Curve (PRC) curves for method 5 trained on the CVC-356-plus and CVC-612-plus datasets as mentioned in the legends. Testing datasets are the CVC-12k and Medico test datasets. Overall, good performance can be observed in both ROC and PR curves. For CVC-12k, the PR curve shows the interesting case of a high random baseline for a biased dataset.

Plots in the first and second columns in Figure 10 show completely different behaviors for the same retraining process when we use different test datasets. The test dataset for the first column comes from the same domain as the training data, and the test dataset for the second column comes from the completely new domain, such as the CVC-12k dataset. To investigate these unusual performance changes, we generated and examined ROC and PR curves for the best DNN model (method 5). The ROC and PR curves for method 5 with the Medico test data (for the plot in Figure 10) are depicted in Figure 11(a) and (c). Similarly, the ROC and PR curves for method 5 with CVC-12k data (for the plot in Figure 10) are plotted in Figure 11(b) and (d).

Analysis of ROC curves is more robust for ML models that are used with balanced datasets, whereas PR curves are more valuable for ML methods when the methods engage with imbalanced datasets. However, we have used both curves in this paper to investigate the behavior of these curves while we are using highly imbalanced datasets. Consequently, the PR curves show completely different baseline values of 0.0483 for the Medico test dataset and 0.8386 for the CVC-12k dataset. The small baseline value arises in the plot in Figure 11(c) as a result of small polyps to the non-polyp proportion in the Medico test dataset. Conversely, the high baseline value in Figure 11(d) appears there as an effect on a high ratio of polyps to non-polyps.

To get a better understanding of the above plots, we selected the plots in Figure 10(i) and (j), and ROC and PR curves in Figure 11. With this selection, first, we analyzed T1 and T2 from the hexagon plots and the

Table 8. Method 5: Training Only the MLP vs. the Complete DNN

Test Data	T	Training Only the MLP						Training the Whole DNN					
		REC	PREC	SPEC	ACC	MCC	F1	REC	PREC	SPEC	ACC	MCC	F1
Medico Test Data	T1	0.9572	0.5859	0.9698	0.9692	0.7357	0.7269	0.9706	0.7773	0.9876	0.9868	0.8623	0.8633
	T2	0.9599	0.7804	0.9879	0.9867	0.8591	0.8609	0.9626	0.7809	0.9879	0.9868	0.8606	0.8623
	T3	0.9626	0.6316	0.9749	0.9744	0.7684	0.7627	0.9599	0.7771	0.9877	0.9865	0.8571	0.8589
CVC-12k	T1	0.6984	0.8972	0.5842	0.6799	0.2184	0.7854	0.8694	0.8507	0.2068	0.7625	0.0802	0.8599
	T2	0.7588	0.8993	0.5583	0.7265	0.2565	0.8231	0.9486	0.8571	0.1778	0.8242	0.1802	0.9005
	T3	0.7614	0.8933	0.5272	0.7236	0.2352	0.8221	0.9278	0.8462	0.1239	0.7981	0.0699	0.8851

T, the additional training dataset that was added to the Medico dataset; T1, Medico dataset + CVC-356-plus; T2, Medico dataset + CVC-612-plus; T3, Medico dataset + CVC-356-plus + CVC-612-plus.

corresponding ROC and PR curves. Although T2 shows a performance loss compared to T1 in Figure 10(i), Figure 10(j) shows that T2 achieves a performance improvement over T1. Next, we look for the reasons for these performance changes.

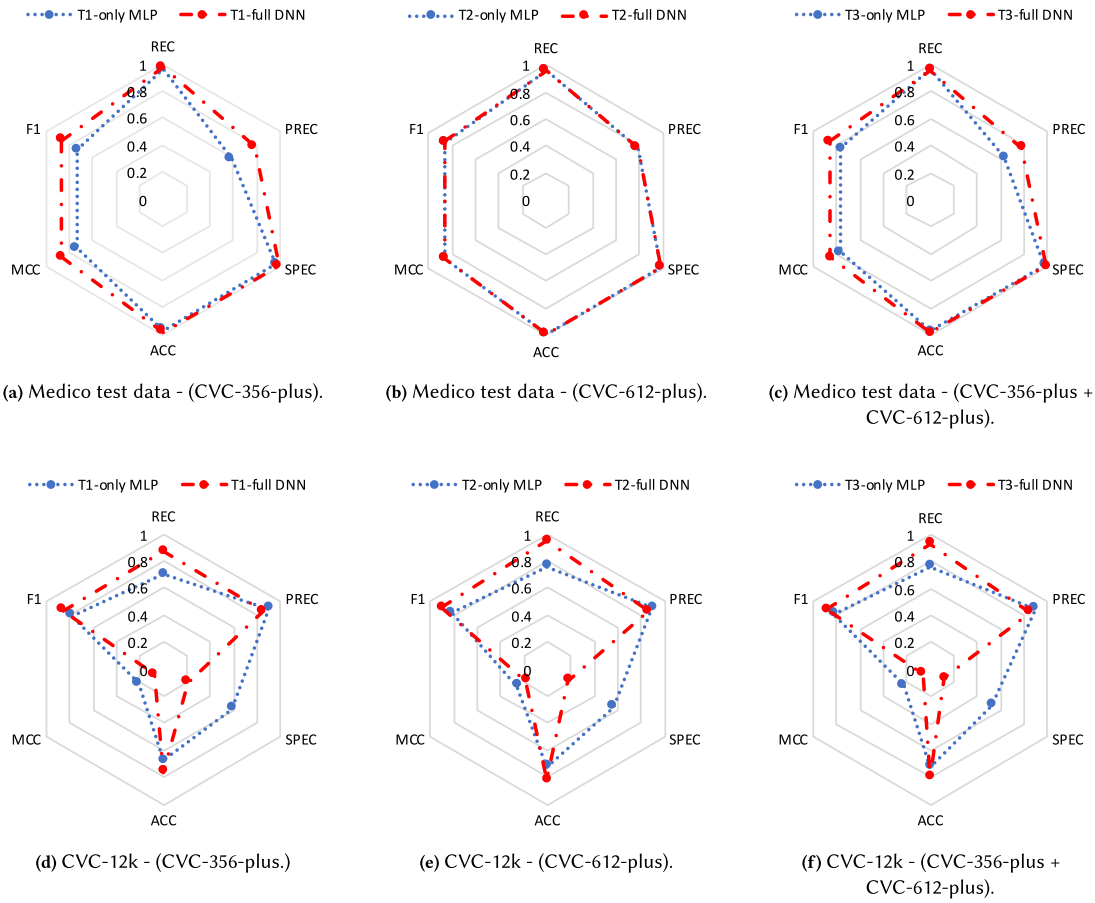
In method 5, the model with the 16 outputs corresponding to T1 has 15 choices to classify non-polyp images. Similarly, the Medico test dataset has more non-polyp images than polyp images. However, the model corresponding to T2 has a 50% chance to classify both polyps and non-polyps. As a result, the model of T1 shows better performance than the model of T2 in Figure 10(i). Because this shows a slight performance change, we cannot see the same difference in ROC and PR curves in Figure 11(a) and (c). In contrast, T2 in the plot in Figure 10(j) shows performance improvement when the model has a 50:50 chance for classifying polyps and non-polyps. This improvement occurred as a result of a large number of polyps in the CVC-12k dataset. The ROC and PR curves in plots in Figure 11(b) and (d) show this performance difference precisely. In other words, the model of T2 has a better chance of classifying polyps compared to the 1/16th chance in the model of T1.

The retrained models corresponding to T3, T4, and T5 do not show considerable performance changes for the Medico test dataset, as we can see from plots in Figure 10(i), (a), and (c). Conversely, the retraining method used in T3, T4, and T5 for the CVC-12k dataset shows large performance changes in the plots in Figure 10(j), (b), and (d). However, these methods show an overall performance loss. More comparisons on these plots are discussed in Section 6.

For the following experiments, we analyzed method 5 even further. The main focus of this analysis is to understand the behavior of the best model for training only the MLP versus training the whole DNN. In this experiment, we collected results for two main test datasets: the Medico test dataset and the CVC-12k dataset. Then, we collected performance measures from the two training mechanisms: training only the MLP and training the whole DNN. Furthermore, results were tabulated in Table 8, and corresponding graphs were depicted in Figure 12 to analyze them.

The first row of Figure 12 shows the differences in the performance of testing with the Medico test data. In the second row, it presents the performance changes for the CVC-12k dataset. The dotted lines in plots in Figure 12 represent training MLP. Similarly, the dashed lines represent training the whole DNN. The three plots of each row represent results of retraining the model with the Medico training data and CVC-356-plus dataset, the Medico training data and CVC-612-plus dataset, and the Medico training data and both CVC-356-plus and CVC-612-plus datasets, respectively.

According to the plots in Figure 12(a) through (c), it is clear that retraining the whole DNN can be used to improve the overall performance of the DNN model because we can see performance improvement in these plots except in Figure 12(b), which shows closely equal performance metrics. However, in test cases with the CVC-12k dataset, it shows a completely new behavior for retraining the whole DNN as depicted in Figure 12(d) through (f). These plots show large changes in the performance hexagons with considerable positive improvements for the



(a) Medico test data - (CVC-356-plus).

(b) Medico test data - (CVC-612-plus).

(c) Medico test data - (CVC-356-plus + CVC-612-plus).

(d) CVC-12k - (CVC-356-plus.)

(e) CVC-12k - (CVC-612-plus).

(f) CVC-12k - (CVC-356-plus + CVC-612-plus).

Fig. 12. Behavior of the complex DNN method (method 5) while training only MLP compared to training the whole DNN. The first row shows the effects for both cases when the test dataset is the Medico test data, and the second row shows the result when the test dataset is the CVC-12k dataset. T1, T2, and T3 represent the training dataset used for the model. (T1, Medico training dataset + CVC-356; T2, Medico training dataset + CVC-612; T3, Medico training dataset + CVC-356 + CVC-612.)

recall and considerable performance loss for the specificity values. This experiment also shows that researchers could be misled by the performance monitoring process of DNN methods using a single dataset. In other words, according to the first row of the figure, researchers may conclude that retraining the whole DNN is a positive factor. However, the results of the second row prove that it is not always true by showing performance losses for the same technique.

The results presented in plots in Figure 12 show difficulties in adapting ML models for cross-dataset generalization with a different perspective. In that experiment, the performance loss in specificity, which is a parameter of reflecting True Negative (TN) detection, shows that method 5 is affected by imbalanced data in the CVC-12k dataset. The main reason for the effect is that the CVC-12k dataset contains a lower percentage of negative images compared to positive ones. This reflects an important factor to take into account when developing generalizable ML models, which is that the ratio of negative and positive findings needs to be taken into account when looking at metrics. Metrics such as MCC are better suited to interpret results. In terms of ROC compared

to the PR curve, the results show that the PR curve reflects the performance of the model more realistically than ROC.

6 DISCUSSION

In this section, we present our findings and point out several important considerations for future research. Our discussion follows the same sequence as our contributions in this article.

In our experiments, combinations of ResNet-152, DenseNet-161, and an additional MLP produced the best result for the Medico 2018 dataset. The reported results from this model for the Medico Task led us to hold second place based on the MCC values calculated by organizers, and there was only a tiny gap of around 0.0029. Furthermore, the winning team of this competition used additional data items that were made by photo-editing tools for the imbalanced classes, such as the out-of-patients class. In contrast to this, our method 5 works well without using manually annotated data items because of the procedure we followed to implement and train that model. The procedure of implementing such a complex model is described step by step in Section 4.3, and anyone can follow these steps to get a well-performing DL model in a classification task.

In addition to the implementation and the procedure used in method 5, the data-filling mechanism used to fill the out-of-patient imbalance class shows impressive performance gain. This method is preferred when one class has a small number of data items in a multi-class classification task. In our work, without annotating more data ourself, which also requires the help of medical experts, we prefer to use random images from the Internet, as described in Section 5.1. This is an efficient way to add more data items without spending more time on manual annotation or creating synthetic data items. The preceding method works because the random images influence the ML models to make a wider range of possibilities to classify images into a particular class.

Dyed-lifted-polyps, dyed-resection-margins, esophagitis, and normal-z-line raised classification conflicts in our best method (method 5). If we could overcome these conflicts, then the model would perform better than the current recorded performance in the 2018 Medico Task. To identify the reasons for these classification conflicts, we manually investigated the images of these classes. If we compare sample images of dyed-lifted-polyps (Figure 1(c)), dyed-resection-margins (Figure 1(d)), esophagitis (Figure 1(e)), and normal-z-line (Figure 1(i)), then we can identify that this conflict was caused as a result of similar texture and shapes of these images. To overcome this problem, researchers can select only the images that made the conflict and train a new DL model to classify them into the correct classes. Then, this model can be added to the model introduced in method 5 using the property of its expandability.

Can we use our best DL model for real systems in hospitals to classify GI findings? Or can we use the state-of-the-art ML classification models introduced by researchers in real applications? Toward answering this question, this article focuses on deep evaluations of the proposed methods as one of the main contributions. Regularly, researchers present the performance of their classification models using only a test dataset, which was reserved from the dataset used to produce the training data. In addition, they measure the performance by selecting only a few measurements out of the REC, PREC, SPEC, ACC, MCC, and F1. However, we emphasize the requirements of an in-depth analysis of all of these six parameters at once to identify the real performance of ML models. Several of the works listed in Table 1 do not use this methodology as part of their evaluations. This makes it difficult to reason about the real-world performance of the proposed methods and how they compare with other methods. In this article, we also consider the importance of evaluating ML models with cross-datasets.

Why do we need cross-dataset evaluations? To explain this requirement, we consider the research work done by Wang et al. [75]. They presented an area under the ROC curve of 0.984 and a per-image sensitivity of 94.38 for polyp detection. In our first look, these results show a good DL model. Similarly, our results in Figure 10(i) and 11(a) and (c) reflect the same impression in the first look because it shows excellent performance as a DL model. However, after analyzing cross-dataset performance for polyp detection with a completely new dataset like CVC-12k, we recognized that performance gain is not enough for applying it in real applications. Therefore, from this work, we emphasize that researchers want to consider cross-dataset evaluations thoroughly before applying

their solutions in real-world applications. Otherwise, the selection bias, the capture bias, and the category bias (label bias) problems may appear in the results. Then, we may end up with the wrong conclusion about research works. All of these facts imply that more research must be performed to improve the generalizability along with the performance improvement on a single dataset or single data source.

7 CONCLUSION

We studied cross-dataset bias and evaluation metrics interpretation in ML using five methods and four different datasets within the field of GI endoscopy as respective use case. In particular, we performed an extensive study of ML models in the context of medical applications based on a use case of GI tract abnormality classification across different datasets. The main conclusion and resulting recommendation is that a multi-center or cross-dataset evaluation is important, if not essential, for ML models in the medical field to obtain a realistic understanding of the performance of such models in real-world settings.

We found that the combination of DNN ResNet-152 and DenseNet-161 with an additional MLP performed best on both the validation and test datasets. This model shows that a combination of multiple pre-trained DNN models can have better capabilities to classify images into the correct classes because of their cumulative decision-making capabilities. We also proposed an evaluation method using six measures: REC, PREC, SPEC, ACC, MCC, and F1. Moreover, we suggest that these measures should be presented all at once using hexagon plots that convey a complete view of real performance. It is our hope that these tools can enable a more realistic evaluation and comparison of ML methods.

Furthermore, we presented cross-dataset evaluations to identify the generalizability of our ML models, emphasizing the fact that achieving high scores for evaluation metrics does not always represent the real performance of ML models and should be interpreted with care. By evaluating the ML models with cross-datasets experiments, we showed the complexity of understanding the real functional performance of the models. The state-of-the-art research works that perform classification cannot be used in practical applications because of their lack of generalizability. Based on the experimental results, we conclude that researchers should focus on implementing and researching generalizable ML models with cross-dataset evaluations. Rather than presenting metrics calculated from a simple training and testing split of the data, we suggest to always rely on cross-dataset evaluation to obtain a real-world representative indication of model performance. This is especially important in a medical context because one has to make sure that the obtained models are reliable and not just perform well on a specific dataset.

Finally, we want to point out that the lack of generalization, as evidenced by the poor result for cross-dataset evaluation presented in this article, rises a very important question: in the context of cross-dataset or multi-center studies, is it really possible to have generalizable ML models? This is something that we ourselves plan to investigate further in future work, and it is ou hope that other researchers in computer science and medicine will do the same or at least have the question in their mind when performing similar studies.

ACKNOWLEDGMENTS

We would like to thank the reviewers for their contributions to the article.

REFERENCES

[1] Taruna Agrawal, Rahul Gupta, Saurabh Sahu, and Carol Y. Espy-Wilson. 2017. SCL-UMD at the Medico Task-MediaEval 2017: Transfer learning based classification of medical images. In *Proceedings of MediaEval 2017*.

[2] Luís A. Alexandre, Nuno Nobre, and João Casteleiro. 2008. Color and position versus texture features for endoscopic polyp detection. In *Proceedings of IEEE BMEI2008*, Vol. 2. 38–42.

[3] Stefan Ameling, Stephan Wirth, Dietrich Paulus, Gerard Lacey, and Fernando Vilarino. 2009. Texture-based polyp detection in colonoscopy. In *Bildverarbeitung für die Medizin 2009*. 346–350.

[4] Quentin Angermann, Jorge Bernal, Cristina Sánchez-Montes, Maroua Hammami, Gloria Fernández-Esparrach, Xavier Dray, Olivier Romain, F. Javier Sánchez, and Aymeric Histace. 2017. Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis. In *Proceedings of CARE and CLIP 2017*. 29–41.

[5] Jorge Bernal, Aymeric Histace, Marc Masana, Quentin Angermann, Cristina Sánchez-Montes, Cristina Rodriguez, Maroua Hammami, et al. 2018. Polyp detection benchmark in colonoscopy videos using GTCreator: A novel fully configurable tool for easy and fast annotation of image databases. In *Proceedings of CARS 2018*.

[6] Jorge Bernal, F. Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics* 43 (2015), 99–111.

[7] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. 2012. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition* 45, 9 (2012), 3166–3182.

[8] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. 2013. Impact of image preprocessing methods on polyp localization in colonoscopy frames. In *Proceedings of IEEE EMBC 2013*. 7350–7354.

[9] Jorge Bernal, Nima Tajkbaksh, Francisco Javier Sánchez, Bogdan J. Matuszewski, Hao Chen, Lequan Yu, Quentin Angermann, et al. 2017. Comparative validation of polyp detection methods in video colonoscopy: Results from the MICCAI 2015 Endoscopic Vision Challenge. *IEEE Transactions on Medical Imaging* 36, 6 (2017), 1231–1249.

[10] Rune Johan Borgli, Pål Halvorsen, Michael Riegler, and Håkon Kvale Stensland. 2018. Automatic hyperparameter optimization in Keras for the MediaEval 2018 Medico Multimedia Task. In *Proceedings of MediaEval 2018*.

[11] Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT 2010*. 177–186.

[12] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* 68, 6 (2018), 394–424.

[13] Da-Chuan Cheng, Wen-Chien Ting, Yung-Fu Chen, and Xiaoyi Jiang. 2011. Automatic detection of colorectal polyps in static images. *Biomedical Engineering: Applications, Basis and Communications* 23, 05 (2011), 357–367.

[14] Torch Contributors. 2018. Torchvision Models. Retrieved May 7, 2020 from https://pytorch.org/docs/stable/torchvision/models.html.

[15] Pieter-Tjerk de Boer, Dirk P. Kroese, Shie Mannor, and Reuven Y. Rubinstein. 2005. A tutorial on the cross-entropy method. *Annals of Operations Research* 134 (2005), 19–67.

[16] Thomas de Lange, Pål Halvorsen, and Michael Riegler. 2018. Methodology to develop machine learning algorithms to improve performance in gastrointestinal endoscopy. *World Journal of Gastroenterology* 24, 45 (2018), 5057.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of IEEE CVPR 2009*. 248–255.

[18] Danielle Dias and Ulisses Dias. 2018. Transfer learning with CNN architectures for classifying gastrointestinal diseases and anatomical landmarks. In *Proceedings of MediaEval 2018*.

[19] Patrick Doetsch, Christian Buck, Pavlo Golik, Niklas Hoppe, Michael Kramp, Johannes Laudenberg, Christian Oberdörfer, Pascal Steingrube, Jens Forster, and Arne Mauser. 2009. Logistic model trees with AUC split criterion for the KDD Cup 2009 Small Challenge. In *Proceedings of KDD-Cup '09*. 77–88.

[20] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. 2009. Pedestrian detection: A benchmark. In *Proceedings of IEEE CVPR 2009*.

[21] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2000. Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *Annals of Statistics* 28, 2 (2000), 337–407.

[22] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter* 11, 1 (2009), 10–18.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of IEEE CVPR 2016*. 770–778.

[24] Steven A. Hicks, Pia H. Smedsrud, Pål Halvorsen, and Michael Riegler. 2018. Deep learning based disease detection using domain specific transfer learning. In *Proceedings of MediaEval 2018*.

[25] Trung-Hieu Hoang, Hai-Dang Nguyen, and Thanh-An Nguyen. 2018. An application of residual network and faster - RCNN for Medico: Multimedia Task at MediaEval 2018. In *Proceedings of MediaEval 2018*.

[26] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of IEEE CVPR 2017*. 2261–2269.

[27] Sae Hwang, JungHwan Oh, Wallapak Tavanapong, Johnny Wong, and Piet C. De Groen. 2007. Polyp detection in colonoscopy video using elliptical shape feature. In *Proceedings of IEEE ICIP 2007*, Vol. 2. 465–468.

[28] Dimitrios K. Iakovidis, Dimitrios E. Maroulis, Stavros A. Karkanis, and A. Brokos. 2005. A comparative study of texture features for the discrimination of gastric polyps in endoscopic video. In *Proceedings of IEEE CBMS 2005*. 575–580.

[29] Yuji Iwahori, Takayuki Shinohara, Akira Hattori, Robert J. Woodham, Shinji Fukui, Manas Kamal Bhuyan, and Kunio Kasugai. 2013. Automatic polyp detection in endoscope images using a Hessian filter. In *Proceedings of MVA 2013*, Vol. 13. 21–24.

[30] Debesh Jha, Pia Smedsrud, Michael Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard Johansen. 2020. Kvasir-SEG: A segmented polyp dataset. In *Proceedings of MMM 2020*. 1–12.

[31] Xiao Jia and Max Q.-H. Meng. 2017. Gastrointestinal bleeding detection in wireless capsule endoscopy images using handcrafted and CNN features. In *Proceedings of IEEE EMBC 2017*. 3154–3157.

[32] Stavros A. Karkanis, Dimitrios K. Iakovidis, Dimitrios E. Maroulis, Dimitris A. Karras, and M. Tzivras. 2003. Computer-aided tumor detection in endoscopic video using color wavelet features. *IEEE Transactions on Information Technology in Biomedicine* 7, 3 (2003), 141–152.

[33] Zeshan Khan and Muhammad Atif Tahir. 2018. Majority voting of heterogeneous classifiers for finding abnormalities in the gastro-intestinal tract. In *Proceedings of MediaEval 2018.*

[34] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei Efros, and Antonio Torralba. 2012. Undoing the damage of dataset bias. In *Proceedings of ECCV 2012.*

[35] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv:1412.6980.

[36] Mathias Kirkerød, Vajira Thambawita, Michael Riegler, and Pål Halvorsen. 2018. Using preprocessing as a tool in medical image detection. In *Proceedings of MediaEval 2018.*

[37] Tobey H. Ko, Zhonglei Gu, and Yang Liu. 2018. Weighted discriminant embedding: Discriminant subspace learning for imbalanced medical data classification. In *Proceedings of MediaEval 2018.*

[38] Niels Landwehr, Mark Hall, and Eibe Frank. 2005. Logistic model trees. *Machine Learning* 59, 1–2 (2005), 161–205.

[39] A. M. Leufkens, M. G. H. van Oijen, F. P. Vleggaar, and P. D. Siersema. 2012. Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. *Endoscopy* 44, 05 (2012), 470–475.

[40] Baopu Li and Max Q.-H. Meng. 2012. Tumor recognition in wireless capsule endoscopy images using textural features and SVM-based feature selection. *IEEE Transactions on Information Technology in Biomedicine* 16, 3 (2012), 323–329.

[41] David Lieberman. 2005. Quality and colonoscopy: A new imperative. *Gastrointestinal Endoscopy* 61, 3 (2005), 392–394.

[42] Mathias Lux, Michael Riegler, Pål Halvorsen, Konstantin Pogorelov, and Nektarios Anagnostopoulos. 2016. LIRE: Open source visual information retrieval. In *Proceedings of ACM MMSys 2016.* 30.

[43] Alexander V. Mamonov, Isabel N. Figueiredo, Pedro N. Figueiredo, and Yen-Hsi Richard Tsai. 2014. Automated polyp detection in colon capsule endoscopy. *IEEE Transactions on Medical Imaging* 33, 7 (2014), 1488–1502.

[44] Yuichi Mori and Shin-Ei Kudo. 2018. Detecting colorectal polyps via machine learning. *Nature Biomedical Engineering* 2, 10 (2018), 713.

[45] Syed Sadiq Ali Naqvi, Shees Nadeem, Muhammad Zaid, and Muhammad Atif Tahir. 2017. Ensemble of texture features for finding abnormalities in the gastro-intestinal tract. In *Proceedings of MediaEval 2017.*

[46] Olga Ostroukhova, Konstantin Pogorelov, Michael Riegler, Duc-Tien Dang-Nguyen, and Pål Halvorsen. 2018. Transfer learning with prioritized classification and training dataset equalization for medical objects detection. In *Proceedings of MediaEval 2018.*

[47] Sun Young Park, Dustin Sargent, Inbar Spofford, Kirby G. Vosburgh, and Y. A-Rahim. 2012. A colon video analysis framework for polyp detection. *IEEE Transactions on Biomedical Engineering* 59, 5 (2012), 1408.

[48] Stefan Petscharnig, Klaus Schöffmann, and Mathias Lux. 2017. An inception-like CNN architecture for GI disease and anatomical landmark classification. In *Proceedings of MediaEval 2017.*

[49] Konstantin Pogorelov, Olga Ostroukhova, Mattis Jeppsson, Håvard Espeland, Carsten Griwodz, Thomas de Lange, Dag Johansen, Michael Riegler, and Pål Halvorsen. 2018. Deep learning and hand-crafted feature based approaches for polyp detection in medical videos. In *Proceedings of IEEE CBMS 2018.* 381–386.

[50] Konstantin Pogorelov, Kristin Ranheim Randel, Thomas de Lange, Sigrun Losada Eskeland, Carsten Griwodz, Dag Johansen, Concetto Spampinato, et al. 2017. Nerthus: A bowel preparation quality video dataset. In *Proceedings of ACM MMSys 2017.* 170–174.

[51] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, et al. 2017. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of ACM MMSys 2017.* 164–169.

[52] Konstantin Pogorelov, Michael Riegler, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Carsten Griwodz, Peter Thelin Schmidt, and Pål Halvorsen. 2017. Efficient disease detection in gastrointestinal videos—Global features versus neural networks. *International Journal of Multimedia Tools and Applications* 76, 21 (2017), 22493–22525.

[53] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Thomas De Lange, Kristin Ranheim Randel, Duc-Tien Dang-Nguyen, Mathias Lux, and Olga Ostroukhova. 2018. Medico multimedia task at MediaEval 2018. In *Proceedings of MediaEval 2018.*

[54] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Carsten Griwodz, Thomas de Lange, Kristin Randel, Sigrun Eskeland, Dang Nguyen, Duc Tien, and Olga Ostroukhova. 2017. A comparison of deep learning with global features for gastrointestinal disease detection. In *Proceedings of MediaEval 2017.*

[55] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Thomas De Lange, Kristin Ranheim Randel, Duc-Tien Dang-Nguyen, Mathias Lux, and Olga Ostroukhova. 2018. Medico multimedia task at MediaEval 2018. In *Proceedings of MediaEval 2018.*

[56] Michael Riegler, Martha Larson, Mathias Lux, and Christoph Kofler. 2014. How 'how' reflects what's what: Content-based exploitation of how users frame social images. In *Proceedings of ACM MM 2014.* 397–406.

[57] Michael Riegler, Mathias Lux, Carsten Griwodz, Concetto Spampinato, Thomas de Lange, Sigrun L. Eskeland, Konstantin Pogorelov, et al. 2016. Multimedia and medicine: Teammates for better disease detection and survival. In *Proceedings of ACM MM 2016.* 968–977.

[58] Michael Riegler, Konstantin Pogorelov, Sigrun Losada Eskeland, Peter Thelin Schmidt, Zeno Albisser, Dag Johansen, Carsten Griwodz, Pål Halvorsen, and Thomas De Lange. 2017. From annotation to computer-aided diagnosis: Detailed evaluation of a medical multimedia system. *ACM Transactions on Multimedia Computing, Communications, and Applications* 13, 3 (2017), 26.

[59] Michael Riegler, Konstantin Pogorelov, Pål Halvorsen, Kristin Randel, Sigrun Losada Eskeland, Duc-Tien Dang-Nguyen, Mathias Lux, Carsten Griwodz, Concetto Spampinato, and Thomas Lange. 2017. Multimedia for medicine: The Medico Task at MediaEval 2017. In *Proceedings of MediaEval 2017*.
[60] Sebastian Ruder. 2016. An overview of gradient descent optimization algorithms. arXiv:1609.04747.
[61] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, et al. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision* 115 (2015), 211–252.
[62] Steven L. Salzberg. 1994. C4. 5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning* 16, 3 (1994), 235–240.
[63] Younghak Shin and Ilangko Balasingham. 2017. Comparison of hand-craft feature based SVM and CNN based deep learning framework for automatic polyp classification. In *Proceedings of IEEE EMBC 2017*. 3277–3280.
[64] Michael Steiner, Mathias Lux, and Pål Halvorsen. 2018. The 2018 Medico Multimedia Task submission of Team NOAT using neural network features and search-based classification. In *Proceedings of MediaEval 2018*.
[65] Marc Sumner, Eibe Frank, and Mark Hall. 2005. Speeding up logistic model tree induction. In *Proceedings of PKDD 2005*. 675–683.
[66] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. 2013. On the importance of initialization and momentum in deep learning. In *Proceedings of ICML 2013*. 1139–1147.
[67] Nima Tajbakhsh, Suryakanth R. Gurudu, and Jianming Liang. 2015. Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks. In *Proceedings of IEEE ISBI 2015*. 79–83.
[68] Nima Tajbakhsh, Suryakanth R. Gurudu, and Jianming Liang. 2016. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Transactions on Medical Imaging* 35, 2 (2016), 630–644.
[69] Nima Tajbakhsh, Jae Y. Shin, Suryakanth R. Gurudu, R. Todd Hurst, Christopher B. Kendall, Michael B. Gotway, and Jianming Liang. 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging* 35, 5 (2016), 1299–1312.
[70] Mario Taschwer, Manfred Jürgen Primus, Klaus Schoeffmann, and Oge Marques. 2018. Early and late fusion of classifiers for the MediaEval Medico Task. In *Proceedings of MediaEval 2018*.
[71] Vajira Thambawita, Debesh Jha, Michael Riegler, Pål Halvorsen, Hugo Lewi Hammer, Håvard D. Johansen, and Dag Johansen. 2018. The Medico-Task 2018: Disease detection in the gastrointestinal tract using global features and deep learning. In *Proceedings of MediaEval 2018*.
[72] A. Torralba and A. A. Efros. 2011. Unbiased look at dataset bias. In *Proceedings of IEEE CVPR 2011*. 1521–1528.
[73] David Vázquez, Jorge Bernal, F. Javier Sánchez, Gloria Fernández-Esparrach, Antonio M. López, Adriana Romero, Michal Drozdzal, and Aaron C. Courville. 2017. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of Healthcare Engineering* 2017 (2017), 4037190.
[74] L. Von Karsa, J. Patnick, and N. Segnan. 2012. European guidelines for quality assurance in colorectal cancer screening and diagnosis. First edition—Executive summary. *Endoscopy* 44, Suppl. 3 (2012), SE1–SE8.
[75] Pu Wang, Xiao Xiao, Jeremy R. Glissen Brown, Tyler M. Berzin, Mengtian Tu, Fei Xiong, Xiao Hu, et al. 2018. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nature Biomedical Engineering* 2, 10 (2018), 741.
[76] Yi Wang, Wallapak Tavanapong, Johnny Wong, JungHwan Oh, and Piet C. De Groen. 2014. Part-based multiderivative edge cross-sectional profiles for polyp detection in colonoscopy. *IEEE Journal of Biomedical and Health Informatics* 18, 4 (2014), 1379–1389.
[77] Yi Wang, Wallapak Tavanapong, Johnny Wong, Jung Hwan Oh, and Piet C. De Groen. 2015. Polyp-Alert: Near real-time feedback during colonoscopy. *International Journal of Computer Methods and Programs in Biomedicine* 120, 3 (2015), 164–179.
[78] Lequan Yu, Hao Chen, Qi Dou, Jing Qin, and Pheng Ann Heng. 2017. Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. *IEEE Journal of Biomedical and Health Informatics* 21, 1 (2017), 65–75.
[79] Yixuan Yuan, Dengwang Li, and Max Q.-H. Meng. 2018. Automatic polyp detection via a novel unified bottom-up and top-down saliency approach. *IEEE Journal of Biomedical and Health Informatics* 22, 4 (2018), 1250–1260.
[80] Matthew D. Zeiler. 2012. ADADELTA: An adaptive learning rate method. arXiv:1212.5701.
[81] Xu Zhang, Fei Chen, Tao Yu, Jiye An, Zhengxing Huang, Jiquan Liu, Weiling Hu, Liangjing Wang, Huilong Duan, and Jianmin Si. 2019. Real-time gastric polyp detection using convolutional neural networks. *PloS One* 14, 3 (2019), e0214133.
[82] Xu Zhang, Weiling Hu, Fei Chen, Jiquan Liu, Yuanhang Yang, Liangjing Wang, Huilong Duan, and Jianmin Si. 2017. Gastric precancerous diseases classification using CNN with a concise model. *PloS One* 12, 9 (2017), e0185508.
[83] Mingda Zhou, Guanqun Bao, Yishuang Geng, Bader Alkandari, and Xiaoxi Li. 2014. Polyp detection and radius measurement in small intestine using video capsule endoscopy. In *Proceedings of IEEE BMEI2014*. 237–241.

Received March 2019; revised December 2019; accepted February 2020

A.15 Paper XV : DDANet: Dual Decoder Attention Network for Automatic Polyp Segmentation

Authors: N. K. Tomar, D. Jha, S. Ali, H. D. Johansen, D. Johansen, M. A. Riegler, and P. Halvorsen

Abstract: Colonoscopy is the gold standard for examination and detection of colorectal polyps. Localization and delineation of polyps can play a vital role in treatment (e.g., surgical planning) and prognostic decision making. Polyp segmentation can provide detailed boundary information for clinical analysis. Convolutional neural networks have improved the performance in colonoscopy. However, polyps usually possess various challenges, such as intra-and inter-class variation and noise. While manual labeling for polyp assessment requires time from experts and is prone to human error (e.g., missed lesions), an automated, accurate, and fast segmentation can improve the quality of delineated lesion boundaries and reduce missed rate. The Endotect challenge provides an opportunity to benchmark computer vision methods by training on the publicly available Hyperkvasir and testing on a separate unseen dataset. In this paper, we propose a novel architecture called "DDANet" based on a dual decoder attention network. Our experiments demonstrate that the model trained on the Kvasir-SEG dataset and tested on an unseen dataset achieves a dice coefficient of 0.7874, mIoU of 0.7010, recall of 0.7987, and a precision of 0.8577, demonstrating the generalization ability of our model.

Published: Proceedings of Pattern Recognition (ICPR Workshop and Challenges)

Candidate contributions: D. Jha assisted in the conceptualization and design of the work. He contributed to the manuscript preparation, provided critical insight, and revised the manuscript subsequently with input from all the co-authors.

Thesis objectives: Objective III

DDANet: Dual Decoder Attention Network for Automatic Polyp Segmentation

Nikhil Kumar Tomar[1], Debesh Jha[1,3], Sharib Ali[4], Håvard D. Johansen[3], Dag Johansen[3], Michael A. Riegler[1], and Pål Halvorsen[1,2]

[1] SimulaMet, Norway
[2] Oslo Metropolitan University, Norway
[3] UIT The Arctic University of Norway
[4] Department of Engineering Science, University of Oxford, Oxford, UK
debesh@simula.no

Abstract. Colonoscopy is the gold standard for examination and detection of colorectal polyps. Localization and delineation of polyps can play a vital role in treatment (e.g., surgical planning) and prognostic decision making. Polyp segmentation can provide detailed boundary information for clinical analysis. Convolutional neural networks have improved the performance in colonoscopy. However, polyps usually possess various challenges, such as intra-and inter-class variation and noise. While manual labeling for polyp assessment requires time from experts and is prone to human error (e.g., missed lesions), an automated, accurate, and fast segmentation can improve the quality of delineated lesion boundaries and reduce missed rate. The Endotect challenge provides an opportunity to benchmark computer vision methods by training on the publicly available Hyperkvasir and testing on a separate unseen dataset. In this paper, we propose a novel architecture called "DDANet" based on a dual decoder attention network. Our experiments demonstrate that the model trained on the Kvasir-SEG dataset and tested on an unseen dataset achieves a dice coefficient of 0.7874, mIoU of 0.7010, recall of 0.7987, and a precision of 0.8577, demonstrating the generalization ability of our model.

Keywords: Polyp segmentation, Deep Learning, Convolutional neural network, Benchmarking

1 Introduction

Colorectal cancer is one of the leading causes of cancer. Colonoscopy is a standard medical procedure for the surveillance examination and treatment. Regular screening and removal of pre-cancerous lesions through colonoscopy is essential for early cancer detection and prevention. Studies suggest that the miss-rate of adenoma is between 6 to 27% [1].

The automatic segmentation of the suspected areas with lesions in colonoscopy images can play a crucial role, and identifying each colon pixel can significantly impact clinical settings. With the increase of publicly available datasets, dominant methodology such as convolutional neural network, improved hardware,

and collaboration between computational and clinical communities to tackle the problems in endoscopic imaging through computer vision tasks is gaining momentum than ever before. An automatic polyp detection or surveillance system can help to achieve low-cost design solutions and save time of clinicians allowing them to use their time to look into more severe cases.

In this respect, the Endotect challenge [8] offers three tasks, namely, detection of Gastrointestinal (GI) tract images, efficient detection on the same images, and automatic polyp segmentation. The detection and efficient detection task are based on the HyperKvasir dataset [5], and the segmentation is based on the Kvasir-SEG dataset [12]. Out of these three tasks, we participated in the "segmentation task", where the goal was to generate an automatic segmentation of the polyps for the unseen dataset.

In this paper, we propose a novel deep learning architecture, called Dual Decoder Attention Network (DDANet), for automatic polyp segmentation. It follows an encoder-decoder scheme and incorporates a single encoder that is shared by two parallel decoders, where the first decoder acts as a segmentation network and the second decoder acts as an autoencoder network. The autoencoder network helps to strengthen the feature maps in the encoder network. It is used as an auxiliary task training, which is used to generate an attention map. This attention map is used in each decoder to improve the semantic representation of the feature maps. This, in turn, helps to improve the performance of the entire network. The proposed DDANet is fed with an RGB input image,where it predicts the segmentation mask and the reconstructed grayscale image. The architecture is efficient in terms of Frame per Second (FPS) and also has a decent evaluation score. These metrics are the requirement for the real-world settings toward developing a Computer Aided Diagnosis (CADx) system.

2 Related Work

Automatic polyp segmentation task is a well-defined computer vision problem. Recently, there have been several competitions [4, 3, 2, 10] and individual efforts [7, 13, 11, 6, 9] toward building a CADx system for the polyp segmentation. With these competitions and individual efforts, polyp segmentation is becoming more and more mature. However, comparing models and results of the many individual approaches is difficult due to the use of diverse (often publicly non-available) datasets and different hardware. In this respect, competitions provide an opportunity to benchmark and compare the designed methods with other competitors' on the same dataset. Moreover, the evaluation metrics are independently calculated by the organizers, including the ranking decision of each team.

The competitions can help us to define the strengths and weaknesses of each method. It also provides us with an opportunity to disseminate methods and discuss the results collectively in the same space. Through this year's Endotect challenge, we provide a novel solution to develop more efficient algorithms that can be useful to build an automatic polyp segmentation system. Our architecture is composed of an autoencoder branch in addition to the segmentation

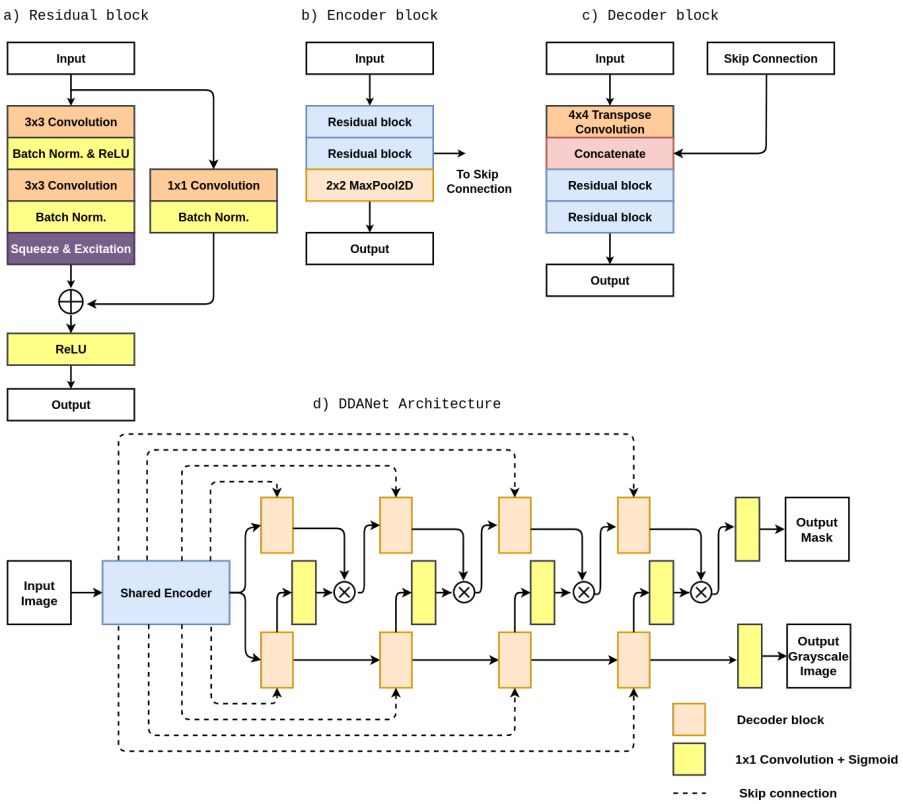


Fig. 1: DDANet architecture and its components.

branch, which is different from other encoder-decoder based network (for example, UNet [14], ResUNet++ [13], DoubleUNet [11]). The benefit of incorporating autoencoder in the network can be seen from the quantitative and qualitative results.

3 DDANet

In this section, we will first describe each component of our DDANet and then detail the overall proposed DDANet architecture.

3.1 Residual block

As the network depth increases, the performance also increases to a certain limit as the gradients can be effectively calculated. However, after a certain depth, the performance of the model may be impacted due to the vanishing or exploding

gradients as the gradients become either zero or too large. By introducing a skip-connection in residual learning, the problem of the vanishing or exploding gradients has been solved. Our residual block (see Figure 1a) consists of two 3×3 convolutions, each followed by a batch normalization and a Rectified linear unit (ReLU) activation function. The residual learning introduces a shortcut connection or identity mapping, which connects the input with the residual block's output. The identity mapping tries to learn an identity function since the input is directly passed to the output. It also helps in a better flow of the gradients during the backpropagation.

3.2 Squeeze and Excitation block

A Convolutional Neural Network (CNN) is used to extract features from an image and then transform the image into a feature map. A problem with CNNs is that they treat every feature channel as equally important. To overcome this problem, we introduce a squeeze and excitation layer, which acts as a channel-wise attention mechanism. It re-weights every feature channel accordingly to create a more accurate feature map. In this way, the overall network becomes more sensitive towards essential features that improve the network performance significantly. The squeeze and excitation network mainly consists of two steps. In the first step, the feature maps are compressed using the global average pooling function to generate a compressed representation for the feature maps. While, in the second step, a 2-layered neural network is used, where features are first reduced and then expanded. This generates a feature vector, which is used to scale the feature channels.

3.3 The DDANet architecture

The proposed architecture named DDANet follows an encoder-decoder design similar to ResUNet++ [13]. The DDANet combines the strength of the residual learning and the squeeze and excitation network. The proposed DDANet is a fully convolutional network that consists of a single encoder shared by dual decoders. The encoder network consists of a 4 encoder block, whereas each decoder network also consists of 4 decoder block (see Figure 1d).

The RGB input image is first fed into the encoder network (see Figure 1b), which encodes it into an abstract feature representation while gradually down-sampling it. The output of the encoder network is fed to both decoders (see Figure 1c), where it is followed by a 4×4 transpose convolution that doubles its spatial dimensions. After that, the image is concatenated with an appropriate feature maps from the encoder network using the skip connection. These skip connections fetch the features from earlier layers at their original resolution, which increases their feature representation strength. The skip connections also act as an alternative path for the gradient flow and are often beneficial for model convergence.

Two residual blocks are then used to learn the necessary feature required by the network during back-propagation. The output of the second decoder

block (autoencoder branch) follows a 1×1 convolution and a sigmoid activation function to generates an attention map. This attention map is multiplied by the output of the first decoder block (segmentation branch), which acts as an input for the next decoder block in the segmentation branch. The final decoder block's output is passed through a 1×1 convolution and a sigmoid activation function, where the first decoder outputs a segmentation mask, and the second outputs the reconstructed grayscale image.

4 Experimental Setup

In this section, we present the implementation details and datasets used in this work.

4.1 Implementation Details

The proposed DDANet architecture is implemented in the PyTorch 1.6 framework[1]. For training the DDANet, we used an NVIDIA DGX-2 machine that uses an Nvidia V100 Tensor Core GPUs.During training, we have used an input image resolution of 512×512. We use a combination of binary cross-entropy and dice loss for calculating the loss between the predicted masks and the ground-truth masks. We have used binary cross-entropy in the case of predicting the grayscale image. An Adam optimizer was used with a learning rate of $1e^{-4}$. The models were trained for 200 epochs.

4.2 Datasets

The Kvasir-SEG [12] dataset was used for training the model. We have used 88% of the dataset for training and the remaining 12% images for development-test-set. Kvasir-SEG consists of 1000 polyp images, ground truth segmentation masks, and bounding boxes. A separate test dataset with 200 images was provided for prediction. However, the ground truth for this dataset was not provided by the organizers. The exact number of images used for the training and testing can also be found in our GitHub repository. More details about the dataset and the baseline results on it can be found in [12].

5 Results

Table 1 shows the results of the DDANet trained and validated on Kvasir-SEG. Additionally, evaluation scores on the test dataset can also be found here. The evaluation metrics for the challenge was Dice Coefficient (DSC). However, we have also calculated other commonly used metrics such as mean Intersection over Union (mIoU), recall, precision, and FPS. The DDANet obtained a DSC of 0.8576, a mIoU of 0.7800, a recall of 0.8880, and a precision of 0.8643. All the metrics suggest that our method performs quite well on the Kvasir-SEG dataset. When we compare the results with our previous results [13, 12], where

[1] `https://github.com/nikhilroxtomar/DDANet`

Table 1: Quantitative results on Kvasir-SEG and unseen (Challenge) dataset.

Dataset	**Method**	**DSC**	**mIOU**	**Recall**	**Precision**	**FPS**
Kvasir-SEG	DDANet	0.8576	0.7800	0.8880	0.8643	69.59
Unseen (**Challenge**)	DDANet	0.7874	0.7010	0.7987	0.8577	70.23

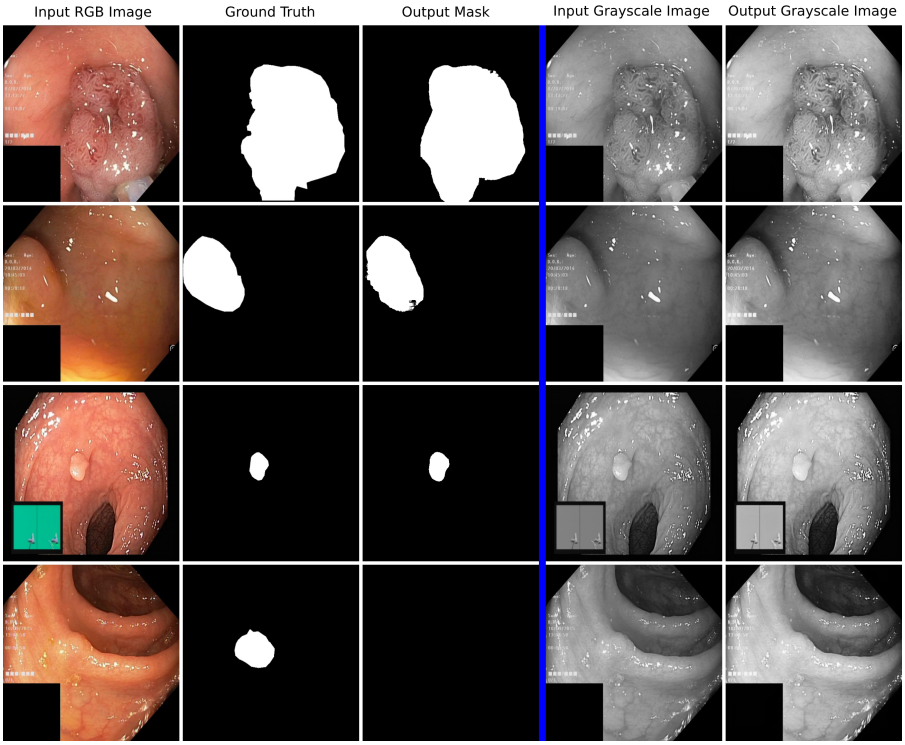


Fig. 2: Qualitative results of the DDANet on the Kvasir-SEG test dataset. The blue line divides the segmentation and the reconstruction part. Columns 4 and 5 show the reconstruction part that was used in the DDANet as an auxiliary task.

the DSC values were 0.8133, and 0.7877, DDANet achieves a higher DSC of 0.8576. However, we can not compare directly with this work with our previous works as a different train-test split of the dataset is used.

Figure 2 shows the qualitative results of the DDANet on Kvasir-SEG. The figure shows that the proposed DDANet is able to segment both larger and smaller polyps. However, the figure also shows the challenges in identifying the flat polyps, which is one of the open issues in the field of development of CADx systems for colonoscopy. From the quantitative results on development and un-

seen test dataset, we can say that the proposed method is comprehensive in producing reliable segmentation output.

6 Discussion

The qualitative results (see Figure 2) show that the proposed model was able to segment polyps ranging from large to small (Figure 2), but still, challenges remain within some polyps (for example, flat or sessile). We can also see a nearly perfect reconstruction of the grayscale image. In the future, we would like to use image super-resolution instead of just a grayscale image reconstruction.

From all the results, we can see that our method achieves high precision and recall evaluation scores on both the Kvasir-SEG validation dataset and on the unseen test dataset (see Table 1). Additionally, we also achieved a DSC of 0.7874 on the unseen dataset. Thus, high DSC, recall, and precision results validate our proposed method. Moreover, our approach is quite fast with an average FPS of 70.23. Thus, the results show that our method can identify polyps in real-time.

7 Conclusion

The Endotect challenge [8] aims to benchmark various computer-vision approaches on the HyperKvasir dataset containing GI images and videos. Here, we have proposed the DDANet architecture for automatic polyp segmentation, and the proposed architecture provides good results in the segmentation task. We have obtained a high precision, recall, DSC, mIoU, and FPS. However, there are large rooms for improvements. We intend to further improve the architecture by applying post-processing and analyzing the optimal parameters in the future.

Acknowledgements

This work is funded in part by the Research Council of Norway, project number 263248 (Privaton) and project number 282315 (AutoCap). We performed all computations in this paper on equipment provided by the Experimental Infrastructure for Exploration of Exascale Computing (eX^3), which is financially supported by the Research Council of Norway under contract 270053.

References

[1] Ahn, S.B., Han, D.S., Bae, J.H., Byun, T.J., Kim, J.P., Eun, C.S.: The miss rate for colorectal adenoma determined by quality-adjusted, back-to-back colonoscopies. Gut and liver 6(1), 64 (2012)

[2] Ali, S., Dmitrieva, M., Ghatwary, N., Bano, S., Polat, G., Temizel, A., Krenzer, A., Hekalo, A., Guo, Y.B., Matuszewski, B., et al.: A translational pathway of deep learning methods in gastrointestinal endoscopy. arXiv preprint arXiv:2010.06034 (2020)

[3] Ali, S., Zhou, F., Braden, B., Bailey, A., Yang, S., Cheng, G., Zhang, P., Li, X., Kayser, M., Soberanis-Mukul, R.D., et al.: An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. Scientific reports 10(1), 1–15 (2020)

[4] Bernal, J., Tajkbaksh, N., Sánchez, F.J., Matuszewski, B.J., Chen, H., Yu, L., Angermann, Q., Romain, O., Rustad, B., Balasingham, I., et al.: Comparative Validation of Polyp Detection Methods in Video Colonoscopy: Results From the MICCAI 2015 Endoscopic Vision Challenge. IEEE transactions on medical imaging 36(6), 1231–1249 (2017)

[5] Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., Johansen, D., Griwodz, C., Stensland, H.K., Garcia-Ceja, E., Schmidt, P.T., Hammer, H.L., Riegler, M.A., Halvorsen, P., de Lange, T.: HyperKvasir, a Comprehensive Multi-Class Image and Video Dataset for Gastrointestinal Endoscopy. Springer Nature Scientific Data 7, Article no. 283 (2020)

[6] Guo, Y.B., Matuszewski, B.: Giana polyp segmentation with fully convolutional dilation neural networks. In: Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. pp. 632–641 (2019)

[7] Guo, Y., Bernal, J., J Matuszewski, B.: Polyp Segmentation with Fully Convolutional Deep Neural Networks—Extended Evaluation Study. Journal of Imaging 6(7), 69 (2020)

[8] Hicks, S.A., Jha, D., Thambawita, V., Halvorsen, P., Hammer, H., Riegler, M.A.: EndoTect: A Competition on Automatic Disease Detection in the Gastrointestinal Tract. In: ICPR 2020 Workshops and Challenges. LNCS, Springer (2020)

[9] Jha, D., Ali, S., Johansen, H.D., Johansen, D., Rittscher, J., Riegler, M.A., Halvorsen, P.: Real-Time Polyp Detection, Localisation and Segmentation in Colonoscopy Using Deep Learning. arXiv preprint arXiv:2006.11392 (2020)

[10] Jha, D., Hicks, S.A., Emanuelsen, K., Johansen, H., Johansen, D., de Lange, T., Riegler, M., Halvorsen, P.: Medico Multimedia Task at MediaEval 2020:Automatic Polyp Segmentation. In: Proceedings of MediaEval CEUR Workshop (2020)

[11] Jha, D., Riegler, M., Johansen, D., Halvorsen, P., Johansen, H.: DoubleU-Net: A Deep Convolutional Neural Network for Medical Image Segmentation. In: Proceedings of the International Symposium on Computer Based Medical Systems (CBMS) (2020)

[12] Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D.: Kvasir-SEG: A Segmented Polyp Dataset. In: Proceedings of the International Conference on Multimedia Modeling (MMM). pp. 451–462 (2020)

[13] Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P., Johansen, H.D.: ResUNet++: An Advanced Architecture for Medical Image Segmentation. In: Proceedings of the International Symposium on Multimedia (ISM). pp. 225–230 (2019)

[14] Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Proceedings of International Conference on Medical image computing and computer-assisted intervention. pp. 234–241 (2015)

A.16 Paper XVI: Improving generalizability in polyp segmentation using ensemble convolutional neural network

Authors: N. Tomar, N. Ibtehaz, D. Jha, P. Halvorsen and S. Ali

Abstract: Medical image segmentation is a crucial task in medical image analysis. Despite near expert-label performance with the application of the deep learning method in medical image segmentation, the generalization of such models in the clinical environment remains a significant challenge. Transfer learning from a large medical dataset from the same domain is a common technique to address generalizability. However, it is difficult to find a similar large medical dataset. To address generalizability in polyp segmentation, we have used an ensemble of four MultiResUNet architectures, each trained on the combination of the different centered datasets provided by the challenge organizers. Our method achieved a decent performance of 0.6172 $\pm$ 0.0778 for the multi-centered dataset. Our study shows that significant work needs to be done to develop a computer-aided diagnosis system to detect and localize polyp of the multi-center datasets, which is essential for improving the quality of the colonoscopy.

Published: 3rd International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV2021) in conjunction with the 18th IEEE International Symposium on Biomedical Imaging ISBI 2021.

Candidate contributions: D. Jha participated in conceptualizing the work. He wrote the manuscript mostly and led the revision process to prepare the final version of the manuscript. Finally, he presented the paper at the conference.

Thesis objectives: Objective III

Improving generalizability in polyp segmentation using ensemble convolutional neural network

Nikhil Kumar Tomar[a], Nabil Ibtehaz[c], Debesh Jha[a,b], Pål Halvorsen[a] and Sharib Ali[d]

[a]SimulaMet, Oslo, Norway
[b]UiT The Arctic University of Norway, Tromsø, Norway
[c]Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
[d]Department of Engineering Science, Big Data Institute, University of Oxford, Oxford, UK

Abstract

Polyp segmentation is crucial for the diagnosis of colorectal cancer. Early detection and removal of polyps can prolong the life of patients and reduce the mortality rate. Despite near expert-label performance with applying the deep learning method in polyp segmentation tasks, the generalization of such models in the clinical environment remains a significant challenge. Transfer learning from a large medical dataset from the same domain is a common technique to address generalizability. However, it is difficult to find a similar large medical dataset. In this work, we investigate the feasibility of building a generalizable model for polyp segmentation using an ensemble of four MultiResUNet architectures, each trained on the combination of the different centered datasets provided by the challenge organizers. Our method achieved a decent performance of 0.6172 ± 0.0778 for the multi-centered dataset. Our findings show that significant work needs to be done to design a robust segmentation model for the development of a clinically acceptable system.

Keywords

Generalization, colonoscopy, convolutional neural network, polyp segmentation, health informatics

1. Introduction

The medical world concerned with the digestive system is currently in the midst of an uprising wave of increased adaption and technology usage for automatic analysis and decision support. With the increase of publicly available datasets, adapted methodologies such as convolutional neural networks, improved hardware, and increased collaboration of computer scientists and medical communities, this development is gaining more momentum than ever before. Global Cancer Statistics 2020 (GLOBOCAN 2020) estimated colorectal cancer as the third most frequently diagnosed cancer. Colorectal cancer accounts for 10.0% of total cancer, which is only 1% below to the most frequently caused cancer, i.e., female breast cancer (11.7%) and lung cancer (11.4%) [1]. Screening and removal of adenomatous polyps and other precancerous anomalies is one of the best working methods for the early detection and avoiding colorectal cancer-based mortality and incidence [2].

Deep learning-based methods have gained popularity in the development of the computer-aided diagnosis (CADx) system for detection of the colorectal polyps [3, 4, 5]. The successful

EndoCV'21: The 3rd International Workshop and Challenge on Computer Vision in Endoscopy (in conjunction with IEEE ISBI 2021), April 13th, 2021, Nice, France
sharib.ali@eng.ox.ac.uk (S. Ali)

deployment of a CAD system for polyp segmentation would require a trained model that achieves high performance on unseen datasets irrespective of different hospitals, cohort populations, and imaging protocol. However, deep learning algorithms are data-driven. The desired generalizable algorithms would require large, high-quality, and diverse datasets samples to train algorithms. Creating such datasets requires expert endoscopists and computer scientists for labeling and pixel-wise annotations. In general, there are only a few publicly available datasets. Although some studies report high performance on a specific dataset, the dataset is not publicly released [5, 3]. Therefore, it is challenging to develop a generalizable polyp segmentation model with a limited or single-center dataset.

Challenges and competitions are a good technique to access and explore new datasets for experimentation. It is also a fair way to compare methods, analyze, and improve the results on provided dataset. Additionally, challenges provide a solution for the lack of dataset availability and help develop reliable and clinically applicable methods. We participated in the EndoCV2021 challenge[1] to explore a multi-center dataset and develop a generalizable polyp segmentation CADx system. Our goal is to investigate and develop a generalizable model, compare our results with the other participants in the challenge, and observe our model's behavior.

The EndoCV2021 challenge offered two different tasks, namely, Detection generalization challenge and Segmentation generalization challenge. We only participated in the **Segmentation generalization challenge**. We used an ensemble model as our solution for the segmentation generalization challenge. The main motivation behind using ensemble methods was that it showed winning results in the different challenges [6, 7]. For our solution, we made an ensemble of four MultiResUNet [8] model. In short, the main contribution of our work are as follows:

- We explore a convolutional neural network-based model for the generalizable polyp segmentation task with a multi-center dataset. In this study, the training dataset is collected from five different medical institutions from five different countries, and test data comes from independent institutions.
- Our work reveals that the proposed deep learning model has significant challenges with the images having bleeding, adenomas, and covered by dyed. The model mostly showed over-segmentation or failed miserably with such scenarios. We highlight these cases that are among the significant challenges for developing a generalizable algorithm for the polyp segmentation task.

The remainder of this paper is organized into five sections. Section 2 provides a short overview of the related work. Section 3 gives an overview Methodology, and Section 4 describes the experimental setup. Section 5 presents the results obtained using the challenge dataset. Finally, we summarize and conclude the paper in Section 6.

2. Related Work

CNN-based architectures for polyp segmentation have been a common strategy for the development of the CADx system. We briefly describe the work on polyp segmentation and generalizability in the below subsection.

[1] https://endocv2021.grand-challenge.org/

2.1. Polyp segmentation

There has been several study on colorectal polyp segmentation [5, 3, 8, 6, 9]. Most of the work have proposed an architecture based on U-Net [10]. There have been also work on improving the segmentation performance on the publicly available dataset [11, 4, 12] to the real-time performance [13, 14, 3]. Although mostly retrospective studies were conducted [5, 13], there has also been work that carried prospective randomized controlled studies [15, 16]. However, most of the studies were conducted on the dataset from a single center. The experiments on multi-center datasets have often been ignored.

2.2. Generalizibility

In medical image analysis, generalization refers to the ability of the machine learning algorithm that is trained on specific interventions in specific health centers should be able to perform well over other interventions or different health center [7]. Poor generalizability has become one of the major issues for the clinical translation of the deep learning methods into clinical practise [17]. Meta-learning under a few-shot setting has gained popularity in developing a generalizable deep learning model and resolve the issue of data scarcity [18, 19].

In our previous study [4], due to the lack of a publicly available multi-center dataset, we have used a trained dataset on one publicly available dataset [20] and tested it against another [21] to observe the generalization capability. Additionally, we have also mixed the datasets from two or more institutions to observe the model's generalization capability. This is our first work where we have the opportunity to train the model with a multi-center dataset (five different center datasets) and benchmark on the completely new dataset.

3. Methodology

To address the generalizability problem in polyp segmentation, we used an ensemble of the four MultiResUNet [8] models. As each folder of the dataset has images from a unique center, we use a different subset of the dataset to train each of the MultiResUNet models. The MultiResUNet is an encoder-decoder architecture, which is an improvement over the existing U-Net [10] architecture. It combines the strength of the U-Net and improving it by replacing the existing components with more effective components such as "MultiRes block" and "Res Path". The MultiResUNet consists of four encoder blocks, four decoder blocks, and a bridge connecting them. The encoder takes the input image, encodes it, and extracts more useful features from it. Later these features are passed to the decoder, where they are upsampled and concatenated with the feature maps from the skip connection. Finally, these features are used to generate a segmentation mask for the input image. The additional block to form MultiResUNet models is briefly described below.

3.1. MultiRes block

The MultiRes block is the major component used in the MultiResUNet [8] architecture. It is the replacement of the convolution block, i.e., two 3×3 convolution used in the U-Net. The

MultiRes block is inspired from the Inception architecture [22] which consists of multiple parallel convolutions with 3×3, 5×5, and 7×7 kernel size. These multiple parallel convolutions help in capturing objects with different shapes and sizes. Using the bigger 5×5 and 7×7 kernel size increases the memory requirement. Therefore, these bigger kernels are factorized and replaced by multiple 3×3 convolutions. The MultiRes block begins with a single 3×3 convolution, which is followed by two 3×3 convolutions which are combined together to get the resultant effect of a 5×5 convolution. Next again are the multiple 3×3 convolutions which are repeated to give the resultant effect of a 7×7 convolution. The outputs from these convolutional blocks are concatenated together to have different scale feature maps. A residual connection is also used, which connects the input to the concatenated output.

3.2. Res path

The introduction of the skip connection in the U-Net architecture proves to be a significant contribution towards improving semantic segmentation performance. These skip connections enable the flow of information from the encoder to the decoder that is lost during the pooling operation. The simple concatenation of the features from the encoders to the decoders is flawed. For example, the first skip connection contains the low-level features from the early layers, which are fused with high-level features in the decoder. Therefore, there is a semantic gap between the features that being merged. To resolve this semantic gap, some convolutional layers and shortcut connections are being introduced as the skip connection in the MultiResUNet, called the "Res path".

3.3. MultiResUNet Architecture

The MultiResUNet [8] architecture begins by feeding the input image to the first encoder, which consists of the MultiRes block, followed by a 2×2 max-pooling with a stride value of 2. The max-pooled feature maps are passed on to the next encoder, and this process is repeated four times. In each step, the number of filters doubles, and the spatial resolution reduces by half. The output of the MultiRes block acts as the skip-connection, which first passes through the Res path and joins the decoder block. Inside each Res path, the number of convolution blocks decreases from 4, 3, 2 to 1 respectively along the four Res paths. The decoder begins with a 2×2 transpose convolution, which doubles the feature maps' spatial dimensions. Next, the feature maps are concatenated with the output of the Res path. Subsequently, the MultiRes block is used to learn the semantic representation. Similarly, the network is followed by three more decoder blocks, where the number of filters decreases and the feature maps resolution increases. It is then followed by a 1×1 convolution with sigmoid activation to generate the binary segmentation mask.

4. Experiment

To evaluate the performance of the ensemble method, we have performed extensive experiments. This section describes the dataset, evaluation metrics, training strategy, and implementation details used in our experimentation. Figure 1 shows the block diagram of the proposed ensemble

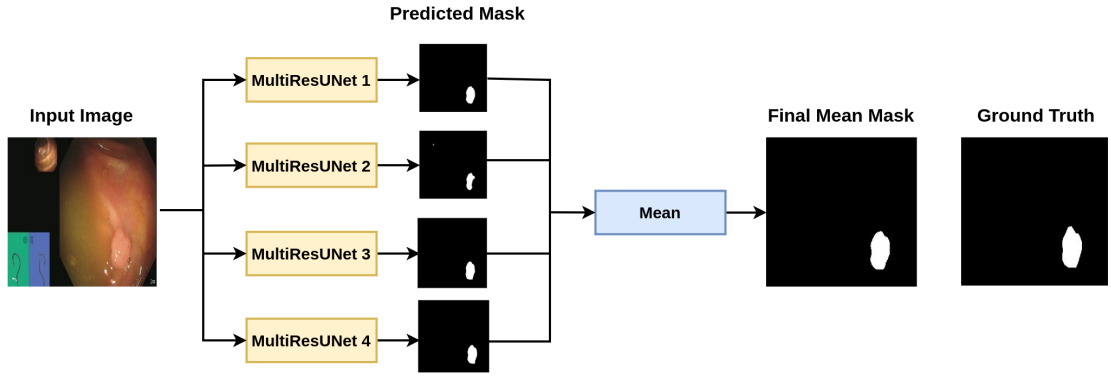


Figure 1: Block diagram of the proposed ensemble architecture

method. As explained in Section 3, the input image is fed to the different MultiResUNet models that produce different segmentation outputs. These predicted outputs from four distinct models are averaged to get the final mean mask.

4.1. Dataset

EndoCV2021 dataset consists of both a single frame dataset and sequence dataset. The dataset is captured from five different institutes. Each center dataset is provided in a separate folder. The training dataset consists of 1452 single image frames. Additionally, the dataset also consists of 165 negative sequence frames and 490 positive sequence frames, in a total of 655 image sequences. The sequence frames are taken from videos. Both positive (polyp) and negative (normal) frames are provided. Each center dataset has a separate image, mask, image with the bounding box, and bounding box information. All the images and their corresponding masks are in jpeg format. The dataset is currently only open to be used for EndoCV2021 challenge participation purpose.

4.2. Evaluation Metrics

The evaluation metric for the detection task is the Average mean precision. Additionally, a mean deviation is also calculated. For the segmentation tasks, the evaluation metrics such as F1-score, mean Intersection over Union (mIoU), recall, precision, F2-score, and overall accuracy is calculated. Additionally, the mean deviation for each of the metrics is also calculated. The procedure for the calculation can be found at GitHub [2].

4.3. Training strategy

For training, the model1, i.e., MultiResUNet1, the subset from center1, center3, and center4 were used. Similarly, we used center2, center1, and center4 for training model2 (MultiResUNet2). Likewise, we used center2, center3, and center1 for training model3. For training model4, we

[2]https://github.com/sharibox/EndoCV2021-polyp_det_seg_gen

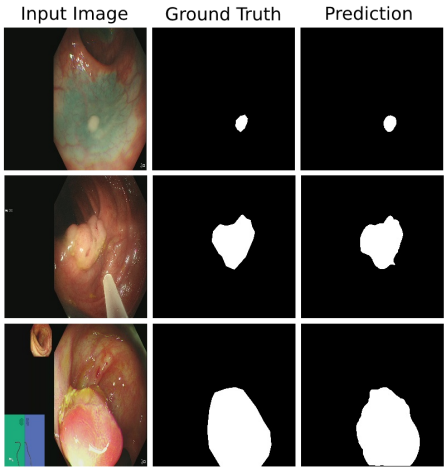


Figure 2: Qualitative results of the four ensemble MultiResUNet[8] models. The example images show that the ensemble models produce high-quality segmentation maps for different polyp shapes and sizes.

used the images from center2, center3, and center 4. We use the dataset from center5 as the validation set.

4.4. Implementation Details

We have implemented the MultiResUNet using the Keras with TensorFlow as a backend. The experiments were run on the Experimental Infrastructure for Exploration of Exascale Computing(eX3), NVIDIA DGX-2 machine. All four models are trained on 100 epochs using the same set of hyperparameters. Each model uses an image size of 256×256 pixels with a batch size of 8. The dice coefficient is used as the loss function with Adam optimizer. The default learning $1e-3$ is used to training the model. We also use the ReduceLROnPlateau callback to reduce further the learning rate for better generalization of the model.

5. Results and Discussion

On the test dataset, we achieved a score of 0.6172 ± 0.0778. Here, 0.6172 is the generalization score and 0.0778 is the generalization deviation. Figure 2 shows the qualitative results of the ensemble MultiResUNet model. The first, second and third column shows the input image, their corresponding ground truth, and the predictions. From the qualitative results, we can see that the model is performing well on polyp of different shapes and sizes (i.e., small, medium, and large-sized polyps).

However, a detailed dissection of the validation results shows that the models produce over-segmentation for the outputs when the input images have bleeding. The model also fails on challenging images such as flat polyps. The model also has a problem with detecting when the input images are covered with dyed. Mostly the models show over-segmentation, and sometimes the model completely fails to produce any segmentation masks. However, a more detailed conclusion can be made when we can visualize the qualitative results on the test dataset.

6. Conclusion

In this paper, we presented a cascaded MultiResUNet based solution for addressing the generalizability in polyp segmentation. The model can automatically segment polyp. The experimental results showed that the ensemble model obtained an evaluation score of 0.6172 ± 0.0778. The research results open a wide range of research directions to build generalizable model on new datasets. Moreover, we showed that ensemble models are not always the best choice for biomedical data science challenges. A deep analysis of the qualitative results showed that the model performs well on polyps of different shapes and sizes. In the future, we plan to explore the transfer learning from both large natural datasets and from biomedical imaging datasets (polyp or similar domain datasets) for improving the results on the polyp segmentation tasks.

Acknowledgment

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

References

[1] H. Sung, et al., Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: a cancer journal for clinicians (2021).

[2] A. M. Wolf, E. T. Fontham, T. R. Church, C. R. Flowers, C. E. Guerra, S. J. LaMonte, R. Etzioni, M. T. McKenna, K. C. Oeffinger, Y.-C. T. Shih, et al., Colorectal cancer screening for average-risk adults: 2018 guideline update from the american cancer society, CA: a cancer journal for clinicians 68 (2018) 250–281.

[3] J. Y. Lee, et al., Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets, Scientific reports 10 (2020) 1–9.

[4] D. Jha, P. H. Smedsrud, D. Johansen, T. de Lange, H. Johansen, P. Halvorsen, M. Riegler, A comprehensive study on colorectal polyp segmentation with resunet++, conditional random field and test-time augmentation, IEEE journal of biomedical and health informatics (2021).

[5] P. Wang, et al., Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy, Nature biomedical engineering 2 (2018) 741–748.

[6] S. Ali, M. Dmitrieva, N. Ghatwary, S. Bano, G. Polat, A. Temizel, A. Krenzer, A. Hekalo, Y. B. Guo, B. Matuszewski, et al., Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy, Medical image analysis (2021) 102002.

[7] T. Roß, A. Reinke, P. M. Full, M. Wagner, H. Kenngott, M. Apitz, H. Hempe, D. Mindroc-Filimon, P. Scholz, T. N. Tran, et al., Comparative validation of multi-instance instrument

segmentation in endoscopy: Results of the robust-mis 2019 challenge, Medical Image Analysis 70 (2021) 101920.
[8] N. Ibtehaz, M. S. Rahman, Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation, Neural Networks 121 (2020) 74–87.
[9] Y. Guo, J. Bernal, B. J Matuszewski, Polyp segmentation with fully convolutional deep neural networks—extended evaluation study, Journal of Imaging 6 (2020) 69.
[10] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Proc. of International Conference on Medical image computing and computer-assisted intervention (MICCAI), 2015, pp. 234–241.
[11] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, H. D. Johansen, Resunet++: An advanced architecture for medical image segmentation, in: Proc. of International Symposium on Multimedia (ISM), 2019, pp. 225–2255.
[12] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, H. D. Johansen, Doubleu-net: A deep convolutional neural network for medical image segmentation, in: Proc. of International Symposium on Computer-Based Medical Systems (CBMS), 2020, pp. 558–564.
[13] N. K. Tomar, D. Jha, S. Ali, H. D. Johansen, D. Johansen, M. A. Riegler, P. Halvorsen, Ddanet: Dual decoder attention network for automatic polyp segmentation, in: Proc. of International Conference on Pattern Recognition (ICPR) Workshop, 2020.
[14] D. Jha, S. Ali, N. K. Tomar, H. D. Johansen, D. Johansen, J. Rittscher, M. A. Riegler, P. Halvorsen, Real-time polyp detection, localization and segmentation in colonoscopy using deep learning, IEEE Access 9 (2021) 40496–40510.
[15] P. Wang, et al., Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study, Gut 68 (2019) 1813–1819.
[16] J.-R. Su, et al., Impact of a real-time automatic quality control system on colorectal polyp and adenoma detection: a prospective randomized controlled study (with videos), Gastrointestinal endoscopy 91 (2020) 415–424.
[17] K. Yasaka, O. Abe, Deep learning and artificial intelligence in radiology: Current applications and future directions, PLoS medicine 15 (2018) e1002707.
[18] P. Zhang, J. Li, Y. Wang, J. Pan, Domain adaptation for medical image segmentation: A meta-learning method, Journal of Imaging 7 (2021) 31.
[19] S. Ravi, H. Larochelle, Optimization as a model for few-shot learning (2016).
[20] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, H. D. Johansen, Kvasir-seg: A segmented polyp dataset, in: Proc. of International Conference on Multimedia Modeling (ISM), 2020, pp. 451–462.
[21] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, F. Vilariño, Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians, Computerized Medical Imaging and Graphics 43 (2015) 99–111.
[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proc. of IEEE conference on computer vision and pattern recognition (CVPR), 2015, pp. 1–9.

A.17 Paper XVII: The EndoTect 2020 challenge: evaluation and comparison of classification, segmentation and inference time for endoscopy

Authors: S. A. Hicks, D. Jha, V. Thambawita, P. Halvorsen, H. Hammer, and M. A Riegler

Abstract: The EndoTect challenge at the International Conference on Pattern Recognition 2020 aims to motivate the development of algorithms that aid medical experts in finding anomalies that commonly occur in the gastrointestinal tract. Using HyperKvasir, a large dataset containing images taken from several endoscopies, the participants competed in three tasks. Each task focuses on a specific requirement for making it useful in a real-world medical scenario. The tasks are (i) high classification performance in terms of prediction accuracy, (ii) efficient classification measured by the number of images classified per second, and (iii) pixel-level segmentation of specific anomalies. Hopefully, this can motivate different computer science researchers to help benchmark a crucial component of a future computer-aided diagnosis system, which in turn, could potentially save human lives.

Published: Proceedings of Pattern Recognition (ICPR Workshop and Challenges)

Candidate contributions: D. Jha led the segmentation challenge, which was a part of EndoTect 2020. He annotated and prepared the dataset for the segmentation tasks. Additionally, he wrote the script for evaluating the segmentation task. Moreover, he participated in drafting and revising the manuscript.

Thesis objectives: Objective I, Objective II, Objective III

The EndoTect 2020 Challenge: Evaluation and Comparison of Classification, Segmentation and Inference Time for Endoscopy

Steven A. Hicks[1,2( )], Debesh Jha[1,3], Vajira Thambawita[1,2], Pål Halvorsen[1,2], Hugo L. Hammer[1,2], and Michael A. Riegler[1]

[1] SimulaMet, Oslo, Norway
steven@simula.no
[2] Oslo Metropolitan University, Oslo, Norway
[3] UIT The Arctic University of Norway, Tromsø, Norway

Abstract. The EndoTect challenge at the International Conference on Pattern Recognition 2020 aims to motivate the development of algorithms that aid medical experts in finding anomalies that commonly occur in the gastrointestinal tract. Using HyperKvasir, a large dataset containing images taken from several endoscopies, the participants competed in three tasks. Each task focuses on a specific requirement for making it useful in a real-world medical scenario. The tasks are (i) high classification performance in terms of prediction accuracy, (ii) efficient classification measured by the number of images classified per second, and (iii) pixel-level segmentation of specific anomalies. Hopefully, this can motivate different computer science researchers to help benchmark a crucial component of a future computer-aided diagnosis system, which in turn, could potentially save human lives.

Keywords: GI endoscopy · Anomaly detection · Segmentation · Accuracy · Efficient processing · Challenge

1 Introduction

The human digestive system is prone to suffer from many different diseases and abnormalities throughout a human lifetime. Some of these may be life-threatening and pose a severe risk to a patient's health and well-being. In most cases, if the detection of lethal disease is done early enough, it can be treated with a high chance of being fully healed. Therefore, it is important that all lesions are identified and reported during a routine investigation of the gastrointestinal (GI) tract. Currently, the gold-standard in performing these investigations is through video endoscopies, which is a procedure involving a small camera attached to a tube that is inserted either orally or rectally. However, there is one major downside to this procedure. The method is highly dependent on the skills and

A. Del Bimbo et al. (Eds.): ICPR 2020 Workshops, LNCS 12668, pp. 263–274, 2021.
https://doi.org/10.1007/978-3-030-68793-9_18

experience of the person operating the endoscope, which in turn results in a high operator variation and performance [18,28,47]. This is one of the reasons for high miss-rates when measuring polyp detection performance, with some miss-rates being as high as 20% [25]. Polyps are small mushroom-like growths that appear on the inner-lining of the GI wall and are the leading cause to colorectal cancer.

Automated detection of GI anomalies has been a research topic for at least two decades, and in the last few years, there have been various AI-based solutions have been proposed using both hand-crafted features and representation learning methods (such as neural networks). However, even though there are many approaches for detecting [1,4,7,13,32,33,35,37,42,44,45,48] and segmenting [14,23,24] GI findings, even some targeting real-time analysis [2,39,40], there is room for improvement. One popular way of benchmarking and improving the state-of-the-art in machine learning is through publicly hosted challenges that motivate researchers to contribute to a use-case they otherwise would not work on. For GI automatic image and video analysis, there have been several such challenges hosted the last few years [3,19,38,41], with each bringing new insights into the current state of the field.

This year, we present three different tasks for participants to complete. The tasks are as follows: (i) The *detection* task which aims for high classification accuracy among 23 different classes, (ii) the *efficient detection* task which targets real-time performance for the same 23 classes of the *detection* task, and (iii) the *segmentation* task that aims to segment polyps in GI images. To participate, the teams had to solve at least one of the provided tasks. Overall, six teams participated, where all participants, in one way or another, utilize deep neural networks to solve the provided tasks. The results vary between teams, but most are able to achieve satisfactory scores in terms of what is suitable for use in clinics [36].

We see this as an opportunity to aid medical doctors by helping them detect lesions through automatic frame analysis done live during endoscopy examinations. The pattern recognition community has a lot of knowledge that could assist in this task, making this challenge a perfect fit for the International Conference on Pattern Recognition (ICPR). The work done in this competition, detecting and segmenting medical findings in the GI tract, has the potential of making a real societal impact, as it directly affects the quality of care that healthcare professionals can provide.

2 Dataset Details

For this challenge, we provided the participants with a development dataset that was to be used to train their algorithms. This year, we provided HyperKvasir [6], which is a large GI dataset consisting of labeled and unlabeled images taken from several different GI endoscopies. The dataset is split into four distinct parts; Labeled image data, unlabeled image data, segmented image data, and annotated video data. In total, the dataset contains 110,079 images (see Fig. 1 for examples) and 374 videos where it captures anatomical landmarks, pathological

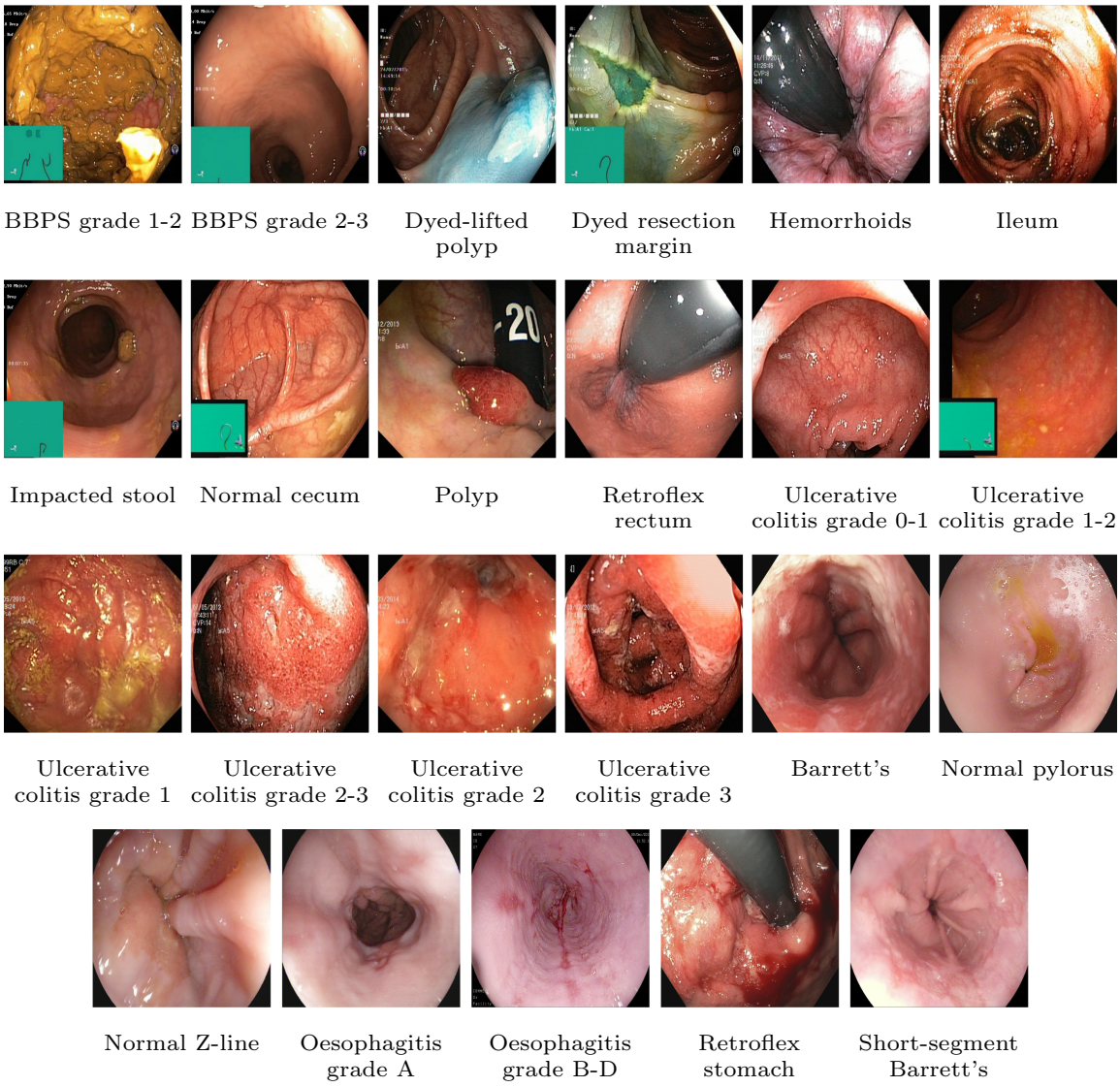


Fig. 1. One example taken from each of the classes contained within the development dataset.

findings, and normal findings. The result is more than one million images and video frames altogether.

For the *detection* and *efficient detection* tasks, participants used the 23 classes provided in the labeled part of the dataset to develop their algorithms. The number of images per class is not balanced, which is a general challenge in the medical field due to the fact that some findings occur more often than others. This adds an additional challenge for researchers since methods applied to the data should also be able to learn from a small amount of training data. The participants could also use the unlabeled part of the dataset to further improve their algorithm by using, for example, a student-teacher approach or the pseudo labels provided in the HyperKvasir GitHub repository[1].

[1] https://github.com/simula/hyper-kvasir.

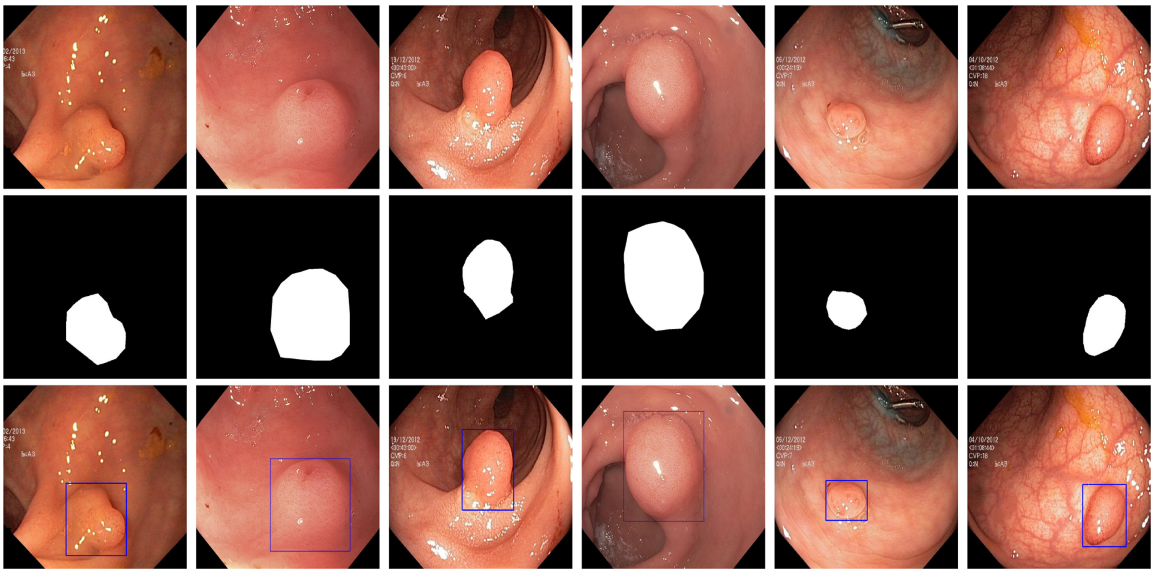


Fig. 2. Some example images of polyps and their corresponding masks and bounding boxes.

For the *segmentation* task, we provide the original image, a segmentation mask, and a bounding box for 1,000 images containing polyps. An example is shown in Fig. 2, where we see six samples taken from the segmentation dataset. For the image masks, the white pixels depict the area of the image containing a polyp, while the black background pixels do not. The bounding box is defined as the outermost pixels of the found polyp.

3 Tasks

With the end-goal of helping medical experts detect more lesions, we present three different tasks that each target a different requirement for in-clinic use. In the following, we give a detailed description of each task and describe how each was evaluated using the appropriate metrics. The script used to evaluate each task is on GitHub[2].

3.1 Detection Task

The detection task stems from the requirement of the high detection accuracy needed to be viable for use in a clinical setting. Participants are asked to develop algorithms that achieve high classification scores on the 23 different classes present in the labeled part of the development dataset (further described in Sect. 2). Submissions to this task was a comma-separated values (CSV) file, where each line contained the filename of the predicted image in the test dataset, the predicted label, and a confidence score ranging from 0 to 1 for the predicted label.

[2] https://github.com/simula/endotect-2020-submission-evaluation.

For this detection task, we use several standard metrics commonly used to evaluate classification tasks. We collect all true and false positives and negatives, and we then calculate metrics such as precision, recall/sensitivity, specificity, F1, and Matthews correlation coefficient (MCC) for multi-classification (also called R_k statistic for multiclass classification). The officially reported metric for evaluating this task is the MCC, which will also be the metric used to rank the submissions.

3.2 Efficient Detection Task

The efficient detection task focuses on the real-time analysis needed to deliver instant feedback to doctors performing endoscopies. To satisfy this requirement, the algorithm must achieve good classification scores while also being able to classify images as fast as they are put on screen, which is approximately 30 frames per second. For the efficient detection task, we asked participants to submit a Docker [31] image so that we can evaluate the speed and efficiency of the proposed algorithm on the same hardware. The Docker image was set up to produce a submission file similar to the one described for the detection task, but in addition to the aforementioned value entries, the classification processing time was also appended to the end of each row. All submissions submitted to this task were run on what could be considered consumer-grade hardware, that is, a computer running Arch Linux with an Intel Core i9-10900K processor, an Nvidia GeForce 1080 Ti graphics processing unit (GPU), and 32 gigabytes of RAM.

As one could generally achieve higher processing speeds with an algorithm with lower prediction accuracy, the evaluation used a combination of the MCC classification score and the number of frames processed per second. The focus here is on the "speed" aspect of the algorithm, so the only requirement from a classification standpoint is that it exceeds a set MCC threshold so that it is still viable for in-clinic use. We set the threshold of 85% as it is considered standard for automatic detection systems for colonoscopies [36].

3.3 Segmentation Task

In the segmentation task, we asked participants to use the segmented images provided in the dataset to generate segmentation masks of polyps automatically. Polyps are clumps of cells that form on the mucosal wall of the GI tract and come in a variety of shapes and sizes. Polyps are among the most critical findings in an endoscopy procedure as they are a precursor to different cancer types, including colorectal cancer, which is one of the most lethal cancer types worldwide [22]. The motivation behind this task is rooted in the requirement for not only detecting that a frame contains a polyp, but also showing where it is so that it can be properly removed. A typical example of a segmented polyp is shown in Fig. 2.

For the evaluation of this task, we use the standard metrics commonly used to evaluate segmentation tasks. This includes precision, recall, the Dice coefficient, and the Intersection over Union (IoU, also known as the Jaccard index). The

metric which will be used to rank submissions will be the IoU. To calculate the metrics, we use the implementation provided by the Python library scikit-learn [34].

4 Participants

This year, we received 26 registrations, of which six submitted results. Each participating team was allowed to submit as many runs to each task as they wished. In the following, we give a short summary of each participant's approach. A more detailed description of each approach can be found in the teams' corresponding challenge papers.

4.1 Team DeepBlueAI

Team *DeepBlueAI* participated in the detection and segmentation tasks. For the detection task, they trained a series of (CNNs), of which the best performing approach is an ensemble network consisting of a ResNet-50 [15] with batch normalization and an EfficientNet B7 [43]. For the segmentation task, they used two different approaches, namely instance and semantic segmentation. The instance segmentation approach used the Mask Scoring R-CNN [21] with ResNeXt-101 [49] as the backbone. As for the semantic segmentation, they used DeepLab V3 plus [9] with multi-scale training. More information on the specific implementation for both tasks can be found in [30].

4.2 Team Spearheads

Team *Spearheads* participated in all three tasks, where two runs were submitted to the detection and efficient detection tasks, and one run to the segmentation task. For the detection and efficient detection task, they used a Tiny Darknet model[3], which was trained using an augmented version of the provided development dataset. For the segmentation task, they used a standard UNet architecture trained on the provided segmentation dataset, which was expanded using augmentation by Augmentor [5]. More information about team *Spearheads* approach can be found in [11].

4.3 Team NKT

Team *NKT* participated in the segmentation task, where they submitted one run. Their approach used a novel CNN-based architecture, which they named Dual Decoder Attention Network (DDANet). The architecture uses a single encoder network together with multiple decoders that use a combination of residual learning [16] and squeeze and excitation networks [20]. A more detailed explanation of the approach can be found in [46].

[3] https://pjreddie.com/darknet/tiny-darknet/.

4.4 Team aggcmab

Team *aggcmab* participated in the detection and segmentation tasks, for which they submitted one run to each. For the detection task, *aggcmab* used a ResNet-50x1 with a BiT-M [27] backbone trained with a hierarchical loss function. For the segmentation task, they use a double encoder-decoder network with a dual path network [10] for the encoders and a Feature-Pyramid [29] for the decoders. More information on the specifics of team *aggcmab's* approach can be found in [12].

Table 1. Results for the best runs from the **detection** task. The table entries are ordered after the best MCC score.

Team name	Macro average			Micro average			MCC (R_K)
	Precision	Recall	F1-score	Precision	Recall	F1-score	
howard	0.683	0.646	0.659	0.913	0.913	0.913	0.903
DeepBlueAI	0.629	0.568	0.590	0.874	0.874	0.874	0.860
aggcmab	0.598	0.533	0.558	0.870	0.870	0.870	0.856
FAST-NU-DS	0.453	0.431	0.413	0.603	0.603	0.603	0.568
Spearheads	0.333	0.220	0.223	0.440	0.440	0.440	0.388

Table 2. Results for the best runs from the **efficient detection** task. Please note that FPS signifies the average FPS calculated over the provided test dataset.

Team name	Macro average			Micro average			MCC (R_K)	FPS
	Precision	Recall	F1-score	Precision	Recall	F1-score		
howard	0.528	0.496	0.503	0.785	0.785	0.785	0.765	129.748
Spearheads	0.333	0.220	0.223	0.440	0.440	0.440	0.388	49.132

Table 3. Results for the best runs from the **segmentation** task. The table entries are ordered after according to the best IoU score.

Team name	Precision	Recall	F1-score/Dice	IoU
aggcmab	0.928	0.937	0.920	0.871
DeepBlueAI	0.907	0.947	0.915	0.861
howard	0.915	0.882	0.879	0.822
NKT	0.858	0.799	0.787	0.701
Spearheads	0.801	0.801	0.754	0.656

4.5 Team FAST-NU-DS

Team *FAST-NU-DS* participated in the detection task, where they submitted three runs. Their approach used bagging with 11 DenseNet169 models, where the final classification was made through hard majority voting. More information on the method can be found in [26].

4.6 Team howard

Team *howard* participated in all three tasks, where they submitted one run to each. For the detection and efficient detection task, they used a CNN based on the ResNet152 [15] architecture trained with a hybrid loss. During training, they also applied some data augmentation, namely, contrast augmentation, color shift, brightness augmentation, flipping, perspective transformation, and blur. For the segmentation task, their solution is based on Cascade Mask R-CNN [8]. More information about their solution can be found in [17].

5 Results and Discussion

Tables 1, 2, and 3 show the results for all tasks in the challenge. Looking at the results for the *detection* task (Table 1), we see that team *howard* achieved the best result with their use of ResNet-152 together with a custom hybrid loss. They achieved an MCC score of 0.903, 0.043 ahead of *DeepBlueAI*, who came in second place. For the *efficient detection* task (Table 2), only two teams participated, but also here, team *howard* achieved the best average frames per second (FPS) while also keeping the classification performance high. None of the teams reached the target MCC threshold of 85%, but team *howard* achieved an MCC of 0.765 at an FPS of 129, far above the real-time requirement. Thus, maybe some speed can be traded for a more complex model, achieving a slightly higher MCC while still reaching a real-time speed of 30 FPS. A common trend in this task was using neural networks with less parameters, like MobileNet or Tiny Darknet, to achieve a higher FPS. For the *segmentation* task (Table 3), team *aggcmab* achieved the highest IoU with their double encoder-decoder network approach. They reached an IoU score of 0.871, which is quite close to the runner up score of 0.861 submitted by team *DeepBlueAI*. Overall, the results prove that deep learning works well for analyzing GI image data and confirms the potential of computer-assisted detection and segmentation of GI anomalies, but they also suggest that there is still some room for improvement.

From an organizational perspective, the challenge went smoothly, without any significant hiccups or sudden difficulties. Docker submissions seem to work well, but may require some extra effort from the participants, which may explain why we only got two submissions to the *efficient detection* task. The difficulty level of the tasks appears to be quite balanced as the different teams achieved a variety of scores. Next year, we plan to hold the challenge again, but this time with an extended evaluation dataset and an additional task for efficient segmentation.

6 Conclusion

This paper described the EndoTect 2020 challenge, which asked participants to build algorithms that automatically detect different findings commonly found in the GI tract. The challenge consisted of three distinct tasks, where participants

were given a large open dataset composed of videos from real endoscopies. We believe that computer scientists can make a real impact on the field of medicine, and the results presented in this paper show that we are at the point where machine learning algorithms have much potential in helping doctors detect more diseases.

References

1. Alammari, A., Islam, A.R., Oh, J., Tavanapong, W., Wong, J., De Groen, P.C.: Classification of ulcerative colitis severity in colonoscopy videos using CNN. In: Proceedings of the ACM International Conference on Information Management and Engineering (ACM ICIME), pp. 139–144 (2017). https://doi.org/10.1145/3149572.3149613
2. Angermann, Q., et al.: Towards real-time polyp detection in colonoscopy videos: adapting still frame-based methodologies for video sequences analysis. In: Cardoso, M.J., et al. (eds.) CARE/CLIP -2017. LNCS, vol. 10550, pp. 29–41. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67543-5_3
3. Bernal, J., Aymeric, H.: MICCAI endoscopic vision challenge polyp detection and segmentation (2017). https://endovissub2017-giana.grand-challenge.org/home/. Accessed 11 Dec 2017
4. Bernal, J., et al.: Polyp detection benchmark in colonoscopy videos using GTCreator: a novel fully configurable tool for easy and fast annotation of image databases. In: Proceedings of Computer Assisted Radiology and Surgery (CARS) (2018). https://hal.archives-ouvertes.fr/hal-01846141
5. Bloice, M.D., Roth, P.M., Holzinger, A.: Biomedical image augmentation using Augmentor. Bioinformatics (Oxford Engl.) **35**(21), 4522–4524 (2019). https://doi.org/10.1093/bioinformatics/btz259
6. Borgli, H., et al.: HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Sci. Data **7** (2020). https://doi.org/10.1038/s41597-020-00622-y. Article no. 283
7. Bychkov, D., et al.: Deep learning based tissue analysis predicts outcome in colorectal cancer. Sci. Rep. **8**(1), 3395 (2018). https://doi.org/10.1038/s41598-018-21758-3
8. Cai, Z., Vasconcelos, N.: Cascade R-CNN: high quality object detection and instance segmentation. IEEE Trans. Pattern Anal. Mach. Intell. (2019)
9. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv:1802.02611 (2018)
10. Chen, Y., Li, J., Xiao, H., Jin, X., Yan, S., Feng, J.: Dual path networks. In: Proceedings of the Annual Conference on Advances in Neural Information Processing Systems (NeurIPS), pp. 4467–4475 (2017)
11. Dutta, A., Bhattacharjee, R.K., Barbhuiya, F.A.: Efficient detection of lesions during endoscopy. In: Proceedings of the ICPR 2020 Workshops and Challenges. LNCS. Springer (2020)
12. Galdran, A., Carneiro, G., Ballester, M.A.G.: A hierarchical multi-task approach to gastrointestinal image analysis. In: Proceedings of the ICPR 2020 Workshops and Challenges. LNCS. Springer (2020)
13. Ghatwary, N.M., Ye, X., Zolgharni, M.: Esophageal abnormality detection using DenseNet based faster R-CNN with gabor features. IEEE Access **7**, 84374–84385 (2019). https://doi.org/10.1109/ACCESS.2019.2925585

14. Guo, Y., Bernal, J., Matuszewski, B.J.: Polyp segmentation with fully convolutional deep neural networks–extended evaluation study. J. Imaging **6**(7), 69 (2020)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
17. He, Q., Bano, S., Stoyanov, D., Zuo1, S.: Hybrid loss with network trimming for disease recognition in digestive endoscopy. In: Proceedings of the ICPR 2020 Workshops and Challenges. LNCS. Springer (2020)
18. Hewett, D.G., Kahi, C.J., Rex, D.K.: Efficacy and effectiveness of colonoscopy: how do we bridge the gap? Gastrointest. Endosc. Clin. **20**(4), 673–684 (2010). https://doi.org/10.1016/j.giec.2010.07.011
19. Hicks, S., et al.: ACM multimedia BioMedia 2019 grand challenge overview. In: Proceedings of the ACM International Conference on Multimedia (ACM MM), pp. 2563–2567 (2019). https://doi.org/10.1145/3343031.3356058
20. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141 (2018). https://doi.org/10.1109/CVPR.2018.00745
21. Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X.: Mask scoring R-CNN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6402–6411 (2019). https://doi.org/10.1109/CVPR.2019.00657
22. International Agency for Research on Cancer - WHO: Cancer fact sheets (2019). https://gco.iarc.fr/today/fact-sheets-cancers. Accessed 16 Dec 2019
23. Jha, D., Riegler, M., Johansen, D., Halvorsen, P., Johansen, H.: DoubleU-Net: a deep convolutional neural network for medical image segmentation. In: Proceeding of the International Symposium on Computer Based Medical Systems (CBMS) (2020)
24. Jha, D., et al.: ResUNet++: an advanced architecture for medical image segmentation. In: Proceedings of the International Symposium on Multimedia (ISM), pp. 225–230 (2019). https://doi.org/10.1109/ISM46123.2019.00049
25. Kaminski, M.F., et al.: Quality indicators for colonoscopy and the risk of interval cancer. N. Engl. J. Med. **362**(19), 1795–1803 (2010). https://doi.org/10.1056/NEJMoa0907667
26. Khan, Z., Tahir, M.A., Memon, S.: Medical diagnostic by data bagging for various instances of neural network. In: Proceedings of the ICPR 2020 Workshops and Challenges. LNCS. Springer (2020)
27. Kolesnikov, A., et al.: Big Transfer (BiT): general visual representation learning. arXiv preprint arXiv:1912.11370, June 2019
28. Lee, S.H., et al.: Endoscopic experience improves interobserver agreement in the grading of esophagitis by Los Angeles classification: conventional endoscopy and optimal band image system. Gut Liver **8**(2), 154 (2014). https://doi.org/10.5009/gnl.2014.8.2.154
29. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2117–2125 (2017)
30. Luo, Z., Che, L., He, J.: A hierarchical multi-task approach to gastrointestinal image analysis. In: Proceedings of the ICPR 2020 Workshops and Challenges. LNCS. Springer (2020)

31. Merkel, D.: Docker: lightweight Linux containers for consistent development and deployment. Linux J. **2014**(239) (2014)
32. Min, M., Su, S., He, W., Bi, Y., Ma, Z., Liu, Y.: Computer-aided diagnosis of colorectal polyps using linked color imaging colonoscopy to predict histology. Sci. Rep. **9**(1), 2881 (2019). https://doi.org/10.1038/s41598-019-39416-7
33. Mori, Y., et al.: Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy: a prospective study. Ann. Intern. Med. **169**(6), 357–366 (2018). https://doi.org/10.7326/M18-0249
34. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
35. Pogorelov, K., et al.: A holistic multimedia system for gastrointestinal tract disease detection. In: Proceedings of the ACM on Multimedia Systems Conference (MMSYS), pp. 112–123 (2017). https://doi.org/10.1145/3193740
36. Pogorelov, K., et al.: Deep learning and hand-crafted feature based approaches for polyp detection in medical videos. In: Proceedings of the IEEE International Symposium on Computer-Based Medical Systems (CBMS). IEEE (2018)
37. Pogorelov, K., et al.: Efficient disease detection in gastrointestinal videos-global features versus neural networks. Multimedia Tools Appl. **76**(21), 22493–22525 (2017). https://doi.org/10.1007/s11042-017-4989-y
38. Pogorelov, K., et al.: Medico multimedia task at mediaeval 2018. In: Proceeding of the MediaEval Benchmarking Initiative for Multimedia Evaluation Workshop (MediaEval) (2018)
39. Pogorelov, K., et al.: GPU-accelerated real-time gastrointestinal diseases detection. In: Proceedings of the International Symposium on Computer-Based Medical Systems (CBMS), pp. 185–190. IEEE (2016). https://doi.org/10.1109/CBMS.2016.63
40. Riegler, M., et al.: EIR - efficient computer aided diagnosis framework for gastrointestinal endoscopies. In: Proceedings of the IEEE International Workshop on Content-Based Multimedia Indexing (CBMI), pp. 1–6 (2016). https://doi.org/10.1109/CBMI.2016.7500257
41. Riegler, M., et al.: Multimedia for medicine: the medico task at MediaEval 2017. In: Proceeding of the MediaEval Benchmarking Initiative for Multimedia Evaluation Workshop (MediaEval) (2017)
42. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. Int. J. Comput. Assist. Radiol. Surg. **9**(2), 283–293 (2014). https://doi.org/10.1007/s11548-013-0926-3
43. Tan, M., Le, Q.V.: EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the International Conference on Machine Learning, pp. 6105–6114 (2019)
44. Thambawita, V., et al.: The medico-task 2018: disease detection in the gastrointestinal tract using global features and deep learning. In: Proceeding of the MediaEval Benchmarking Initiative for Multimedia Evaluation Workshop (MediaEval) (2018)
45. Thambawita, V.L., et al.: An extensive study on cross-dataset bias and evaluation metrics interpretation for machine learning applied to gastrointestinal tract abnormality classification. ACM Trans. Comput. Healthcare **1** (2020)
46. Tomar, N.K., Jha, D., Ali, S., Johansen, H.D.J.D., Riegler, M.A., Halvorsen, P.: DDANet: dual decoder attention network for automatic polyp segmentation. In: Proceedings of the ICPR 2020 Workshops and Challenges. LNCS. Springer (2020)

47. Van Doorn, S.C., et al.: Polyp morphology: an interobserver evaluation for the Paris classification among international experts. Am. J. Gastroenterol. **110**(1), 180–187 (2015). https://doi.org/10.1038/ajg.2014.326
48. Wang, Y., Tavanapong, W., Wong, J., Oh, J.H., De Groen, P.C.: Polyp-Alert: near real-time feedback during colonoscopy. Comput. Methods Programs Biomed. **120**(3), 164–179 (2015). https://doi.org/10.1016/j.cmpb.2015.04.002
49. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. arXiv preprint arXiv:1611.05431 (2016)

A.18 Paper XVIII : The medico-task 2018: Disease detection in the gastrointestinal tract using global features and deep learning

Authors: V. Thambawita, D. Jha, M. Riegler, P. Halvorsen, H. L. Hammer, H. D Johansen, D. Johansen,

Abstract: In this paper, we present our approach for the 2018 Medico Task classifying diseases in the gastrointestinal tract. We have proposed a system based on global features and deep neural networks. The best approach combines two neural networks and the reproducible experimental results signify the efficiency of the proposed model with an accuracy rate of 95.80%, a precision of 95.87%, and an F1-score of 95.80%.

Published: In Proceedings MediaEval 2018

Candidate contributions: D. Jha conceptualized this work. He developed and analyzed the two machine learning models (simple logistic classifier and logistic model tree classifier) using global features extracted from images. Additionally, he prepared and revised the manuscript with V. Thambawita.

Thesis objectives: Objective II

The Medico-Task 2018: Disease Detection in the Gastrointestinal Tract using Global Features and Deep Learning

Vajira Thambawita[1,3], Debesh Jha[1,4], Michael Riegler[1,3,5], Pål Halvorsen[1,3,5],
Hugo Lewi Hammer[2], Håvard D. Johansen[4], and Dag Johansen[4]
[1]Simula Research Laboratory, Norway [2]Oslo Metropolitan University, Norway [3]Simula Metropolitan, Norway
[4]University of Tromsø, Norway [5]University of Oslo, Norway
Contact:vajira@simula.no,debesh@simula.no

ABSTRACT

In this paper, we present our approach for the 2018 Medico Task classifying diseases in the gastrointestinal tract. We have proposed a system based on global features and deep neural networks. The best approach combines two neural networks, and the reproducible experimental results signify the efficiency of the proposed model with an accuracy rate of 95.80%, a precision of 95.87%, and an F1-score of 95.80%.

1 INTRODUCTION

Our main goal for the Medico Task [15] is to classify findings in images from the Gastrointestinal (GI) tract. This task provides two types of input data: Global Features (GFs) and original images. The 2017 Medico Task consisted of a balanced dataset with only 8 classes [12] whereas the current task consists of a highly imbalanced dataset with 16 classes [11, 12], i.e., making this years task more complicated. Different approaches have been used in the last year medico task [5, 7, 9, 10, 14, 17] based on GFs extractions and Convolutional Neural Networks (CNN) methods. We extend upon these solutions and present our solutions based on both GFs and transfer learning mechanisms using CNN. We achieve best results combining two CNNs and using an extra multilayer perceptron to combine the outputs of the two networks.

2 APPROACHES

We approach the problem of GI tract disease detection with small training datasets using five different methods: two based on GF extractions, and three based on CNN with transfer learning described below.

2.1 Global-feature-based approaches

Method 1 and **Method 2** use the concept of GFs. For the extraction of GFs, we use Lucence Image Retrieveal (LIRE) [6]. GFs are easy and fast to calculate, and can also be used for image comparison, image collection search and distance computing [14]. Based on [13, 16], we use Joint Composite feature (JCD), Tamura, Color layout, Edge Histogram, Auto Color Correlogram and Pyramid Histogram of Oriented Gradients (PHOG). These features represent the overall properties of the images. Adding more GFs is possible, but it may increase the redundant information which can reduce the overall classification performance.

The extracted features are sent to the different machine learning classifier for the multi-class classification. **Method 1** makes the use of extracted GFs that are sent to SimpleLogistic (SL) classifier. We input the same selected set of features to the logistic model tree (LMT) classifier in **Method 2**.

2.2 Transfer learning based approaches

Our CNN approaches use transfer learning mechanism with pre-trained models using the ImageNet dataset [18]. Resnet-152 [3] and Densenet-161 [4] have been selected, and this selection is based on top 1-error and top-5-errors rate of pre-trained networks in the Pytorch [8] deep learning framework.

One of the main problems of the given dataset is the "out of patient"-category which has only four images while other classes have a considerable number. The colour distribution of this class shows a completely different colour domain compared to the other categories. We identified this difference via manual investigations of the dataset and moved all four images of this category into the corresponding validation set folder. Then, the training set folder is filled with random Google images which are not related to the GI tract. To overcome the problems of stopping training in a local minima, we use the stochastic gradient descent [1] method with dynamic learning rate scheduling. The losses (loss 1 and loss 2 in Figure 1) of CNN methods were calculated for each network separately. Additionally, horizontal flips, vertical flips, rotations and re-sizing data augmentations have been applied to overcome the problem of over-fitting.

Method 3 uses transfer learning with Resnet-152 which has the top-1-error and top-5-error rates. The last fully connected layer of Resnet-152, which is originally designed to classify 1000 classes of the ImageNet dataset, has been changed to classify the 16 classes in the MEdico task. Usually, the transfer learning freezes pre-trained layers to avoid back propagation of large errors. This is because of newly added layers with random weights. However, we did not freeze the pre-trained layers, because modifying only the last layer cannot propagate huge errors backwards in transfer learning. The network was trained until it reached to the maximum validation accuracy of the validation dataset.

Method 4 extends Method 3 by using two parallel pre-trained models, Resnet-152 and Densenet-161, to get a cumulative decision at the end as depicted in Figure 1. The classification is based on an average of the two output probability vectors. Finally, one loss value was calculated and propagated for updating weights. However, this yields a restriction of updating weights of networks Resnet-152 and Densenet-161 separately as they required. Therefore, we calculated two different loss values (loss 1 and loss 2 in Figure 1) from each network to update their weights separately. Both

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

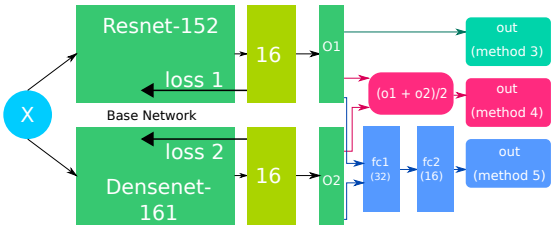


Figure 1: Block diagram of the CNN methods

networks were trained simultaneously until it reached to the best validation accuracy by changing hyper-parameters manually.

Method 5 was constructed to overcome the limitation of calculating the average of the probabilistic output of the two networks used in Method 4. Instead of calculating the average using the simple mathematical formula, another multilayer perceptron (MLP) has been merged with the above network to identify complex mathematical formula to get the cumulative decision as illustrated in Figure 1. Therefore, we passed the probability output of two networks (16 probabilities from each network) to a new MLP with 32 inputs, 16 outputs (via sigmoid layer) and one hidden layer with 32 units. In this, we used pre-trained Resnet-152 and Densenet-161 using the dataset and froze them before training the MLP. Then, we trained only the MLP to identify the best mathematical formula to get the cumulative decision.

3 RESULTS AND ANALYSIS

We have divided the development dataset into a training set (70%) and a validation set (30%). For the GFs based approach, ensembles of six extracted GFs were fetched to all the available machine learning classifiers (with different parameters) using WEKA[2] library. The SL and LMT classifiers outperform all other available classifiers for the dataset. The other promising classifier were Sequential minimal optimization (RBF kernel), and a combination of PCA with LibSVM (RBF) classifier.

On validation set, all the CNN methods (3-5) show accuracies of around 95% and specificities of around 99%. These are always better than the GFs based extraction methods (1,2) which have accuracies of around 82% and specificities of around 98%. According to the task organizers' evaluation results of the test dataset, Methods 3 to 5 show accuracies and specificities of around 99% again,which demonstrates our CNN methods are not overfitted with validation dataset.

Method 5 and 4 with Resnet-152 and Densenet-161 performs better compared to the Method 3 which has only Resnet-152 because of the capability of deciding the final answer based on two answers generated from two deep learning networks. However, getting a cumulative decision based on simple averaging function (Method 4) shows poor performance than the decision taken from a MLP (Method 5). As a result, Method 5 shows better results than method 4 by increasing the accuracy from 0.955 to 0.958. Therefore, Method 5 has been selected as our best method and confusion matrix represented in Table 1 was generated. An overview of the individual results obtained from five different experiments along with their performance metrics is presented in Table 2. Results obtained from the organizers for the test dataset is presented in the Table 3.

Table 1: The Confusion Matrix of Method 5 in our study

A:blurry-nothing, **B**:colon-clear, **C**:dyed-lifted-polyps, **D**:dyed-resection-margins, **E**:esophagitis,**F**:instruments, **G**:normal-cecum, **H**:normal-pylorus, **I**:normal-z-line, **J**:out-of-patient, **K**:polyps, **L**:retroflex-rectum, **M**:retroflex-stomach, **N**:stool-inclusions, **O**:stool-plenty, **P**:ulcerative-colitis

Actual class \ Predicted class	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
A	53															
B		81														
C			130	7												1
D			3	122												
E					115				19							
F						10					1					
G							125									
H								132								
I					11				121							
J						1				3						
K		1					6	2			172					
L							1					71				
M												2	118			
N														39		
O					1	1	2				4	1			110	
P																129

Table 2: Validation results

Method	REC	PREC	SPEC	ACC	MCC	F1	FPS
1	0.855	0.793	0.989	0.816	0.814	0.823	79
2	0.816	0.817	0.984	0.816	0.800	0.815	12
3	0.9536	0.9543	0.9968	0.9536	0.9498	0.9535	64
4	0.9555	0.9563	0.9969	0.9555	0.9519	0.9554	29
5	0.9580	0.9587	0.9971	0.9580	0.9546	0.9580	29

Table 3: Official results

Method	REC	PREC	SPEC	ACC	MCC	F1
1	0.8457	0.8457	0.9897	0.9807	0.8353	0.8456
2	0.8457	0.8457	0.9897	0.9807	0.8350	0.8457
3	0.9376	0.9376	0.9958	0.9922	0.9335	0.9376
4	0.9400	0.9400	0.9960	0.9925	0.9360	0.9400
5	0.9458	0.9458	0.9964	0.9932	0.9421	0.9458

The main considerable point in the confusion matrix in Table 1 is misclassification between categories E: esophagitis and I: normal-z-line. A large number of misclassifications like 30 images from the validation set occurred and a manual investigation was done to identify the reason. We notice that the images of these two categories were very similar to each other because of the close location in the GI tract, and identifying these is also a challeng for physicians.

4 CONCLUSION

In this paper, we presented five different methods for the multi-class classification of GI tract diseases. The proposed approach are based on the GFs, and pre-trained CNN with transfer learning mechanism. The combination of Resnet-152 and Densenet-161 with an additional MLP achieved the highest performance with both the validation dataset and the test dataset provided by the task organizers. We show that a combination of pre-trained deep neural models on ImageNet has better capabilities to classify images into the correct classes because of cumulative decision-making capabilities. For future work, we will combine deeper CNNs parallelly to add more cumulative decision taking capabilities for classifying multi-class objects. In addition to that, Generative Adversarial Network (GAN) methods can be utilized to handle imbalance dataset by generating more data to train deep neural networks.

REFERENCES

[1] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*. Vol. 1. MIT press Cambridge.

[2] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter (SIGKDD Explor. Newsl.)* 11, 1 (2009), 10–18.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 770–778.

[4] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2261–2269.

[5] Yang Liu, Zhonglei Gu, and William K Cheung. 2017. HKBU at MediaEval 2017 Medico: Medical multimedia task. In *Working Notes Proceedings of the MediaEval 2017 Workshop (MediaEval 2017)*.

[6] Mathias Lux, Michael Riegler, Pål Halvorsen, Konstantin Pogorelov, and Nektarios Anagnostopoulos. 2016. LIRE: open source visual information retrieval. In *Proceedings of the 7th International Conference on Multimedia Systems (MMSys)*. ACM, 30.

[7] Syed Sadiq Ali Naqvi, Shees Nadeem, Muhammad Zaid, and Muhammad Atif Tahir. 2017. Ensemble of Texture Features for Finding Abnormalities in the Gastro-Intestinal Tract. *Working Notes Proceedings of the MediaEval 2017 Workshop (MediaEval 2017)*.

[8] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *Proceedings of 31st Conference on Neural Information Processing Systems (NIPS)*.

[9] Stefan Petscharnig and Klaus Schöffmann. 2018. Learning laparoscopic video shot classification for gynecological surgery. *An International Journal of Multimedia Tools and Applications* 77, 7 (2018), 8061–8079.

[10] Stefan Petscharnig, Klaus Schöffmann, and Mathias Lux. 2017. An Inception-like CNN Architecture for GI Disease and Anatomical Landmark Classification. In *Working Notes Proceedings of the MediaEval 2017 Workshop (MediaEval 2017)*.

[11] Konstantin Pogorelov, Kristin Ranheim Randel, Thomas de Lange, Sigrun Losada Eskeland, Carsten Griwodz, Dag Johansen, Concetto Spampinato, Mario Taschwer, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. 2017. Nerthus: A Bowel Preparation Quality Video Dataset. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSYS)*. ACM, 170–174.

[12] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, and others. 2017. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSYS)*. ACM, 164–169.

[13] Konstantin Pogorelov, Michael Riegler, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Carsten Griwodz, Peter Thelin Schmidt, and Pål Halvorsen. 2017. Efficient disease detection in gastrointestinal videos–global features versus neural networks. *An International Journal Multimedia Tools and Applications* 76, 21 (2017), 22493–22525.

[14] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Carsten Griwodz, Thomas de Lange, Kristin Ranheim Randel, Sigrun Eskeland, Dang Nguyen, Duc Tien, Olga Ostroukhova, and others. 2017. A comparison of deep learning with global features for gastrointestinal disease detection. In *Working Notes Proceedings of the MediaEval 2017 Workshop (MediaEval 2017)*.

[15] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Thomas De Lange, Kristin Ranheim Randel, Duc-Tien Dang-Nguyen, Mathias Lux, and Olga Ostroukhova. 2018. Medico Multimedia Task at MediaEval 2018. In *Working Notes Proceedings of the MediaEval 2018 Workshop*.

[16] Michael Riegler, Konstantin Pogorelov, Sigrun Losada Eskeland, Peter Thelin Schmidt, Zeno Albisser, Dag Johansen, Carsten Griwodz, Pål Halvorsen, and Thomas De Lange. 2017. From annotation to computer-aided diagnosis: Detailed evaluation of a medical multimedia system. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 13, 3 (2017), 26.

[17] Michael Riegler, Konstantin Pogorelov, Pål Halvorsen, Carsten Griwodz, Thomas Lange, Kristin Ranheim Randel, Sigrun Eskeland, Dang Nguyen, Duc Tien, Mathias Lux, and others. 2017. Multimedia for medicine: the medico Task at mediaEval 2017. In *Working Notes Proceedings of the MediaEval 2017 Workshop (MediaEval 2017)*.

[18] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and others. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* (2015).

A.19 Paper XIX : Artificial Intelligence in Medicine: Gastroenterology

Authors: I. Strümke, S. A. Hicks, V. Thambawita, D. Jha, S. Parasa, M. A. Riegler, and P. Halvorsen

Abstract: The holy grail in endoscopy examinations has for a long time been assisted diagnosis using artificial intelligence. Recent developments in computer hardware are now enabling technology to equip clinicians with promising tools for computer-assisted diagnosis (CAD) systems. However, creating viable models or architectures, training them, and assessing their ability to diagnose at a human level, are complicated tasks. This is currently an active area of research, and many promising methods have been proposed. In this chapter, we give an overview of the topic. This includes a description of current medical challenges followed by a description of the most commonly used methods in the field. We also present example results from research targeting some of these challenges, and a discussion on open issues and ongoing work is provided. Hopefully, this will inspire and enable readers to future develop CAD systems for gastroenterology.

Published: Nature Springer, 2021

Candidate contributions: D. Jha contributed to drafting the manuscript. He wrote a few sections of the manuscript. Additionally, he prepared the tables presented in the manuscript and contributed to the revision procedure.

Thesis objectives: Objective II, Objective III

Artificial Intelligence in Medicine: Gastroenterology

Inga Strümke[1], Steven A. Hicks[1,2], Vajira Thambawita[1,2], Debesh Jha[1,3], Sravanthi Parasa[4], Michael A. Riegler[1], and Pål Halvorsen[1,2]

[1] SimulaMet, Norway
[2] Department of Computer Science, Oslo Metropolitan University, Norway
[3] Department of Computer Science, UIT The Arctic University of Norway
[4] Department of Gastroenterology, Swedish Medical Group, WA, USA

Abstract. The holy grail in endoscopy examinations has for a long time been assisted diagnosis using Artificial Intelligence (AI). Recent developments in computer hardware are now enabling technology to equip clinicians with promising tools for computer-assisted diagnosis (CAD) systems. However, creating viable models or architectures, training them, and assessing their ability to diagnose at a human level, are complicated tasks. This is currently an active area of research, and many promising methods have been proposed. In this chapter, we give an overview of the topic. This includes a description of current medical challenges followed by a description of the most commonly used methods in the field. We also present example results from research targeting some of these challenges, and a discussion on open issues and ongoing work is provided. Hopefully, this will inspire and enable readers to future develop CAD systems for gastroenterology.

Keywords: Gastrointestinal endoscopy · Artificial Intelligence · Neural Networks · Hand-crafted features · Anomaly detection · Semantic segmentation · Performance

1 Introduction

Numerous abnormal mucosal findings, ranging from minor annoyances to highly lethal diseases, can be found in the human Gastrointestinal (GI) tract. For example, according to the International Agency for Research on Cancer, about 3.5 million luminal GI (esophageal, stomach, colorectal) cancers are detected yearly in the world [41]. These cancers represent a substantial health challenge for society, with a mortality rate of about $63 - 65\%$, resulting in around 2.2 million deaths per year [19, 41]. Overall, Colorectal cancer (CRC) is the third most common cause of cancer mortality for women and men combined [104], and the other most frequently occurring GI cancers are stomach, liver, pancreatic and esophageal cancers [18].

For diagnosis and treatment of GI diseases, GI endoscopy is the gold-standard procedure used to examine the tract for anomalies, and to a certain extent, the

GI diseases may be prevented by improved endoscopic performance and high-quality systematic screening in high incidence areas [19]. However, despite the substantial technical improvement of endoscopes over the last two decades, a major limitation of the endoscopic examinations is the endoscope operator variation, depending on the procedural skill, perceptual factors, personality characteristics, experience, knowledge, and attitude deficits [34]. This translates to a substantial inter-observer variation in the detection and assessment of mucosal lesions [64, 108]. This causes, for example, an average 20% polyp miss-rate during colonoscopies [52]. All these factors could potentially, to some extent, be alleviated by substantial educational efforts, but not eliminated [88].

In this context, assisted diagnosis using computers has for a long time been a holy grail. Developments in computer hardware have enabled computationally demanding yet promising technologies like AI, more specifically its sub-field Machine Learning (ML), to provide the clinicians with potentially highly accurate and efficient Computer aided diagnosis (CAD) systems, giving healthcare professionals the tools needed to provide quality care at a large scale [102, 86]. At its core, machine learning involves using algorithms to parse data, learn from it, and then make predictions, in the medical domain this means detect, segment, assess or classify a disease. However, there exist several issues which need to be addressed, both for creating and improving automated diagnosis algorithms. Developing and assessing a computer's ability to diagnose at a human level are complicated tasks, and a potential success depends on various factors which goes beyond simply determining the accuracy of an algorithm. These challenges have been an active area of research for about a decade, and a large number of promising results have been published.

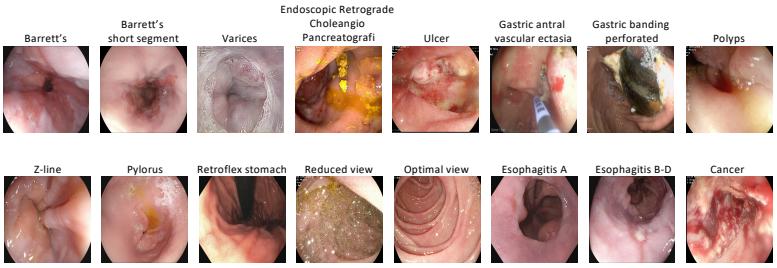
In this chapter, we describe current challenges on the way towards effective computer-based digital assistant systems. In particular, we focus on GI endoscopy. We provide examples of proposed methods and tools employing various techniques, identify current challenges and give hints for future development and assessment of CAD systems.

2 GI Endoscopy

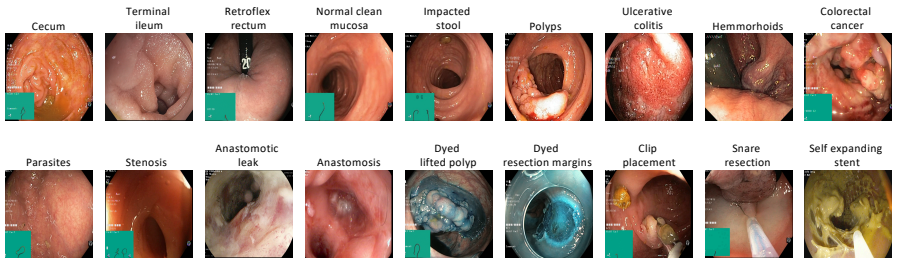
To examine the esophagus, stomach, duodenum (upper GI), and the large bowel and rectum (lower GI), a long, flexible tube is inserted into the mouth and rectum, respectively. A tiny video camera at the tip of the tube allows the doctor to view inside of the GI tract in real-time, where findings, as depicted in figures 1a and 1b can be found.

The small bowel is, due to its anatomical location, less accessible for inspection by such flexible endoscopes. To easier access these areas of the GI tract, Video Capsule Endoscopy (VCE) [22] has been introduced as an alternative examination method [25]. A VCE consists of a small capsule containing one or more wide-angle cameras. The capsule is swallowed by the patient, and it captures a video as it moves through the GI tract. The video is extracted, and a

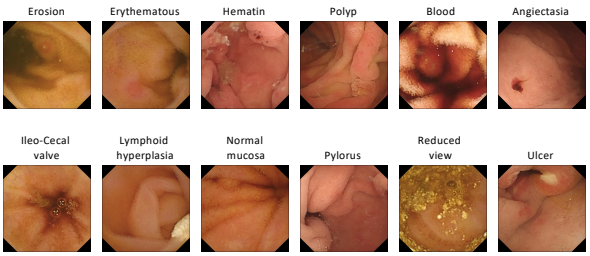
medical expert assesses it in a potentially tedious and time-consuming process after the procedure, searching for findings like the ones shown in figure 1c.



(a) Upper GI tract, esophagus, stomach (Gastroscopy).



(b) Lower GI tract, large bowel (Colonoscopy).



(c) Lower GI tract, small bowel (Capsule endoscopy).

Fig. 1: Examples of various findings in the GI tract including anatomical landmarks, pathological findings, normal mucosa, therapeutic interventions and medical instruments [14, 96].

Even though these examination procedures allow clinicians to detect GI anomalies, there is still ample scope for improvements. Looking at the possible findings depicted in figure 1, it is obvious that it can be hard to detect and classify the various anomalies potentially found in the various parts of the

GI tract, either live during a gastroscopy or colonoscopy, or in a post-analysis of the VCE video. Moreover, there are large operator variations and anomaly miss-rates reported for both regular endoscopies [34, 52, 64, 108] and capsule endoscopies [20, 88].

Hence, the hope is that automated analysis can *assist* medical experts in real-time anomaly detection, removing variations and increasing detection rates. Moreover, analysing hours of VCE video, there is also a large potential in saving medical expert time, by analysing the $4-12$ hour long videos in a few minutes by a fast computer, compared to the usual $45-60$ minutes error-prone, fast-forward analysis performed by medical personnel today. From an analysis point of view, there are two important requirements for such CAD systems:

1. *High detection or segmentation performance* in the analysis is important in order to address the large human miss-rates and variabilities. It is often measured in terms of metrics like precision, sensitivity (recall), specificity, accuracy, F1 score, Matthews correlation coefficient (MCC) or similar [98]. This requirement aims at finding all anomalies correctly, i.e., detecting all findings without false positives or negatives. A more detailed discussion on metrics is given in section 5.3.
2. An often neglected requirement is *fast processing* in order to give real-time feedback during the endoscopy examination, or in the case of VCE, higher scale of the analysis and a faster feedback on the same amount of processing resources.

Furthermore, in order to be deployable in a clinical environment, all components need to be integrated in a pipeline capturing videos or frames from the endoscopy equipment, via an automatic analysis, to give the clinicians a visual feedback (and potentially also assisting in generating an examination report according to medical standards). The system must also be easily integrated into and usable with the current examination procedures, and of course, the various components must meet the medical privacy and security regulations.

3 Existing Methods

As mentioned above, a large number of algorithms and models for automated analysis of GI video and images have already been proposed. In this respect, when we discuss CAD systems for the GI tract today, people often interchangeably talk about detection, localization and segmentation. Here, we therefore first try to distinguish between the terms as follows:

- *Detection* is the operation of detecting whether an image belongs to a certain classification or not. This can be a binary "yes or no" for questions whether the image or video frame contains a polyp or not. It also includes systems that *classifies* the input into multiple classes.
- *Localization* is to point into the image where the object is located, e.g., using some type of point markers or making a bounding box around the object of interest.

– *Segmentation* is yet another step further where one determines pixel-wise whether the pixel belongs to a finding or not, e.g., generating an exact segmentation mask of the finding.

Figure 2 shows an example of detection, localization and detection. As localization is often mixed into both detection and segmentation, we here focus on detection and segmentation.

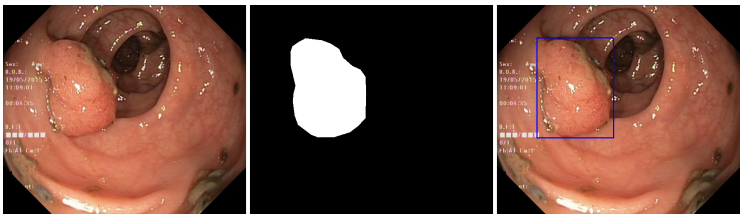


Fig. 2: Various ways of indicating a finding (*left*: detection "just" showing the image; *center*: full segmentation mask showing in white all pixels part of the finding; and *right*: bounding box making a rectangle around the finding)

3.1 Hand-crafted-feature-based approaches

Automatic detection of GI anomalies has been a topic of research long before the success of AI and deep neural networks, using what is nowadays often called traditional computer vision and ML methods, as found in libraries such as *OpenCV* [16] and *LIRE* [69]. Already in 1998, Krishnan et al. [59] proposed detecting polyps using shape-features in a curvature analysis. In the subsequent decade, various approaches using a mix of shape, edge, texture, and color features appeared. For example, Alexandre et al. [2] detected polyps using a support vector machine (SVM) on color patterns. Further, using SVMs, Ameling et al. [5] combined texture and colors, and Park et al. [75] used shape and texture features in a conditional random field classifier.

Two more recent approaches using hand-crafted techniques are Polyp-Alert [111] and EIR [85], where the authors also measured analysis time, with the goal of being able to give real-time feedback during the examination. The Polyp-Alert [111] system combines edge and texture features. The polyp edge detection algorithm mainly relies on edge features obtained from the part-based multi-derivative edge cross-section profile [110]. The EIR [85] system combines a content-based similarity search with statistical classifiers from the training data. A large number of image features are tested [87], ending up with a combination of the joint composite descriptor feature and the Tamura features, due to a good trade-off between the precision and sensitivity (recall), and the speed of the algorithm. A search-based classifier is then used to determine if an image contains a finding of a certain class.

A detailed overview containing earlier example approaches can be found in [111, 85]. However, lately, deep learning approaches have outperformed these hand-crafted approaches and replaced them entirely.

3.2 Deep learning-based approaches

Already in 2001, Karkanis et al. [53] aimed for the detection of lesions in endoscopic video using textural descriptors on the wavelet domain supported by artificial neural network architectures, albeit not using deep architectures. Such early approaches were tested on tiny data sets, in this case 8 images [53]. More recent approaches are usually based on deep learning architectures where Convolutional Neural Networks (CNNs) are clearly the most popular ones.

Where hand-crafted features rely on extracting predefined properties of an image, such as color, texture, or shape, CNNs are neural network architectures using convolutions and pooling operations to automatically learn which features are most relevant. CNNs perform well on many different tasks like image classification, object detection in images, and image generation [56]. Although they are mostly used for image analysis, they have also proven useful in time-series research and video analysis. In medicine, architectures like U-Net [89] have shown promising results in areas like cardiology, colonoscopy, and radiology [119, 122, 74]. This also includes gastroenterology, where CNNs are currently state-of-the-art for analysing colonoscopy videos. The most common application is the detection and segmentation of polyps, where many CNN-based approaches have shown excellent results [17, 113, 50]. These approaches have expanded to other findings as well, like detecting and segmenting ulcers [31]. Furthermore, due to limited access to medical image and video data, most approaches use transfer learning. In transfer learning, pre-trained models are used as a starting point, and refined for the given data set by retraining with some layers trainable and some frozen [82].

An automated CAD system for the GI endoscopic image segmentation is a step further than providing "just" detection of anomalies. A predicted segmentation mask (see figure 2) can help point out the area of interest in the images (frames) that need to be further examined. However, making such per-pixel predictions is also a more complex task. In this respect, there has been a considerable amount of work done so far, especially targeting polyps [109, 45, 50, 47, 48, 71, 32, 100, 81], artefacts [3], and endoscopic instruments [90]. In general, CNN-based approaches perform well with the larger polyps. However, still the major challenges issues in the field are related to adenomatous polyps or small and flat polyps. Recent studies are targeting smaller polyps [63, 50], however, it is yet an open-challenge to solve.

3.3 Unsupervised and semi-supervised approaches

The above presented approaches fall into the category of supervised learning, meaning that we train the models on a data set with an existing ground truth.

In this section, we give a glance at newly emerging unsupervised and semi-supervised methods.

Generative Adversarial Networks (GANs), which were introduced by Goodfellow et al. [30] in 2014, are becoming increasingly popular in the medical domain for generating synthetic data. Different advancements to the original GAN architecture, such as conditional GAN [72], pix2pix [43], CycleGAN [123], StyleGANs [54, 55], to mention a few, present different methods, ranging from domain transformation to high definition image generation. ML researchers in the medical domain can use GAN models to generate synthetic data to tackle challenges related to privacy, data deficiency, and data annotation. For example, Younghak et al. [93] use a conditional GAN architecture to generate synthetic polyp images to improve the performance of a deep learning system detecting polyps in the colon. This methodology is still in its early stages, and it has yet to be shown to which extent generated data can replace real data and help to improve performance and shareability.

Another emerging method in the field of medical image analysis, is semi-supervised learning. Here, the goal is to learn from a small set of labelled data combined with a larger amount of unlabelled data. Examples include [116, 7, 70, 67]. These models produce promising results, and could also help overcome the challenge of insufficient labelled data faced by many data-hungry methods. However, these approaches still struggle with challenges such as low accuracy and high entropy during early stages of the training process. The models are also regularized towards high entropy predictions, making it hard to achieve a high accuracy [120, 117]. It will be interesting to see whether these challenges can be overcome, and how useful the results will prove to be in the medical domain.

4 Example results

High detection or segmentation rates are important in order to be clinically relevant, and the typical way the performance is compared. However, due to factors like different data sets and different equipment, the pure numbers cannot be directly compared. Still, to give some indications of the state-of-the-art performance, we give a set of, by far from complete, examples using standard metrics like precision, sensitivity (recall), specificity, accuracy, F1 score and MCC for detection; and Dice similarity coefficient (DSC), Intersection over Union (IoU), precision and sensitivity for segmentation. A substantial overview of existing approaches can be found in [61], containing 138 different studies. An explanation of the different metrics is given Table 1 and further discussed in section 5.3. Another source for exploring and comparing different approaches, are the popular GI detection, classification and segmentation challenges discussed in section 5.6.

A selection of performance examples are given in Table 2. Looking at the numbers, we see that in the specific tested cases, the computer should be at the level of the best experts with scores above 90%, i.e., potentially being a helpful digital assistant during a GI endoscopy examination. Likewise, example

Table 1: List of commonly used metrics. To define each metric, TP, FP, TN, and FN represent true positives, false positives, true negatives, and false negatives, respectively.

Formula	**Description**
accuracy $= \frac{TP+TN}{TP+FP+TN+FN}$	Rate of correct classification. Ratio between correctly classified samples and all samples.
precision $= \frac{TP}{TP+FP}$	Proportion of retrieved samples which are relevant. Ratio between correctly classified positive samples and all samples classified as positive.
sensitivity (also known as recall) $= \frac{TP}{TP+FN}$	Proportion of relevant samples which are retrieved. Ratio between correctly classified positive samples and all positive samples.
specificity $= \frac{TN}{TN+FP}$	Negative class sensitivity. Ratio between correctly classified negative samples and all negative samples.
F1 $= \frac{2\times TP}{2\times TP+FP+FN}$	Harmonic mean of the precision and sensitivity (recall).
MCC $= \frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$	Pearson's correlation coefficient [23] for binary classification.
IoU (also known as Jaccard) $= \frac{TP}{TP+FP+FN}$	Similarity between sets from the size of the intersection divided by the size of the union.
DSC $= \frac{2TP}{2TP+FP+FN}$	Quotient of similarity of two sets. Semi-metric as it doesn't satisfy the triangle inequality. Related to the IoU via $\frac{S}{2-S}$.

results for lesion segmentation are provided in Table 3, and the numbers are again encouraging in terms of proving that the used models could be of use in a medical setting. However, while the results achieved are promising, there are still several open challenges, including generalizability, overfitting, cross data set testing and explainability of the results. Moreover, as indicated in the Tables, hardly any existing research report the speed of the system, meaning that it is hard to assess the system's capability to provide a live analysis in the clinic.

5 Open issues and ongoing research

Despite impressive results presented in many of the published papers, even exceeding what are reported as average detection rates from clinicians, there are still challenges and open issues. First, for example, Thambawita et al. [98] presented the issue of overfitting to specific data sets and a lack of generalizability. This means that a model that performs well on one data set may not perform at all on another. Furthermore, like other deep neural networks, CNNs are black boxes, and it is not easy to understand why one input gives a particular result. There is also a lack of large open data sets that contain annotations for un-

Table 2: Examples of **detection** performance from different approaches. The results show promising performance with numbers above 90%. Unfortunately, speed is not commonly reported.

Paper / system	Data set used	Sensitivity (recall)	Specificity	Accuracy	Precision	F1	MCC	Speed (fps)
Boughorbel [17]	MICCAI-challenge data sets	86.3	-	-	73.6	-	-	-
Kundu [60]	30 Own data set	95.2	98.3	97.9	88.4	-	-	-
Cho [21]	Seoul National University Hospital	>87	-	>93	-	-	-	
Ghosh [29]	VCE videos data set	99.4	99.2	97.9	95.8	-	-	-
Bell [8]	CTC generating 4000 images per patients	89.8	75.5	-	-	-	-	-
Pogorelov [76]	Kvasir	83.9	98.5	97.2	84.1	85.6	82.8	46
Billah [13]	Colonoscopy & Endoscopy vision data set	98.7	98.2	98.3	-	-	-	-
Thambawita [99]	Kvasir	95.8	99.7	95.8	95.9	95.8	95.3	29

Table 3: Some examples of different **segmentation** approaches applied to different data sets. We can clearly see that the performance overall is quite promising (with all metrics in the range of 70 to 95. Speed is unfortunately not commonly reported.

Paper / system	Data set used	DSC	IoU (Jaccard)	Sensitivity (recall)	Precision	Speed (fps)
U-Net [89]	MICCAI-PhC-U373	-	92.0	-	-	-
PraNet [26]	CVC-ClinicDB	89.9	84.0	-	-	-
PolypSegNet [71]	CVC-ClinicDB	91.5	86.2	91.1	96.2	-
ResUNet++ [50]	CVC-ClinicDB	79.6	79.6	70.2	87.9	-
PraNet [26]	Kvasir-SEG	89.8	84.0	-	-	-
PolypSegNet [71]	Kvasir-SEG	88.7	82.5	84.5	91.7	-
ResUNet++ [50]	Kvasir-SEG	81.3	79.3	70.6	87.7	-
Double-UNet [45]	CVC-ClinicDB	92.4	86.1	84.6	96.0	-

common abnormalities and rarely documented findings to support data-hungry algorithms like CNNs. Here, we elaborate on a few of these open issues.

5.1 Limited data availability

Available medical data is scarce. However, modern deep learning approaches usually require a lot of data to perform well, and often, the more variations in the data, the better the model gets, especially for supervised learning models. Table 4 shows the data sets available in the field of GI endoscopy. Evidently, the number of images used for training and testing is small when compared to

Table 4: Summary of available endoscopic data sets. A further discussion about data sets are found in [14, 96].

Data set	Findings	Location	#Images	#Videos	Bounding box?	Segmentation mask?	Input size	VCE data?	Endoscopic device
Kvasir [77]	various	⇑⇓	8,000	-	-	-	variable	-	†
Nerthus [78]	stool (cleanness)	⇓	5,525	-	-	-	720 × 576	-	†
HyperKvasir [14]	various	⇑⇓	110,079	373	-	-	variable	-	†
KvasirInstrument [46]	instruments	⇓	590	-	✓	✓	variable	-	†
Kvasir-SEG [49]	polyps	⇓	1,000	-	✓	✓	variable	-	†
ASU-Mayo [97]	polyps	⇓	18,781	-	-	-	variable	-	-
CVC-ClinicDB [10]	polyps	⇓	612	-	-	-	384 × 288	-	‡◇
CVC-ColonDB [11]	polyps	⇓	380	-	-	-	574 × 500	-	-
ETIS Larib Polyp DB [95]	polyps	⇓	196	-	-	-	1225 × 966	-	‡◇
SUN Colonoscopy Video DB [73]	polyps	⇓	158,690	-	✓	-	1080 × 1240	-	?
CVCVideoClinicDB [6, 12]	polyps	⇓	11,954	-	-	-	384 × 288	-	-
CAD-CAP [65]	various		25,000	-	-	-	-	✓	-
KID [58]	various		2,371	47	-	-	-	✓	-
KvasirCapsule [96]	various	⇓	4,820,739	118	✓	-	variable	✓	•

Location: ⇑ = upper GI ⇓ = lower GI

Device: † = ScopeGuide, Olympus ‡ = Olympus Q160ALandQ165L
◇ = Exera IIvideoprocessor • = Olympus EC-S10 endocapsule

the data set from the natural images. This is because it is difficult to obtain data from the medical domain. The data is often protected and unavailable due to legal restrictions and lack of medical personnel for the tedious process of manually extracting and labeling training data. This calls either for better data sharing processes and culture, or methods more capable of handling small amounts of data.

This gives rise to several basic challenges: The amount of data is too small to train a robust model, and the presented results might appear deceivingly good due to overfitting. Moreover, it is hard to compare results if all experiments are performed on different data, and practically impossible to reproduce them. Thus, it is almost impossible to conclude whether one model is better than another. We must therefore aim for more and open data sets. Table 4 contains an overview of know available data sets at the time of writing, making a good starting point for future experiments. Still, more data is needed, especially data containing pathological outcomes.

5.2 Generalizability

One of the open issues in the field is the GI endoscopy is the generalizability of ML models, i.e., their ability to perform well on previously unseen data regardless of source, equipment, etc. Such data can be from either the same distribution as

the model was trained on, or from a different distribution. Which of the two a new data sample represents, is not always clear [103, 101]. Although some recent studies address generalizability of ML models for polyp classification [112, 45], this must be addressed for any model or system to be deployed into clinical practice.

Evaluating whether a model is reliable for real world use also requires cross data set testing, to avoid accepting a model which coincidentally works well on one specific set of data. The model developers should in general not have access to the final test data, to avoid bias during testing and development. This process, known as data blinding, is an important tool in many fields of research, including medicine [80]. Ideally, the model should be tested for robustness on data collected separately from the data used during model development and testing.

Furthermore, distinction should be made between data annotated by medical experts, referred to as *soft ground truth*, and data labelled based on a medical test, referred to as *hard ground truth*, e.g., pathological examination of a polyp. The quality of soft ground truth data is limited by how well the medical annotator is trained, and such data is most useful for training models intended to automate processes. On the other hand, data with hard ground truth labels can also be used for automating processes, with the added benefit of avoiding annotator error or bias into the model, but it can furthermore be used for obtaining new knowledge. Note that while, as mentioned above, annotating each image is time consuming, collecting hard ground truth data is even more demanding, resulting in a scarcity of such data sets.

In current endoscopy practices, different hospitals use different endoscope system for diagnosis and therapy. The most common globally available endoscope systems are Olympus (Japan), Pentax 90i series (Japan), Fujinon (Japan), and Karl Storz (Germany) [57]. Moreover, different medical institutes have different protocols. Therefore, designing generalizable CAD systems is essential for performing well on a variety of institutes. Such systems should always be tested on several data sets. Discussions regarding challenges and advantages associated with cross-dataset testing can be found in [98].

5.3 Metrics and evaluation

Evaluating performance is an important step when creating models for clinical use, and depends strongly on the choice of metric. As shown in Table 1, commonly used metrics are precision, sensitivity (recall), specificity, accuracy and F1 score. Some papers also report AUROC (area under the receiver operating characteristics). There are several reasons for going beyond the aforementioned metrics [98]. One challenge frequently encountered in association with medical data sets, is their tendency to be imbalanced between classes, often having far more normal images than images with lesions. Because of this, certain metrics can provide an overly optimistic impression of the actual performance. For instance, a binary classifier can achieve a high accuracy on a data set containing

few negative instances, by assigning all instances to the positive class. The AUROC is also known to be deceptive for imbalanced classification [91]. In such cases, the correlation coefficient between the true and predicted classes can be more informative [15], although no single metric is universally informative or suited for any imbalanced data problem. Moreover, for detection purposes it is also a question whether one report per-frame performance, i.e., giving a decision for every frame in the video, or per-lesion, i.e., giving a correct prediction for at least one of the frames in the video sequence. Looking at the results from a technical point of view, a per-frame analysis of often desired, but from the medical point of view, a per-lesion analysis is often sufficient to notify the clinician of the finding once.

For segmentation performance, commonly used metrics are DSC and the IoU, also known as the Jaccard index. In clinical use, medical experts are usually interested in pixel-wise detail information about the potential lesion. DSC and IoU can be used to compare the pixel-wise similarity between the predicted segmentation maps and the ground truth. In addition, precision and sensitivity are used to evaluate under-segmentation or over-segmentation, where under-segmentation implies that the model predicts less relevant content in some portion of the image compared to the ground truth, and over-segmentation that the predicted image covers more pixels than the ground truth.

As observed in Tables 2 and 3, little research has until now focused on the required real-time capabilities in order to provide live feedback to clinicians during the endoscopy examinations. However, there seems to be reported systems that analyse data faster than the frame-rate threshold, and it has also been given attention in some of the arranged competitions (see section 5.6). Nonetheless, it is often a trade-off between speed (model complexity) and detection performance, indicating that this is still an important issue in future research and development of CAD systems.

5.4 Automatic report generation

After the endoscopist finishes an endoscopy, a high-quality report should be generated. This often a time-consuming process, where research shows that approximately one-sixth of U.S. physicians working time is spent on administrative tasks, taking time away from direct-patient care and lessening job satisfaction [115]. Moreover, there are large variations in endoscopists' interpretations of findings as well as reporting styles. This can, and often does, lead to inconsistencies in the final decision [37]. Hence, automated report generation could both save clinical time and help standardize endoscopy reports, and recent development in natural language processing is expected to open up new possibilities in automatic report generation [86].

A method proposed by Jing et. al. [51] uses neural image captioning to create reports from x-ray images. In [121], images are analysed by a neural network, and example images of findings similar to the one at hand and attention maps are combined to reports. Most approaches focus on image analysis as a basis,

and combine this with additional information [118, 33, 24]. This of course depends on access to a database containing correct information which can be used in combination with the images. A significant challenge is different reporting standards between countries or even hospitals, making it practically impossible to create a widely adoptable software.

However, for medical experts, automatic text creation might not even be the most crucial feature of such a software: A more important aspect is their ability to understand the reasoning and decision of the underlying model, enabling them to include it in their assessment. This is discussed in the next section.

5.5 Explainability

A well-known challenge associated with deep learning based CAD systems, is limited explainability due to their inherent complexity [4]. This property has caused their notoriety as black boxes whose decision-making processes are unknown, especially to end-users [35]. The need for understanding and explaining how the systems work and which roles the different data features play in the decisions, addresses different needs in the different stages of the system's development and use. The developer of the system needs to understand how data and methods are working together, as understanding and interpretability of the output helps to determine errors in the data as well as enabling targeted failure analysis. Particularly, in the context of this AIM, the medical experts require an explanation of the system's decision to assure that it concurs with the relevant medical knowledge.

Deep learning based systems, such as CNNs, have no inherent ways of providing explanations, meaning that they must either be extended to contain explanation generators, or explanations must be obtained post hoc [38, 1, 39]. A brief overview of approaches to model explanations is shown in Table 5. Models can be designed to provide justifications for their decisions as an additional task, e.g., via a text justification generator as part of the model architecture [62]. Given a model without such a design, different approaches are available: Those which explain the properties of the decision making system itself, and those which treat the system like a black box and provide explanations based on its emergent behaviour, referred to as model dependent and model agnostic approaches, respectively. One example of the former is displaying the values of the Deep Neural Network (DNN)'s internal parameters as a heat map superimposed on the classification instance [35]. Interpreted correctly, this can provide an understanding of the system's internal decision making process. Such an approach can also be extended to include information regarding the system induced decision uncertainty (meaning the part of the uncertainty not associated with the data collection and selection process), see [114].

Among the model-agnostic methods, the explanation concept LIME (Locally Interpretable Model-agnostic Explanations) approximates the black-box model using an interpretable model, such as a linear model, decision tree, or falling rule list [84]. This is done in the neighborhood of the instance to explain, making the resulting explanation a local one, given that it applies to a single outcome and

is based on the particular instance's characteristics, as is also the case for the aforementioned model-dependent explanations.

In contrast, global explanations capture and explain the model at large, such as feature importance ranking. One class of methods capable of producing global explanations, are those based on the game-theoretic concept of Shapley values [92], which are currently enjoying a surge of interest in the statistics and machine learning literature [66, 44, 40, 27]. Shapley values are obtained by evaluating the model using all possible combinations of the data features. Hence, the computational complexity increases with the number of features $|f|$ as $2^{|f|}$, and the calculation involves re-training the model for each subset of features. The latter is problematic as re-training would result in different model parameters, highlighting that Shapley values are merely model agnostic, not *independent*. The widely used SHAP (SHapley Additive exPlanations) package [68] circumvents these challenges in different ways for various model architectures, by calculating approximate values using background samples from the data, and for deep architectures using a similar approach as the per node attribution rules from DeepLIFT [94]. The Shapley decomposition can be computed both globally and locally, and can be formulated [68] as a special case of LIME. Shapley values can also be used to obtain model-independent explanations [28].

Table 5: The different model explanation approaches regarding when they are applied: During the model development (in-model) or after the model is finished (post-model). Explanation methods provide insight into model behaviour either locally (around a particular prediction) or globally.

Category		Description	Ex.
In-model		Justification text generator as part of model architecture	[62]
Post-model	Model dependent	GradCam: Display DNN activations on image	[35]
	Model agnostic	LIME: Yields a locally interpretable model approximating the full model	[84]
		SHAP: Shapley decomposition of a conditional expectation function of the full model	[68]
	Model independent	Global non-parametric Shapley decomposition	[28]

5.6 Competitions and challenges

There have been a series of different challenges related to automatic analysis of endoscopy data [9, 79, 36], where CNN-based approaches have been the top-performing methods for the last few years. The various tasks given have been to benchmark and develop automated systems to accurately detect, localize, and segment the abnormalities inside the GI tract. These challenges targeted different tasks from detection, localization, and segmentation of GI anomalies, colorectal polyps to artifacts presence in the GI tract (see Table 6). These regular competitions can help the research community in the field to find to find common

standards for evaluating models, benchmarking state-of-the-art methods and tools, and finding new directions to bring the field forward together.

Table 6: List of GI detection, classification and segmentation challenge examples.

Challenge name	URL
MICCAI 2015 Endoscopic Vision	https://polyp.grand-challenge.org/databases/
Medico 2017	http://www.multimediaeval.org/mediaeval2017/medico/
Medico 2018	http://www.multimediaeval.org/mediaeval2018/medico/
GIANA 2018	https://giana.grand-challenge.org/Home/
EAD 2019	https://ead2019.grand-challenge.org/
Biomedia 2019	https://github.com/kelkalot/biomedia-2019
Medico 2020	https://multimediaeval.github.io/editions/2020/tasks/medico/
EndoTect 2020	https://github.com/simula/icpr-endotect-2020
EDD Challenge 2020	https://edd2020.grand-challenge.org/
EndoCV 2020	https://endocv.grand-challenge.org

5.7 Clinical verification and emerging commercial systems

Many research groups have presented promising research results and good performance indicators, and several AI-based commercial systems have emerged, some of which are listed in Table 7. The status of these are mostly unknown, but for example, the GI Genius system is CE marked, but still lacks US Food and Drug Administration (FDA) approval, and EndoBRAIN-EYE is approved only in Japan. For CAD systems to be deployed for real-time examinations in clinical examination rooms, or to be used for VCE data post analysis, clinical verification is strictly necessary. Still, at the time of writing, such studies are very limited. In August 2020, Repici et al. [83] presented a randomized multi-center trial, concluding that the AI-based CAD increases the adenoma detection rate (ADR), i.e., the percentage of patients with at least one histologically proven adenoma or carcinoma, demonstrating the potential of such systems. They examined 685 patients: 341 patients using the CAD system and 344 patients using only the traditional manual examination. The system achieved an ADR of 54.8%, and the control group 40.4%. This demonstrates that AI-based systems can help detect adenomas, but that further improvements are required to increase detection rates, and to detect a larger number of sessile serrated lesions (at all). Considering the limitations of the study as well as the presented performance, it is clear that there are still improvements to made, and more clinical studies are in order.

Despite significant interest from the industry, proper standards regarding evaluation methods and reproducibility are widely lacking. In addition, industry applications seem not to have focused on model explainability or model output interpretability. These are all crucial ingredients of trustworthy applications, and industry development will hopefully follow current research trends and focus more on these in the future.

Table 7: Emerging commercial products

Product	Vendor	Year	URL
GI Genius AI	Medtronic/ Cosmo Pharma	2019	https://www.cosmopharma.com/products/gi-genius
EndoBRAIN-EYE	Cybernet	2020	https://www.cybernet.jp/english/documents/pdf/news/press/2020/20200129.pdf
CAD-Eye	Fujifilm	2020	https://www.fujifilm.eu/eu/cadeye
Ai4Gi	Ai4Gi	2016?	https://ai4gi.com
UltiVision	DocBot	2018	https://www.docbot.co/gastroenterology-and-health
DISCOVERY	Pentax	2020	https://www.pentaxmedical.com/pentax/en/95/2/DISCOVERY-new
ENDO-AID	Olympus	2020	https://www.olympus.no/medical/en/Products-and-Solutions/Products/Product/ENDO-AID.html
SOMA	Augere Medical	2018	https://augere.md

Finally, when a high-performing (research) prototype has been build and tested, meeting the requirements above, it must be approved for medical use. Robust evaluation of AI based software before implementation is needed to reduce patient and health system risk, establish trust to facilitate wide-spread adoption. The common term used for such products is AI based software as a medical device (SaMD). Regulators of the SaMD applications, including the FDA in the United States, have been guided by the Global Harmonization Task Force and International Medical Device Regulators Forum (IMDRF). The IMDRF has proposed 4 different risk categories for SaMD each with a different set of requirements for assessing scientific and clinical validity of the technology [42]. Within gastroenterology, CADe and CADx technologies have not yet been classified. The current FDA process for SaMD is derived from its approval process for medical devices and will be categorized into 3 risk categories: Classes I, II and III (highest risk) [105]. After risk classification, premarket submission as a 510(k) pathway or de novo pathway might be relevant to GI based AI technologies similar to Osteoidetect [107]. Moreover, given that the AI algorithms are rapidly iterative and continuously learning, it can pose a challenge to the current regulatory process. The FDA proposed a new system of regulation for AI technologies in its Digital Health Innovation Action Plan, focused on AI technologies that rely on continuous learning and adaptation [106]. Regulators around the world have also recognized the challenges involved with AI algorithms when applied to medicine and most countries have initiated efforts to develop policies tailored for SaMD. Many of them share the core principles of designation of risk, review clinical evidence to demonstrate efficacy and safety, practices to incorporate evolving AI systems.

6 Summary & Conclusions

In this work, we have introduced the application of automated data analysis for GI endoscopy, and presented an overview on detection and segmentation

based approaches to tackle challenges like large lesion miss-rates and inter-observer variability. Recent studies have shown that deep computer vision-based approaches seem to have the potential of improving the accuracy and overall performance in GI endoscopy by providing fully automated CAD systems acting as an additional digital eye. Nevertheless, there are still several open issues and challenges which need to be addressed before automatic analyses can be usefully integrated into clinical practice. These should be regarded as issues requiring research attention in the field.

Acknowledgements

This work is funded in part by the Research Council of Norway, project number 282315 (AutoCap).

References

[1] Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access 6, 52138–52160 (2018)

[2] Alexandre, L.A., Casteleiro, J., Nobreinst, N.: Polyp detection in endoscopic video using svms. In: Proceeding of Knowledge Discovery in Databases (PKDD). pp. 358–365. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)

[3] Ali, S., Zhou, F., Braden, B., Bailey, A., Yang, S., Cheng, G., Zhang, P., Li, X., Kayser, M., Soberanis-Mukul, R., Albarqouni, S., Wang, X., Wang, C., Watanabe, S., Oksuz, I., Ning, Q., Yang, S., Khan, M.A., Gao, X., Rittscher, J.: An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. Scientific Reports 10 (02 2020)

[4] Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I., Precise4Q consortium: Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC medical informatics and decision making 20(1), 310 (November 2020), https://europepmc.org/articles/PMC7706019

[5] Ameling, S., Wirth, S., Paulus, D., Lacey, G., Vilarino, F.: Texture-based polyp detection in colonoscopy. In: Bildverarbeitung für die Medizin, pp. 346–350. Springer (2009)

[6] Angermann, Q., Bernal, J., Sánchez-Montes, C., Hammami, M., Fernández-Esparrach, G., Dray, X., Romain, O., Sánchez, F.J., Histace, A.: Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis. In: Cardoso, M.J., Arbel, T., Luo, X., Wesarg, S., Reichl, T., González Ballester, M.Á., McLeod, J., Drechsler, K., Peters, T., Erdt, M., Mori, K., Linguraru, M.G., Uhl, A., Oyarzun Laura, C., Shekhar, R. (eds.) Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures. pp. 29–41. Springer International Publishing, Cham (2017)

[7] Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P.M., Rueckert, D.: Semi-supervised Learning for Network-Based Cardiac MR Image Segmentation. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 253–260. Lecture Notes in Computer Science, Springer International Publishing (2017)

[8] Bell, L.T., Gandhi, S.: A comparison of computer-assisted detection (CAD) programs for the identification of colorectal polyps: performance and sensitivity analysis, current limitations and practical tips for radiologists. Clinical Radiology pp. 1–8 (2018), https://doi.org/10.1016/j.crad.2018.02.009

[9] Bernal, J., Tajkbaksh, N., Sánchez, F.J., Matuszewski, B.J., Chen, H., Yu, L., Angermann, Q., Romain, O., Rustad, B., Balasingham, I., Pogorelov, K., Choi, S., Debard, Q., Maier-Hein, L., Speidel, S., Stoyanov, D., Brandao, P., Córdova, H., Sánchez-Montes, C., Gurudu, S.R., Fernández-Esparrach, G., Dray, X., Liang, J., Histace, A.: Comparative validation of polyp detection methods in video colonoscopy: Results from the miccai 2015 endoscopic vision challenge. IEEE Transactions on Medical Imaging 36(6), 1231–1249 (2017)

[10] Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computeri. Med. Imag. and Graph. 43, 99–111 (2015)

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

[12] Bernal, J.J., Histace, A., Masana, M., Angermann, Q., Sánchez-Montes, C., Rodriguez, C., Hammami, M., Garcia-Rodriguez, A., Córdova, H., Romain, O., Fernández-Esparrach, G., Dray, X., Sanchez, J.: Polyp Detection Benchmark in Colonoscopy Videos using GTCreator: A Novel Fully Configurable Tool for Easy and Fast Annotation of Image Databases. In: Proceedings of 32nd CARS conference. Berlin, Germany (Jun 2018)

[13] Billah, M., Waheed, S.: Gastrointestinal polyp detection in endoscopic images using an improved feature extraction method. Biomedical Engineering Letters 8(1), 69–75 (2018)

[14] Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., Johansen, D., Griwodz, C., Stensland, H.K., Garcia-Ceja, E., Schmidt, P.T., Hammer, H.L., Riegler, M.A., Halvorsen, P., de Lange, T.: HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Springer Nature Scientific Data 7, 283 (2020), https://doi.org/10.1038/s41597-020-00622-y

[15] Boughorbel, S., Jarray, F., El-Anbari, M.: Optimal classifier for imbalanced data using matthews correlation coefficient metric. PloS one 12(6), e0177678 (2017)

[16] Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000)
[17] Brandao, P., Mazomenos, E., Ciuti, G., Caliò, R., Bianchi, F., Menciassi, A., Dario, P., Koulaouzidis, A., Arezzo, A., Stoyanov, D.: Fully convolutional neural networks for polyp segmentation in colonoscopy. In: Medical Imaging 2017: Computer-Aided Diagnosis. vol. 10134, p. 101340F. International Society for Optics and Photonics (2017)
[18] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A.: Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians 68(6), 394–424 (2018)
[19] Brenner, H., Kloor, M., Pox, C.P.: Colorectal cancer. Lancet 383(9927), 1490–502 (2014)
[20] Cave, D.R., Hakimian, S., Patel, K.: Current controversies concerning capsule endoscopy. Digestive Diseases and Sciences 64(11), 3040–3047 (Nov 2019)
[21] Cho, M., Kim, J.H., Kong, H.J., Hong, K.S., Kim, S.: A novel summary report of colonoscopy: timeline visualization providing meaningful colonoscopy video information. International Journal of Colorectal Disease 33(5), 549–559 (2018)
[22] Costamagna, G., Shah, S.K., Riccioni, M.E., Foschia, F., Mutignani, M., Perri, V., Vecchioli, A., Brizi, M.G., Picciocchi, A., Marano, P.: A prospective trial comparing small bowel radiographs and video capsule endoscopy for suspected small bowel disease. Gastroenterology 123(4), 999–1005 (2002)
[23] Cramer, H.: Mathematical methods of statistics. Princeton University Press Princeton (1946)
[24] Daniels, Z.A., Metaxas, D.N.: Exploiting visual and report-based information for chest x-ray analysis by jointly learning visual classifiers and topic models. In: Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1270–1274. IEEE (2019)
[25] Enns, R.A., Hookey, L., Armstrong, D., Bernstein, C.N., Heitman, S.J., Teshima, C., Leontiadis, G.I., Tse, F., Sadowski, D.: Clinical practice guidelines for the use of video capsule endoscopy. Gastroenterology 152(3), 497–514 (2017)
[26] Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: PraNet: Parallel Reverse Attention Network for Polyp Segmentation. arXiv preprint arXiv:2006.11392 (2020)
[27] Frye, C., Rowat, C., Feige, I.: Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability (2020)
[28] Fryer, D., Strümke, I., Nguyen, H.: Explaining the data or explaining a model? shapley values that uncover non-linear dependencies. ArXiv abs/2007.06011 (2020)
[29] Ghosh, T., Fattah, S.A., Wahid, K.A.: CHOBS: Color Histogram of Block Statistics for Automatic Bleeding Detection in Wireless Capsule Endoscopy Video. IEEE Journal of Translational Engineering in Health and Medicine 6(May 2017) (2018)

[30] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)

[31] Goyal, M., Yap, M.H., Reeves, N.D., Rajbhandari, S., Spragg, J.: Fully convolutional networks for diabetic foot ulcer segmentation. In: Proceedings of the IEEE international conference on systems, man, and cybernetics (SMC). pp. 618–623 (2017)

[32] Guo, Y.B., Matuszewski, B.: GIANA Polyp Segmentation with Fully Convolutional Dilation Neural Networks. In: Proceedings of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. pp. 632–641 (2019)

[33] Han, Z., Wei, B., Leung, S., Chung, J., Li, S.: Towards automatic report generation in spine radiology using weakly supervised framework. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 185–193. Springer (2018)

[34] Hewett, D.G., Kahi, C.J., Rex, D.K.: Efficacy and effectiveness of colonoscopy: how do we bridge the gap? Gastrointestinal Endoscopy Clinics 20(4), 673–684 (2010)

[35] Hicks, S., Riegler, M., Pogorelov, K., Anonsen, K.V., de Lange, T., Johansen, D., Jeppsson, M., Ranheim Randel, K., Losada Eskeland, S., Halvorsen, P.: Dissecting deep neural networks for better medical image classification and classification understanding. In: Proceedings of IEEE International Symposium on Computer-Based Medical Systems (CBMS). pp. 363–368 (2018)

[36] Hicks, S., Petlund, A., de Lange, T., Schmidt, P., Halvorsen, P., Riegler, M., Smedsrud, P., Haugen, T., Randel, K., Pogorelov, K., Stensland, H., Dang Nguyen, D.T., Lux, M.: Acm multimedia biomedia 2019 grand challenge overview. In: Proceedings of the ACM International Conference on Multimedia (ACM MM). pp. 2563–2567 (2019)

[37] Hicks, S., Smedsrud, P., Riegler, M., de Lange, T., Petlund, A., Eskeland, S., Pogorelov, K., Schmidt, P., Halvorsen, P.: Deep learning for automatic generation of endoscopy reports. Gastrointestinal Endoscopy 89, AB77 (06 2019)

[38] Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable ai systems for the medical domain? arXiv preprint arXiv:1712.09923 (2017)

[39] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H.: Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9(4), e1312 (2019)

[40] Huettner, F., Sunder, M.: Axiomatic arguments for decomposiing goodness of fit according to Shapley and Owen values. Electronic Journal of Statistics 6, 1239–1250 (2012)

[41] International Agency for Research on Cancer, World Health Organization: Cancer Fact Sheets. https://gco.iarc.fr/today/fact-sheets-cancers (2020), `https://gco.iarc.fr/today/fact-sheets-cancers`

[42] International Medical Device Regulators Forum (IMDRF): Software as a Medical Device (SaMD): Key Definitions. http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-131209-samd-key-definitions-140901.pdf (2013)
[43] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
[44] Israeli, O.: A Shapley-based decomposition of the R-square of a linear regression. Journal of Economic Inequality 5, 199–212 (2007)
[45] Jha, D., Riegler, M.A., Johansen, D., Halvorsen, P., Johansen, H.D.: Doubleu-net: A deep convolutional neural network for medical image segmentation. In: Proceedings of the IEEE International Symposium on Computer-Based Medical Systems (CBMS). pp. 558–564 (2020)
[46] Jha, D., Ali, S., Emanuelsen, K., Hicks, S.A., Thambawita, V., Garcia-Ceja, E., Riegler, M.A., de Lange, T., Schmidt, P.T., Johansen, H.D., Johansen, D., Halvorsen, P.: Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy (2020)
[47] Jha, D., Ali, S., Johansen, H.D., Johansen, D., Rittscher, J., Riegler, M.A., Halvorsen, P.: Real-Time Polyp Detection, Localisation and Segmentation in Colonoscopy Using Deep Learning. arXiv preprint arXiv:2006.11392 (2020)
[48] Jha, D., Hicks, S.A., Emanuelsen, K., Johansen, H.D., Johansen, D., de Lange, T., Riegler, M.A., Halvorsen, P.: Medico multimedia task at mediaeval 2020:automatic polyp segmentation. In: Proceedings of the MediaEval 2020 Workshop (2020)
[49] Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D.: Kvasir-SEG: A Segmented Polyp Dataset. In: Proceedings of the International Conference on Multimedia Modeling (MMM). pp. 451–462 (2020)
[50] Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P., Johansen, H.D.: ResUNet++: An advanced architecture for medical image segmentation. In: Proceedings of International Symposium on Multimedia (ISM). pp. 225–2255 (2019)
[51] Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. arXiv preprint arXiv:1711.08195 (2017)
[52] Kaminski, M.F., Regula, J., Kraszewska, E., Polkowski, M., Wojciechowska, U., Didkowska, J., Zwierko, M., Rupinski, M., Nowacki, M.P., Butruk, E.: Quality indicators for colonoscopy and the risk of interval cancer. New England Journal of Medicine 362(19), 1795–1803 (2010)
[53] Karkanis, S.A., Iakovidis, D.K., Karras, D.A., Maroulis, D.E.: Detection of lesions in endoscopic video using textural descriptors on wavelet domain supported by artificial neural network architectures. In: Proceedings the International Conference on Image Processing. pp. 833–836 (2001)
[54] Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 4401–4410 (2019)

[55] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (2020)
[56] Khan, A., Sohail, A., Zahoora, U., Qureshi, A.S.: A survey of the recent architectures of deep convolutional neural networks. Artificial Intelligence Review 53(8), 5455–5516 (2020)
[57] Ko, W.J., An, P., Ko, K.H., Hahm, K.B., Hong, S.P., Cho, J.Y.: Image quality analysis of various gastrointestinal endoscopes: why image quality is a prerequisite for proper diagnostic and therapeutic endoscopy. Clinical endoscopy 48(5), 374 (2015)
[58] Koulaouzidis, A., Iakovidis, D.K., Yung, D.E., Rondonotti, E., Kopylov, U., Plevris, J.N., Toth, E., Eliakim, A., Johansson, G.W., Marlicz, W., Mavrogenis, G., Nemeth, A., Thorlacius, H., Tontini, G.E.: Kid project: an internet-based digital video atlas of capsule endoscopy for research purposes. Endoscopy international open 5(6), E477–E483 (May 2017)
[59] Krishnan, S.M., Yang, X., Chan, K.L., Kumar, S., Goh, P.M.Y.: Intestinal abnormality detection from endoscopic images. In: Proceedings of the IEEE Annual International Conference of the Engineering in Medicine and Biology Society. pp. 895–898 (1998)
[60] Kundu, A.K., Fattah, S.A., Rizve, M.N.: An Automatic Bleeding Frame and Region Detection Scheme for Wireless Capsule Endoscopy Videos Based on Interplane Intensity Variation Profile in Normalized RGB Color Space. Journal of Healthcare Engineering 2018 (2018)
[61] Le Berre, C., Sandborn, W.J., Aridhi, S., Devignes, M.D., Fournier, L., Smaïl-Tabbone, M., Danese, S., Peyrin-Biroulet, L.: Application of artificial intelligence to gastroenterology and hepatology. Gastroenterology 158(1), 76–94 (2020)
[62] Lee, H., Kim, S.T., Ro, Y.M.: Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis. In: Suzuki, K., Reyes, M., Syeda-Mahmood, T., Glocker, B., Wiest, R., Gur, Y., Greenspan, H., Madabhushi, A. (eds.) Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support. pp. 21–29. Springer International Publishing, Cham (2019)
[63] Lee, J.Y., Jeong, J., Song, E.M., Ha, C., Lee, H.J., Koo, J.E., Yang, D.H., Kim, N., Byeon, J.S.: Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets. Scientific Reports 10(1), 1–9 (2020)
[64] Lee, S., Jang, B., Kim, K.O., Jeon, S., Kwon, J., Kim, E., Jung, J., Park, K., Cho, K., Kim, E.S., Park, C., Yang, C.: Endoscopic experience improves interobserver agreement in the grading of esophagitis by los angeles classification: Conventional endoscopy and optimal band image system. Gut and liver 8, 154–9 (03 2014)
[65] Leenhardt, R., Li, C., Mouel, J.P., Rahmi, G., Sabourin, J.C., Cholet, F., Boureille, A., Amiot, X., Delvaux, M., Duburque, C., Leandri, C., Gerard,

R., Lecleire, S., Mesli, F., Nion-Larmurier, I., Romain, O., Sacher-Huvelin, S., Simon-Shane, C., Vanbiervliet, G., Dray, X.: Cad-cap: a 25,000-image database serving the development of artificial intelligence for capsule endoscopy. Endoscopy international open 8 (03 2020)

[66] Lipovetsky, S., Conklin, M.: Analysis of regression in game theory approach. Applied Stochastic Models in Business and Industry 17, 319–330 (2001)

[67] Liu, Q., Yu, L., Luo, L., Dou, Q., Heng, P.A., Heng, P.A.: Semi-supervised Medical Image Classification with Relation-driven Self-ensembling Model. IEEE Transactions on Medical Imaging pp. 1–1 (2020)

[68] Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[69] Lux, M., Chatzichristofis, S.A.: Lire: Lucene image retrieval: An extensible java cbir library. In: Proceedings of the ACM International Conference on Multimedia (ACM MM). p. 1085–1088 (2008)

[70] Madani, A., Moradi, M., Karargyris, A., Syeda-Mahmood, T.: Semi-supervised learning with generative adversarial networks for chest X-ray classification with ability of data domain adaptation. In: Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1038–1042 (2018)

[71] Mahmud, T., Paul, B., Fattah, S.A.: PolypSegNet: A Modified Encoder-Decoder Architecture for Automated Polyp Segmentation from Colonoscopy Images. Computers in Biology and Medicine p. 104119 (2020)

[72] Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)

[73] Misawa, M., ei Kudo, S., Mori, Y., Hotta, K., Ohtsuka, K., Matsuda, T., Saito, S., Kudo, T., Baba, T., Ishida, F., Itoh, H., Oda, M., Mori, K.: Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). Gastrointestinal Endoscopy (2020)

[74] Norman, B., Pedoia, V., Majumdar, S.: Use of 2d u-net convolutional neural networks for automated cartilage and meniscus segmentation of knee mr imaging data to determine relaxometry and morphometry. Radiology 288(1), 177–185 (2018)

[75] Park, S.Y., Sargent, D., Spofford, I., Vosburgh, K.G., A-Rahim, Y.: A colon video analysis framework for polyp detection. IEEE Transactions on Biomedical Engineering 59(5), 1408–1418 (2012)

[76] Pogorelov, K., Riegler, M., Halvorsen, P., Griwodz, C., Lange, T., Randel, K., Eskeland, S., Dang-Nguyen, D.T., Ostroukhova, O., Lux, M., Spampinato, C.: A comparison of deep learning with global features for gastroin-

testinal disease detection. In: CEUR Workshop Proceedings - MediaEval. vol. 1984, pp. 8–10 (2017)
[77] Pogorelov, K., Randel, K., Griwodz, C., de Lange, T., Eskeland, S., Johansen, D., Spampinato, C., Dang Nguyen, D.T., Lux, M., Schmidt, P., Riegler, M., Halvorsen, P.: Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of ACM Multimedia Systems (MMSYS) (2017)
[78] Pogorelov, K., Randel, K., de Lange, T., Eskeland, S., Johansen, D., Griwodz, C., Spampinato, C., Taschwer, M., Lux, M., Schmidt, P., Riegler, M., Halvorsen, P.: Nerthus: A bowel preparation quality video dataset. In: Proceedings of ACM Multimedia Systems (MMSYS) (2017)
[79] Pogorelov, K., Riegler, M., Halvorsen, P., Hicks, S., Randel, K.R., Dang Nguyen, D.T., Lux, M., Ostroukhova, O., de Lange, T.: Medico multimedia task at mediaeval 2018. In: CEUR Workshop Proceedings - MediaEval (2018)
[80] Polit, D.F.: Blinding during the analysis of research data. International Journal of Nursing Studies 48(5), 636 – 641 (2011), http://www.sciencedirect.com/science/article/pii/S0020748911000496
[81] Qadir, H.A., Balasingham, I., Solhusvik, J., Bergsland, J., Aabakken, L., Shin, Y.: Improving automatic polyp detection using cnn by exploiting temporal dependency in colonoscopy video. IEEE Journal of Biomedical and Health Informatics 24(1), 180–193 (2020)
[82] Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging. In: Proceedings of Advances in neural information processing systems (NeurIPS). pp. 3347–3357 (2019)
[83] Repici, A., Badalamenti, M., Maselli, R., Correale, L., Radaelli, F., Rondonotti, E., Ferrara, E., Spadaccini, M., Alkandari, A., Fugazza, A., Anderloni, A., Galtieri, P.A., Pellegatta, G., Carrara, S., Di Leo, M., Craviotto, V., Lamonaca, L., Lorenzetti, R., Andrealli, A., Antonelli, G., Wallace, M., Sharma, P., Rosch, T., Hassan, C.: Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology 159(2), 512 – 520 (2020), http://www.sciencedirect.com/science/article/pii/S0016508520305837
[84] Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 1135–1144. KDD '16, Association for Computing Machinery, New York, NY, USA (2016), https://doi.org/10.1145/2939672.2939778
[85] Riegler, M., Pogorelov, K., Halvorsen, P., Lange, T.d., Griwodz, C., Schmidt, P.T., Eskeland, S.L., Johansen, D.: EIR — Efficient computer aided diagnosis framework for gastrointestinal endoscopies. In: Proceeding of the International Workshop on Content-Based Multimedia Indexing (CBMI). pp. 1–6 (Jun 2016)
[86] Riegler, M., Lux, M., Griwodz, C., Spampinato, C., de Lange, T., Eskeland, S.L., Pogorelov, K., Tavanapong, W., Schmidt, P.T., Gurrin, C., Johansen, D., Johansen, H., Halvorsen, P.: Multimedia and medicine: Teammates

for better disease detection and survival. In: Proceedings of the ACM International Conference on Multimedia (ACM MM). pp. 968–977 (2016), http://doi.acm.org/10.1145/2964284.2976760

[87] Riegler, M., Pogorelov, K., Eskeland, S.L., Schmidt, P.T., Albisser, Z., Johansen, D., Griwodz, C., Halvorsen, P., Lange, T.D.: From annotation to computer-aided diagnosis: Detailed evaluation of a medical multimedia system. ACM Trans. Multimedia Comput. Commun. Appl. (May 2017), https://doi.org/10.1145/3079765

[88] Rondonotti, E., Soncini, M., Girelli, C.M., Russo, A., Ballardini, G., Bianchi, G., Cantù, P., Centenara, L., Cesari, P., Cortelezzi, C.C., Gozzini, C., Lupinacci, G., Maino, M., Mandelli, G., Mantovani, N., Moneghini, D., Morandi, E., Putignano, R., Schalling, R., Tatarella, M., Vitagliano, P., Villa, F., Zatelli, S., Conte, D., Masci, E., de Franchis, R.: Can we improve the detection rate and interobserver agreement in capsule endoscopy? Digestive and Liver Disease 44(12), 1006 – 1011 (2012), http://www.sciencedirect.com/science/article/pii/S1590865812002368

[89] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proceeding of the International Conference on Medical image computing and computer-assisted intervention (MICCAI). pp. 234–241. Springer (2015)

[90] Ross, T., Reinke, A., Full, P.M., Wagner, M., Kenngott, H., Apitz, M., Hempe, H., Filimon, D.M., Scholz, P., Tran, T.N., Bruno, P., Arbeláez, P., Bian, G.B., Bodenstedt, S., Bolmgren, J.L., Bravo-Sánchez, L., Chen, H.B., González, C., Guo, D., Halvorsen, P., Heng, P.A., Hosgor, E., Hou, Z.G., Isensee, F., Jha, D., Jiang, T., Jin, Y., Kirtac, K., Kletz, S., Leger, S., Li, Z., Maier-Hein, K.H., Ni, Z.L., Riegler, M.A., Schoeffmann, K., Shi, R., Speidel, S., Stenzel, M., Twick, I., Wang, G., Wang, J., Wang, L., Wang, L., Zhang, Y., Zhou, Y.J., Zhu, L., Wiesenfarth, M., Kopp-Schneider, A., Müller-Stich, B.P., Maier-Hein, L.: Robust medical instrument segmentation challenge 2019 (2020)

[91] Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PloS one 10(3), e0118432 (2015)

[92] Shapley, L.S.: A value for n-person games. Contributions to the Theory of Games 2(28), 307–317 (1953)

[93] Shin, Y., Qadir, H.A., Balasingham, I.: Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance. IEEE Access 6, 56007–56017 (2018)

[94] Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. CoRR abs/1704.02685 (2017), http://arxiv.org/abs/1704.02685

[95] Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. Int. Jour. of Comput. Assis. Radiol. and Surg. 9(2), 283–293 (2014)

[96] Smedsrud, P.H., Gjestang, H.L., Nedrejord, O.O., Næss, E., Thambawita, V., Hicks, S.A., Borgli, H., Jha, D., Berstad, T.J.D., Eskeland, S.L., Lux,

M., Espeland, H., Petlund, A., Dang-Nguyen, D.T., Garcia-Ceja, E., Johansen, D., Schmidt, P.T., Hammer, H.L., de Lange, T., Riegler, M., Halvorsen, P.: Kvasir-capsule, a video capsule endoscopy dataset. OSF Preprints (Aug 2020), https://doi.org/10.31219/osf.io/gr7bn

[97] Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans. Med. Imag. 35(2), 630–644 (2015)

[98] Thambawita, V., Jha, D., Hammer, H.L., Johansen, H.D., Johansen, D., Halvorsen, P., Riegler, M.A.: An extensive study on cross-dataset bias and evaluation metrics interpretation for machine learning applied to gastrointestinal tract abnormality classification. ACM Transactions on Computing for Healthcare 1(3) (Jun 2020), https://doi.org/10.1145/3386295

[99] Thambawita, V., Jha, D., Riegler, M., Halvorsen, P., Hammer, H.L., Johansen, H., Johansen, D.: The Medico-Task 2018: Disease Detection in the Gastrointestinal Tract using Global Features and Deep Learning. In: CEUR Workshop Proceedings - MediaEval (2018)

[100] Tomar, N.K., Jha, D., Ali, S., Johansen, H.D., Johansen, D., Riegler, M.A., Halvorsen, P.: DDANet: Dual Decoder Attention Network forAutomatic Polyp Segmentation. arXiv preprint arXiv:2006.11392 (2020)

[101] Tommasi, T., Tuytelaars, T.: A testbed for cross-dataset analysis. In: European Conference on Computer Vision. pp. 18–31. Springer (2014)

[102] Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine 25(1), 44–56 (2019)

[103] Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: Proceedings of the International Conference on Pattern Recongition (CVPR). pp. 1521–1528. IEEE (2011)

[104] Torre, L.A., Bray, F., Siegel, R.L., Ferlay, J., Lortet-Tieulent, J., Jemal, A.: Global cancer statistics, 2012. CA: a cancer journal for clinicians 65(2), 87–108 (2015)

[105] U.S. Food and Drug Administration: Learn if a medical device has been cleared by FDA for marketing. https://www.fda.gov/medical-devices/consumers-medical-devices/learn-if-medical-device-has-been-cleared-fda-marketing (2017)

[106] U.S. Food and Drug Administration: Digital health innovation action plan. https://www.fda.gov/media/106331/download (2018)

[107] U.S. Food and Drug Administration: FDA permits marketing of artificial intelligence algorithm for aiding providers in detecting wrist fractures. https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-algorithm-aiding-providers-detecting-wri (2018)

[108] Van Doorn, S.C., Hazewinkel, Y., East, J.E., Van Leerdam, M.E., Rastogi, A., Pellisé, M., Sanduleanu-Dascalescu, S., Bastiaansen, B.A., Fockens, P., Dekker, E.: Polyp morphology: an interobserver evaluation for the paris classification among international experts. The American journal of gastroenterology 110(1), 180 (2015)

[109] Wang, P., Xiao, X., Brown, J., Berzin, T., Tu, M., Xiong, F., Hu, X., Liu, P., Song, Y., Zhang, D., Yang, X., Li, L., He, J., Yi, X., Liu, J., Liu, X., Lai, L.: Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. Nature Biotechnology pp. 741–748 (10 2018)

[110] Wang, Y., Tavanapong, W., Wong, J., Oh, J., de Groen, P.C.: Part-based multiderivative edge cross-sectional profiles for polyp detection in colonoscopy. IEEE Journal of Biomedical and Health Informatics 18(4), 1379–1389 (2014)

[111] Wang, Y., Tavanapong, W., Wong, J., Oh, J.H., De Groen, P.C.: Polyp-alert: Near real-time feedback during colonoscopy. Comp. Meth. Progr. Biomed. 120(3), 164–179 (2015)

[112] Wei, J., Suriawinata, A., Vaickus, L., Ren, B., Liu, X., Lisovsky, M., Tomita, N., Abdollahi, B., Kim, A., Snover, D., Baron, J., Barry, E., Hassanpour, S.: Evaluation of a deep neural network for automated classification of colorectal polyps on histopathologic slides. JAMA Network Open 3, e203398 (04 2020)

[113] Wickstrøm, K., Kampffmeyer, M., Jenssen, R.: Uncertainty modeling and interpretability in convolutional neural networks for polyp segmentation. In: Rroceedings of the IEEE international workshop on machine learning for signal processing (mlsp). pp. 1–6. IEEE (2018)

[114] Wickstrøm, K., Kampffmeyer, M., Jenssen, R.: Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. Medical Image Analysis 60, 101619 (2020), http://www.sciencedirect.com/science/article/pii/S1361841519301574

[115] Woolhandler, S., Himmelstein, D.U.: Administrative work consumes one-sixth of u.s. physicians' working hours and lowers their career satisfaction. International Journal of Health Services 44(4), 635–42 (2014)

[116] Wu, H., Prasad, S.: Semi-Supervised Deep Learning Using Pseudo Labels for Hyperspectral Image Classification. IEEE Transactions on Image Processing 27(3), 1259–1270 (2018)

[117] Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with Noisy Student improves ImageNet classification. arXiv (2020), http://arxiv.org/abs/1911.04252

[118] Xue, Y., Xu, T., Long, L.R., Xue, Z., Antani, S., Thoma, G.R., Huang, X.: Multimodal recurrent model with attention for automated radiology report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 457–466. Springer (2018)

[119] Yang, J., Faraji, M., Basu, A.: Robust segmentation of arterial walls in intravascular ultrasound images using dual path u-net. Ultrasonics 96, 24–33 (2019)

[120] Zhang, C., Tavanapong, W., Wong, J., de Groen, P.C., Oh, J.: Real Data Augmentation for Medical Image Classification. In: Cardoso, M.J., Arbel, T., Lee, S.L., Cheplygina, V., Balocco, S., Mateus, D., Zahnd, G., Maier-Hein, L., Demirci, S., Granger, E., Duong, L., Carbonneau, M.A., Albarqouni, S., Carneiro, G. (eds.) Intravascular Imaging and Computer

Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, vol. 10552, pp. 67–76. Springer International Publishing (2017), http://link.springer.com/10.1007/978-3-319-67534-3_8

[121] Zhang, Z., Xie, Y., Xing, F., McGough, M., Yang, L.: Mdnet: A semantically and visually interpretable medical image diagnosis network. In: Proc. of IEEE CVPR. pp. 6428–6436 (2017)

[122] Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 3–11. Springer (2018)

[123] Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)

A.20 Paper XX : Comparative validation of multi-instance instrument segmentation in endoscopy: Results of the ROBUST-MIS 2019 challenge

Authors: Tobias Roß, Annika Reinke, Peter M. Full, Martin Wagner, Hannes Kenngott, Martin Apitz, Hellena Hempe, Diana Mindroc-Filimon, Patrick Scholz, Thuy Nuong Tran, Pierangela Bruno, Pablo Arbeláez, Gui-BinBian, Sebastian Bodenstedt, Jon Lindström Bolmgren, Laura Bravo-Sánchez, Hua-Bin Chen, Cristina González, Dong Guo, Pål Halvorsen, Pheng-Ann Heng, Enes Hosgor, Zeng-Guang Hou, Fabian Isensee, Debesh Jha, Tingting Jiang, Yueming Jin, Kadir Kirtac, Sabrina Kletz, Stefan Leger, Zhixuan Li, Klaus H. Maier-Hein, Zhen-Liang Ni, Michael A. Riegler, Klaus Schoeffmann, Ruohua Shi, Stefanie Speidel, Michael Stenzel, Isabell Twick, Gutai Wang, Jiacheng Wang, Liansheng Wang, Lu Wang, Yujie Zhang, Yan-Jie Zhou, Lei Zhu, Manuel Wiesenfarth, Annette Kopp-Schneider, Beat P.Müller-Stich, Lena Maier-Hein

Abstract: Intraoperative tracking of laparoscopic instruments is often a prerequisite for computer and robotic-assisted interventions. While numerous methods for detecting, segmenting and tracking of medical instruments based on endoscopic video images have been proposed in the literature, key limitations remain to be addressed: Firstly, robustness, that is, the reliable performance of state-of-the-art methods when run on challenging images (e.g. in the presence of blood, smoke or motion artifacts). Secondly, generalization; algorithms trained for a specific intervention in a specific hospital should generalize to other interventions or institutions.

In an effort to promote solutions for these limitations, we organized the Robust Medical Instrument Segmentation (ROBUST-MIS) challenge as an international benchmarking competition with a specific focus on the robustness and generalization capabilities of algorithms. For the first time in the field of endoscopic image processing, our challenge included a task on binary segmentation and also addressed multi-instance detection and segmentation. The challenge was based on a surgical data set comprising 10,040 annotated images acquired from a total of 30 surgical procedures from three different types of surgery. The validation of the competing methods for the three tasks (binary segmentation, multi-instance detection and multi-instance segmentation) was performed in three different stages with an in-

creasing domain gap between the training and the test data. The results confirm the initial hypothesis, namely that algorithm performance degrades with an increasing domain gap. While the average detection and segmentation quality of the best-performing algorithms is high, future research should concentrate on detection and segmentation of small, crossing, moving and transparent instrument(s) (parts).

Published: Medical Image Analysis, Vol. 70, 2021.

Candidate contributions: D. Jha participated in the binary segmentation task from the Robust Medical Instrument challenge. He proposed a solution for binary instrument segmentation. His team was 9th for the binary instrument segmentation task. He shared his method's information with the challenge organizers in a 2-page paper, and he also participated in a survey where the challenge organizers asked participants to fill all the critical information in as a summary. The challenge organizers conceptualized, designed the study, wrote and revised the manuscript to prepare the full version with the participating team's information collected.

Thesis objectives: Objective III

Medical Image Analysis 70 (2021) 101920

Contents lists available at ScienceDirect

Medical Image Analysis

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

Challenge Report

Comparative validation of multi-instance instrument segmentation in endoscopy: Results of the ROBUST-MIS 2019 challenge

Tobias Roß[a,b,1,1,*,*], Annika Reinke[a,b,1], Peter M. Full[b,c], Martin Wagner[d], Hannes Kenngott[d], Martin Apitz[d], Hellena Hempe[a], Diana Mindroc-Filimon[a], Patrick Scholz[a,e], Thuy Nuong Tran[a], Pierangela Bruno[a,v], Pablo Arbeláez[f], Gui-Bin Bian[g,h], Sebastian Bodenstedt[i,j,k], Jon Lindström Bolmgren[l], Laura Bravo-Sánchez[f], Hua-Bin Chen[g,h], Cristina González[f], Dong Guo[m], Pål Halvorsen[n,o], Pheng-Ann Heng[p], Enes Hosgor[l], Zeng-Guang Hou[g,h], Fabian Isensee[b,c], Debesh Jha[n,q], Tingting Jiang[r], Yueming Jin[p], Kadir Kirtac[l], Sabrina Kletz[s], Stefan Leger[i,j,k], Zhixuan Li[r], Klaus H. Maier-Hein[c], Zhen-Liang Ni[g,h], Michael A. Riegler[n], Klaus Schoeffmann[s], Ruohua Shi[r], Stefanie Speidel[i,j,k], Michael Stenzel[l], Isabell Twick[l], Gutai Wang[m], Jiacheng Wang[t], Liansheng Wang[t], Lu Wang[m], Yujie Zhang[t], Yan-Jie Zhou[g,h], Lei Zhu[p], Manuel Wiesenfarth[u], Annette Kopp-Schneider[u], Beat P. Müller-Stich[d], Lena Maier-Hein[a]

[a] *Computer Assisted Medical Interventions (CAMI), German Cancer Research Center, Im Neuenheimer Feld 223, 69120, Heidelberg, Germany*
[b] *University of Heidelberg, Germany, Seminarstraße 2, 69117 Heidelberg, Germany*
[c] *Division of Medical Image Computing (MIC), Im Neuenheimer Feld 223, 69120 Heidelberg, Germany*
[d] *Department for General, Visceral and Transplantation Surgery, Heidelberg University Hospital, Im Neuenheimer Feld 110, 69120 Heidelberg, Germany*
[e] *HIDSS4Health – Helmholtz Information and Data Science School for Health, Im Neuenheimer Feld 223, 69120 Heidelberg, Germany*
[f] *Universidad de los Andes, Cra. 1 No 18A - 12, 111711 Bogotá, Colombia*
[g] *University of Chinese Academy Sciences, 52 Sanlihe Rd., Beijing, China*
[h] *State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, 100864 Beijing, China*
[i] *National Center for Tumor Diseases (NCT), Partner Site Dresden, Germany: German Cancer Research Center, Im Neuenheimer Feld 460, 69120 Heidelberg, Germany*
[j] *Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Fetscherstraße 74, 01307 Dresden, Germany*
[k] *Helmholtz Association/Helmholtz-Zentrum Dresden - Rossendorf (HZDR), Bautzner Landstraße 400, 01328 Dresden, Germany*
[l] caresyntax, *Komturstraße 18A, 12099 Berlin, Germany*
[m] *School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Shahe Campus:No.4, Section 2, North Jianshe Road, 610054 | Qingshuihe Campus:No.2006, Xiyuan Ave, West Hi-Tech Zone, 611731, Chengdu, China*
[n] *SimulaMet, Pilestredet 52, 0167 Oslo, Norway*
[o] *Oslo Metropolitan University (OsloMet), Pilestredet 52, 0167 Oslo, Norway*
[p] *Department of Computer Science and Engineering, The Chinese University of Hong Kong, Chung Chi Rd, Ma Liu Shui, Hong Kong, China*
[q] *Department of Informatics, UIT The Arctic University of Norway, Hansine Hansens vei 54, 9037 Tromsø, Norway*
[r] *Institute of Digital Media (NELVT), Peking University, 5 Yiheyuan Rd, Haidian District, 100871 Peking, China*
[s] *Institute of Information Technology, Klagenfurt University, Universitätsstraße 65-67, 9020 Klagenfurt, Austria*
[t] *Department of Computer Science, School of Informatics, Xiamen University, 422 Siming South Road, 361005 Xiamen, China*
[u] *Division of Biostatistics, German Cancer Research Center, Im Neuenheimer Feld 581, Heidelberg, Germany*
[v] *Department of Mathematics and Computer Science, University of Calabria, 87036 Rende, Italy*

ARTICLE INFO

Article history:
Received 20 May 2020
Revised 22 September 2020
Accepted 24 November 2020
Available online 28 November 2020

ABSTRACT

Intraoperative tracking of laparoscopic instruments is often a prerequisite for computer and robotic-assisted interventions. While numerous methods for detecting, segmenting and tracking of medical instruments based on endoscopic video images have been proposed in the literature, key limitations remain to be addressed: Firstly, *robustness*, that is, the reliable performance of state-of-the-art methods

* Corresponding author.
E-mail address: t.ross@dkfz-heidelberg.de (T. Roß).
[1] Contributed equally to this paper.

https://doi.org/10.1016/j.media.2020.101920

Keywords:
Multi-instance instrument
Minimally invasive surgery
Robustness and generalization
Surgical data science

when run on challenging images (e.g. in the presence of blood, smoke or motion artifacts). Secondly, *generalization*; algorithms trained for a specific intervention in a specific hospital should generalize to other interventions or institutions.

In an effort to promote solutions for these limitations, we organized the *Robust Medical Instrument Segmentation (ROBUST-MIS) challenge* as an international benchmarking competition with a specific focus on the robustness and generalization capabilities of algorithms. For the first time in the field of endoscopic image processing, our challenge included a task on binary segmentation and also addressed multi-instance detection and segmentation. The challenge was based on a surgical data set comprising 10,040 annotated images acquired from a total of 30 surgical procedures from three different types of surgery. The validation of the competing methods for the three tasks (binary segmentation, multi-instance detection and multi-instance segmentation) was performed in three different stages with an increasing domain gap between the training and the test data. The results confirm the initial hypothesis, namely that algorithm performance degrades with an increasing domain gap. While the average detection and segmentation quality of the best-performing algorithms is high, future research should concentrate on detection and segmentation of small, crossing, moving and transparent instrument(s) (parts).

1. Introduction

Minimally invasive surgery has become increasingly common over the past years (Siddaiah-Subramanya et al., 2017). However, issues such as limited view, a lack of depth information, haptic feedback and increased difficulty in handling instruments have increased the complexity for the surgeons. Surgical data science applications (Maier-Hein et al., 2017) could help the surgeon to overcome those limitations and to increase patient safety. These applications, e.g. surgical skill assessment (Law et al., 2017; Lin et al., 2019), augmented reality (Wang et al., 2017; Burström et al., 2019), assistance robots (Amini Khoiy et al., 2016; Zhang and Gao, 2020), vision-based force estimation (Su et al., 2018) or depth enhancement (De Paolis and De Luca, 2019), are often based on the segmentation and/or tracking of medical instruments during surgery. Currently, commercial tracking systems usually rely on optical or electromagnetic markers and, therefore, also require additional hardware (Bianchi et al., 2019; Zhou et al., 2019), which are expensive, need extra space and require technical knowledge. Alternatively, with the recent success of deep learning methods in the medical domain (Esteva et al., 2019) and first surgical data science applications (Fawaz et al., 2019; Nguyen et al., 2019), video-only based approaches offer new opportunities to handle difficult image scenarios such as bleeding, light over-/underexposure, smoke and reflections (Bodenstedt et al., 2018). Video-only based approaches offer new opportunities to handle difficult image scenarios such as bleeding, light over-/underexposure, smoke and reflections (García-Peraza-Herrera et al., 2016; Kurmann et al., 2017; Laina et al., 2017; Pakhomov et al., 2019; Zhao et al., 2019). In turn, the tracking information may directly affect the instructions provided to the surgeon to navigate the surgical instruments. Furthermore, unreliable algorithms potentially reduce the acceptance on the part of the surgical team, and thus, the chances for translation into the clinical routine (Panch et al., 2019; Qayyum et al., 2020).

As validation and evaluation of image processing methods is usually performed on the researchers' individual data sets, finding the best algorithm suited for a specific use case is a difficult task. Consequently, reported publication results are often difficult to compare (Ioannidis, 2005; Armstrong et al., 2009). In order to overcome this issue, we can implement *challenges* to find algorithms that work best on specific problems. These international benchmarking competitions aim to assess the performance of several algorithms on the same data set, which enables a fair comparison to be drawn across multiple methods (Maier-Hein et al., 2018; 2019).

One international challenge which takes place on a regular basis is the Endoscopic Vision (EndoVis) Challenge[2]. It hosts sub-challenges with a broad variety of tasks in the field of endoscopic image processing and and has been held annually at the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI) since 2015 (exception: 2016). However, data sets provided for instrument detection/tracking/segmentation in previous EndoVis editions (e.g., (Allan et al., 2019, 2020)) comprised a relatively small number of cases (between ~500 to ~4,000) and generally represented best cases scenarios (e.g. with clean views, limited distortions in videos) which did not comprehensively reflect the challenges in real-world clinical applications. Although these competitions enabled primary insights and comparison of the methods, the information gained on robustness and generalization capabilities of methods were limited.

To remedy these issues, we present the Robust Medical Instrument Segmentation (ROBUST-MIS) challenge 2019, which was part of the 4th edition of EndoVis at MICCAI 2019. We introduced a large data set comprising more than 10,000 image frames for instrument segmentation and detection, extracted from daily routine surgeries. The data set contained images which included all types of difficulties and was annotated by medical experts according to a pre-defined labeling protocol and subjected to a quality control process. The challenge addressed methods with a projected application in minimally invasive surgeries, in particular the tracking of medical instruments in the abdomen, with a special focus on the generalizibility and robustness. This was achieved by introducing three stages with increase in difficulty in the test phase. To emphasize the robustness of methods, we used a ranking scheme that specifically measures the worst-case performance of algorithms.

Section 2 outlines the challenge design as a whole, including the data set. The results of the challenge are presented in Section 3 with a discussion following in Section 4. The appendix includes challenge design choices regarding the organization (see Appendix A), the labeling and submission instructions (see Appendix B and Appendix C), the rankings across all stages (see Appendix D) and the complete challenge design document (see Appendix F).

2. Methods

The ROBUST-MIS 2019 challenge was organized as a sub-challenge of the Endoscopic Vision Challenge 2019 at MICCAI 2019

[2] https://endovis.grand-challenge.org/.

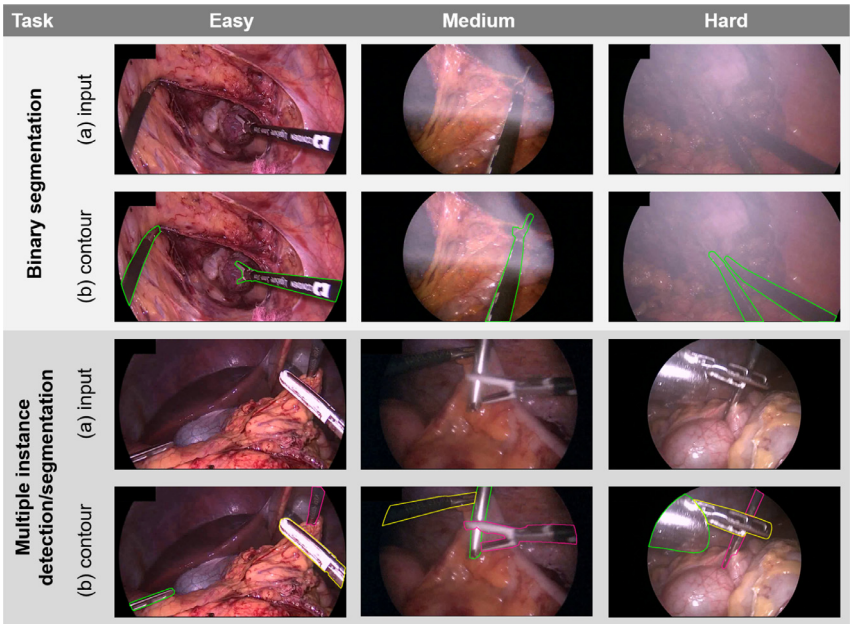


Fig. 1. Various levels of difficulty represented in the challenge data for the binary segmentation (two upper rows) and multi-instance detection/segmentation tasks (two lower rows). Input frames (a) are shown along with the reference segmentation masks for all tasks. The latter are shown as contours (b).

in Shenzhen, China. Details of the challenge organization can be found in Appendix A and Appendix F. The objective of the challenge, the challenge data sets and the assessment method used to evaluate the participating algorithms are presented in the following.

2.1. Mission of the challenge

The goal of the ROBUST-MIS 2019 challenge was to benchmark algorithms designed for instrument detection and segmentation in videos of minimally invasive surgeries. Specifically, we were interested in (1) identifying robust methods for instrument detection and segmentation, (2) assessing the generalization capabilities of the methods proposed and (3) identifying the image properties (e.g. smoke, bleeding, motion artifacts) that make images particularly challenging. The challenges' metrics and ranking schemes were designed to assess these properties (see Section 2.3).

The challenge was divided into three different tasks with separate evaluations and leaderboards (see Fig. 1). For the binary segmentation task, participants had to provide precise contours of instruments, using binary masks, with '1' indicating the presence of a surgical instrument in a given pixel and '0' representing the absence thereof. Analogously, for the multi-instance segmentation task, participants had to provide image masks by allotting numbers '1', '2', etc. which represented different instances of medical instruments. In contrast, the multi-instance detection task merely required participants to detect and roughly locate instrument instances in video frames in which the location could be represented by arbitrary forms, such as bounding boxes.

As detailed in Section 2.3, the generalizability and performance of all participating algorithms was assessed in three stages with increasing levels of difficulty:

- **Stage 1:** Test data was taken from the procedures (patients) from which the training data were extracted.
- **Stage 2:** Test data was taken from the exact same type of surgery as the training data but from procedures (patients) not included in the training
- **Stage 3:** Test data was taken from a different but similar type of surgery (and different patients) compared to the training data.

Before the algorithms were submitted to the challenge, participants were only informed of the surgery types for stages 1 and 2 (rectal resection and proctocolectomy, see Section 2.2.1). For the third stage, the surgery type (sigmoid resection) was referred to as *unknown surgery* to enable the generalizability to be tested.

2.2. Challenge data set

2.2.1. Data recording

All data was recorded with a Karl Storz Image 1 laparoscopic camera (Karl Storz SE & Co. KG, Tuttlingen, Germany), with a 30° optic lens. The Karl Storz Xenon 300 was used as a light source. Data acquisition was executed during daily routine procedures at the Heidelberg University Hospital, Department of Surgery in the integrated operating room (Karl Storz OR1 FUSION®). Whenever parts of the video showed the outside of the abdomen, these frames were manually excluded for the purpose of anonymization. To reduce storage and memory usage, image resolution was reduced from 1920 × 1080 pixels (HD) in the primary video to 960 × 540. Videos from 30 minimally invasive surgical procedures taken in three different types of surgery, namely 10 *rectal resection* procedures, 10 *proctocolectomy* procedures and 10 procedures of *sigmoid resection* procedures, served as a basis for this challenge. A total of 10,040 images were extracted from these 30 procedures according to the procedure summarized in Section 2.2.2.

2.2.2. Data extraction

The frames were selected according to the following procedures: Initially, whenever the camera was outside the abdomen, the corresponding frames were removed to ensure anonymization. Next, all videos were sampled at a rate of 1 frame/sec, eliciting 4,456 extracted frames. To increase this number, additional frames were extracted during the surgical phase transitions, resulting in a total of 10,040 frames. Labels for the surgical phases were available from the previous challenge *EndoVis Surgical Workflow Analysis in the SensorOR*[3]. All of these frames were annotated as described in 2.2.3.

[3] https://endovissub2017-workflow.grand-challenge.org/.

Table 1
Case distribution of the data with frames per stage and surgery. Empty frames (ef) were classed as the % of frames in which an instrument did not appear.

PROCEDURE	TRAINING	TESTING		
		Stage 1	Stage 2	Stage 3
proctocolectomy	2,943 (2% ef.)	325 (11% ef.)	225 (11% ef.)	0
rectal resection	3,040 (20% ef.)	338 (20% ef.)	289 (15% ef.)	0
sigmoid resection*	0	0	0	2,880 (23% ef.)
TOTAL	5,983 (17% ef.)	663 (15% ef.)	514 (13% ef.)	2,880 (23% ef.)

* *unknown surgery*

2.2.3. Label generation

As stated in the introduction, a labeling mask was created for each of the 10,040 extracted endoscopic video frames. The assignment of instances was done per frame, not per video. The instrument labels were generated according to the following procedure: First, the company Understand AI[4] performed initial segmentations on the extracted frames. Following this, the challenge organizers analyzed the annotations, identified inconsistencies and agreed on an annotation protocol (see Appendix B). A team of 14 engineers and four medical students reviewed all of the annotations and, if necessary, refined them according to the annotation protocol. In ambiguous or unclear cases, a team of two engineers and one medical student generated a consensus annotation. For quality control, a medical expert went through all of the refined segmentation masks and reported potential errors. The final decision on the labels was made by a team comprised of a medical expert and an engineer.

2.2.4. Training and test case definition

A training case comprised a 10 second video snippet in the form of 250 endoscopic image frames and a reference annotation for the last frame. For training cases, the entire video was provided as context information along with information on the surgery type. Test cases were identical in format but did not include a reference annotation.

For the division of the data into training and test data, in accordance with the described testing scheme, all sigmoid resection procedures were reserved for stage 3. The two shortest videos per procedure (20%) were selected from the remaining 20 videos for stage 2 in order to have as much training data as possible. Finally, every 10th annotated frame from the remaining 16 videos was used for stage 1 testing. All other frames were released as training data.

No validation cases for hyperparameter tuning were provided by the organizers; hence, it was up to the challenge participants to split the training cases into training and validation data. In summary, this led to a case distribution as shown in Table 1.

2.3. Assessment method

2.3.1. Metrics

The following metrics[5] were used to assess performance:

- Binary Segmentation: Dice Similarity Coefficient (*DSC*) (Dice, 1945) and Normalized Surface Dice (*NSD*)[6] (Nikolov et al., 2018),
- Multi-instance Detection: F1-score (other name for the *DSC*)(Dice, 1945),
- Multi-instance Segmentation: Multi-instance Dice Similarity Coefficient (*MI_DSC*) and multi-instance Normalized Surface Dice (*MI_NSD*).

The *DSC* is a widely used overlap metric for segmentation (Cardoso, 2018; Everingham et al., 2015) and detection challenges (e.g., the Cerebral Aneurysm Detection (CADA)[20]). It is defined as the harmonic mean of precision and recall:

$$DSC(Y,\hat{Y}) := \frac{2|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|}, \tag{1}$$

where Y denotes the reference annotation and $\hat{Y}$ the corresponding prediction of an image frame.

The *NSD* served as a distance-based measurement for assessing performance. In contrast to the *DSC*, which measures the overlap of volumes, the *NSD* measures the overlap of two surfaces (mask borders) (Nikolov et al., 2018). Furthermore, the metric uses a threshold that is related to the inter-rater variability of the annotators. In our case, the inter-rater variability was computed by a pairwise comparison of a total of 5 annotators over $n = 20$ training images, which resulted in a threshold of $\tau := 13$. Further analysis revealed that thresholds above 10 had no effect on rankings.

According to the challenge design, the indices of instrument instances between the references and predictions did not necessarily match. The only requirement was that each instance was assigned a unique instrument index. Thus, all multi-instance tasks required the prediction and references to be matched, which was computed by applying the Hungarian algorithm (Kuhn, 1955).

To compute the *MI_DSC* and *MI_NSD*, matches of instrument instances were computed. Afterwards, the resulting performance scores for each instrument instance per image have been aggregated by the mean. The choice of the metrics *(MI_)DSC* and *(MI_)NSD* were based on the Medical Segmentation Decathlon challenge (Cardoso, 2018) for the binary segmentation and the multi instance tasks.

Finally, the *F1-score* for the detection task requires the definition of true positives (TP), false negatives (FN) and false positives (FP), where $F1(Y,\hat{Y}) := \frac{2TP}{2TP+FN+FP}$. The assignment of matching candidates was done using the Hungarian algorithm. For this purpose, the intersection over union (*IoU*) was computed for each possible pair of reference and prediction instances, which simply measures the overlap of two areas, divided by their union:

$$IoU(Y,\hat{Y}) := \frac{|Y \cap \hat{Y}|}{|Y \cup \hat{Y}|}, \tag{2}$$

where in both cases Y denotes the reference annotation and $\hat{Y}$ the corresponding prediction of an image frame. Similar to the *MI_DSC* computation, the Hungarian algorithm (Kuhn, 1955) was used to assign matching pairs of references and predictions. Assigned pairs of references and predictions $(Y,\hat{Y})$ were defined as TP if their $IoU(Y,\hat{Y}) > \xi := 0.3$. Reference instances without or with a smaller

[4] https://understand.ai.

[5] The implementation of all metrics can be found here: https://phabricator.mitk.org/source/rmis2019/.

[6] https://github.com/deepmind/surface-distance.

[20] https://cada.grand-challenge.org

prediction than ξ were defined as FN. All instances that could not be assigned to a reference instance were assigned to FP.

2.3.2. Rankings

Separate rankings for accuracy and robustness were computed for stage 3 of the challenge in order to address multiple aspects of the challenge purpose. To investigate accuracy, a significance ranking[7] as recently applied in the MSD (Cardoso, 2018) and described in Algorithm 1 was computed. The robustness ranking specifically

Algorithm 1 Ranking scheme for the binary and multi-instance segmentation tasks.

Let $T = \{t_1, \ldots, t_N\}$ be the test cases for the given task.
for all participating algorithms a_i **do**
  Determine the performance $m(a_i, t_j)$ of algorithm a_i for each test case t_j
  if $m(a_i, t_j) == N/A$ **then**
    $m(a_i, t_j) = 0$
  end if
  Aggregate metric values $m(a_i, t_j)$ with the following two aggregation methods:

1. **Accuracy:** Compute the *significance ranking*. For each pair of algorithms, perform one-sided Wilcoxon signed rank tests with a significance level of $\alpha = 0.05$ to assess differences in the metric values. The accuracy rank $r_a(a_i)$ for algorithm~a_i is based on the number of significant test results for each algorithm (Maier-Hein et~al., 2018; Cardoso, 2018).
2. **Robustness:** Compute the *5% percentile* of all $m(a_i, t_j)$ to get the robustness rank $r_r(a_i)$ for algorithm~a_i.

end for

focused on the worst case performance of methods. For this reason, the 5% percentile was computed instead of aggregating metric values with the mean or median. The computation of the *F1-score* naturally included a ranking as the TP, FN, FP were aggregated across all test cases. This led to a global metric value for each participant which was used to create the ranking. Please note both that the number of test cases and the number of algorithms were generally differed for each task and stage. For the binary and multi-instance segmentation tasks, the rankings were computed for both metrics, namely *(MI_)DSC* and *(MI_)NSD*, as shown in Algorithm 1.

These procedures produced nine rankings in total, namely four separate rankings (accuracy and robustness ranking for the *(MI_)DSC* and the *(MI_)NSD*) for the binary and the multi-instance segmentation task respectively and one ranking for multi-instance detection. In every ranking scheme, missing cases were set to the worst possible value, namely 0 for all metrics.

2.3.3. Statistical analyses

The stability of the rankings was investigated via bootstrapping as this approach was identified as appropriate for quantifying ranking variability (Maier-Hein et al., 2018). The analysis was performed using the R package *challengeR* (Wiesenfarth et al., 2019b; 2019a). The package was further used to create plots that visualize (1) the absolute frequency of test cases in which each algorithm achieved the different ranks and (2) the bootstrap results for each algorithm.

[7] Please note that an algorithm A with a higher rank (according to the significance ranking) than algorithm B did not necessarily perform significantly better than algorithm B, as detailed in Wiesenfarth et al. (2019b).

2.3.4. Further analyses

Expert baseline Given the imperfect reference (no perfect ground truth) resulting from human annotation, it is typically difficult to determine a plausible upper bound (optimal) performance. To address this knowledge gap, one additional labeling expert, a medical student with six years of experience in labeling (henceforth denoted 'expert') annotated all images from stage 2. Inspired by a human vs. algorithms analysis for natural image multi-label classification from Shankar et al. (2020), we used the additional data in two principal ways. Firstly, we considered the expert as an additional team and generated new rankings for both the binary and the multi-instance segmentation task using the *(MI_)DSC*. Secondly, we analyzed his performance as a function of the number of instruments present in the image. *Worst case analysis* The influence of the image artifacts and the size and number of instruments were analyzed. For this purpose, the 100 cases with the worst performance were analyzed to investigate which image artifacts cause the main failures of the algorithms.

3. Results

In total, 75 participants registered on the Synapse challenge website (Roß et al., 2019b) before the submission deadline. Aside from one team that decided to be excluded from the rankings, all teams with a working docker[8] submission were included in this paper. Their participation over the three challenge tasks and the total amount of submissions is summarized in Table 2.

3.1. Method descriptions of participating algorithms

In the following, the participating algorithms are briefly summarized based on a description provided by the participants upon submission of the challenge results. Further details can be found in Table 3.

Team caresyntax*: Single network fits all*

The *caresyntax* team's core idea for multi-instance segmentation was to apply a Mask R-CNN (He et al., 2017) based on a single network with shared convolutional layers for both branches. They hypothesized that it would help the network to generalize better if it was only provided with limited training data. The team decided to use a pre-trained version of the Mask R-CNN without including any temporal information from the videos. In their results, they reported that their approach outperformed a U-Net-based model by a significant margin. The team worked out that tuning pixel-level and mask-level confidence thresholds on the predictions played an important role. Furthermore, they acknowledged the importance that the training set size had for improved predictions, both qualitatively and quantitatively. The team participated in all three tasks using the same method. They produced the same output for the multi-instance segmentation and detection tasks and binarized the output of the multi-instance segmentation for the binary segmentation task.

Team CASIA_SRL*: Dense pyramid attention network for robust medical instrument segmentation*

The *CASIA_SRL* team proposed a network named Dense Pyramid Attention Network (Ni et al., 2020) for multi-instance segmentation. They mainly focused on two problems: Changes in illumination and surgical instruments scale changes. They proposed that an attention module should be used, which was able to capture second-order statistics, with the goal of covering semantic dependencies between pixels and capturing the global context

[8] https://www.docker.com/.

Table 2
Overview of selected participating teams over the three tasks, namely binary segmentation (BS), multi-instance detection (MID) and multi-instance segmentation (MIS).

Team identifier	BS	MID	MIS	Affiliations
caresyntax	x	x	x	[1] caresyntax, Berlin, Germany
CASIA_SRL	x		x	[1] University of Chinese Academy Sciences, Beijing, China [2] State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China
Djh	x			[1] SimulaMet, Oslo, Norway [2] Arctic University of Norway (UiT), Tromsø, Norway [3] Oslo Metropolitan University (Oslomet), Oslo, Norway
fisensee	x	x	x	[1] University of Heidelberg, Germany [2] Division of Medical Image Computing (MIC), German Cancer Research Center, Heidelberg, Germany
haoyun	x			[1] Department of Computer Science, School of Informatics, Xiamen University, Xiamen, China and School of Mechanical [2] Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China
NCT	x			[1] National Center for Tumor Diseases (NCT), Partner Site Dresden, Germany: German Cancer Research Center (DKFZ), Heidelberg, German [2] Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany [3] Helmholtz Association/Helmholtz-Zentrum Dresden - Rossendorf (HZDR), Dresden, Germany
SQUASH	x	x	x	[1] Institute of Information Technology, Klagenfurt University, Austria
Uniandes	x	x	x	[1] Universidad de los Andes, Bogotá, Colombia
VIE	x	x	x	[1] Institute of Digital Media (NELVT), Peking University, Peking, China
www	x	x	x	[1] Department of Computer Science, School of Informatics, Xiamen University, Xiamen, China [2] Department of Computer Science Engineering, The Chinese University of Hong Kong, Hong Kong, China
valid submissions	10	6	7	
invalid submissions	2	1	1	
TOTAL	12	7	8	

(Ni et al., 2020). As the scale of surgical instruments constantly changes as they move, the team introduced dense connections across scales to capture multi-scale features for surgical instruments. The team did not use the provided videos to complement the information contained in the individual frames. The team participated in the binary and multi-instance segmentation tasks. They produced the same output for the multi-instance segmentation and detection tasks and binarized the output of the multi-instance segmentation for the binary segmentation task.

Team Djh*: A RASNet-based deep learning approach for the binary segmentation task*

The *Djh* team only participated in the binary segmentation task. They used the Refined Attention Segmentation Network (Ni et al., 2019) and put a large amount of effort into data augmentation and hyperparameter tuning. Their motivation for using this architecture was its U-shape design which consists of contracting and expanding paths like the ResUNet++ (Jha et al., 2019). The RASNet is able to capture low-level and higher-level features. The team did not use the videos provided to complement the information contained in the individual frames.

Team fisensee*: OR-UNet*

Team *fisensee's* core idea was to optimize a binary segmentation algorithm and then adjust the output with a connected component analysis in order to solve the multi-instance segmentation and detection tasks (Isensee and Maier-Hein, 2020). Inspired by the recent successes of the nnU-Net (Isensee et al., 2018), the authors used a simple established baseline architecture (the U-Net (Ronneberger et al., 2015)) and iteratively improved the segmentation results through hyperparameter tuning. The method, referred to as optimized robust residual 2D U-Net (OR-UNet), was trained with the sum of *DSC* and cross-entropy loss and a multi-scale loss. During training, extensive data augmentation was used to increase robustness. For the final prediction, they used an ensemble of eight models. They hypothesized that ensembles perform better than a single network. In their report, the team wrote that they attempted to use the temporal information by stacking previous frames but did not observe a performance gain. Additionally, they noticed that in many cases, instruments did not touch thus they used a connected component analysis (Shapiro, 1996) to separate instrument instances.

Team haoyun*: Robust medical instrument segmentation using enhanced DeepLabV3+*

The *haoyun* team only participated in the binary segmentation task. They based their work on the DeepLabV3+ (Chen et al., 2018) architecture in order to focus on high-level information. To enrich the receptive fields, they used a pre-trained ResNet-101 (He et al., 2016) with dilated convolutions as encoder. To train their network, the team combined the *DSC* with the focal loss (Lin et al., 2017) in order to focus more on less accurate pixels and challenging images. In addition, the team used a 5-fold cross validation to improve both generalization and stability of the network. They did not use the provided videos to complement the information contained in the individual frames.

Team NCT*: Robust medical instrument segmentation in robot-assisted surgery using deep convolutional neuronal network*

The *NCT* team only participated in the binary segmentation task. They used a TernausNet with a pre-trained VGG16 network (Iglovikov and Shvets, 2018) as TernausNet had already showed promising results in two previous MICCAI EndoVis segmentation challenges from 2017 and 2018 (Allan et al., 2019). The team did not use the provided videos to complement the information contained in the individual frames.

Table 3
Overview of submitted methods. Abbreviations are as follows: Stochastic gradient descent (SGD) (Kiefer et al., 1952), adaptive moment estimation (Adam) (Kingma and Ba, 2014).

Team	Basic architecture	Video data used?	Additional data used?	Loss functions	Data augmentation	Optimizer
caresyntax	Mask R-CNN (He et al., 2017) (backbone: ResNet-50 (He et al., 2016))	No	ResNet-50 pre-trained on MS-COCO (Lin et al., 2014)	Smooth L1 loss, cross entropy loss, binary cross entropy loss	Applied in each epoch: Random flip (horizontally) with probability 0.5	SGD (Kiefer et al., 1952)
CASIA_SRL	Dence Pyramid Attention Network (Ni et al., 2020) (backbone: ResNet-34 (He et al., 2016))	No	ResNet-34 backbone pre-trained on ImageNet (Russakovsky et al., 2015)	Hybrid loss: cross entropy $-\alpha \log(Jaccard)$	Data augmented once before training: Random rotation, shifting, flipping	Adam (Kingma and Ba, 2014)
Djh	RASNet (Ni et al., 2019)	No	ResNet50 (He et al., 2016) pre-trained on ImageNet (Russakovsky et al., 2015)	*DSC* coefficient loss	Applied on the fly on each batch: Crop (random and center), flip (horizontally and vertically), scale, cutout, greyscale	Adam (Kingma and Ba, 2014)
fisensee	2D U-Net (Ronneberger et al., 2015) with residual encoder	No	No	Sum of *DSC* and cross-entropy loss	Randomly applied on the fly on each batch: Rotation, elastic deformation, scaling, mirroring, Gaussian noise, brightness, contrast, gamma	SGD (Kiefer et al., 1952)
haoyun	DeepLabV3+ (Chen et al., 2018) with ResNet-101 (He et al., 2016)) encoder	No	ResNet-101 pre-trained on ImageNet (Russakovsky et al., 2015)	Logarithmic *DSC* loss	Applied on the fly on each batch: Flip (vertically), crop (random)	Adam (Kingma and Ba, 2014)
NCT	TernausNet (Iglovikov and Shvets, 2018), replaced ReLU with eLU (Clevert et al., 2015)	No	VGG16 pre-trained on ImageNet (Russakovsky et al., 2015)	Weighted binary cross entropy in combination with Jaccard Index	Applied on the fly on each batch: Flips (horizontally and vertically), rotations of $[-10, 10]^\circ$, image contrast manipulations (brightness, blur, motion-blur)	Adam (Kingma and Ba, 2014)
SQUASH	Mask R-CNN (He et al., 2017) (backbone: ResNet-50 (He et al., 2016))	*Yes, to* estimate the probability that last frame of video shows instrument instance	No	ResNet-50: Focal loss, Mask R-CNN: Mask R-CNN loss + cross entropy loss	35% of total input for classification: Gaussian blur, sharpening, gamma contrast enhancement; additional 35% of images: Mirroring (along x- and y-axes); minority class: Translation (horizontally); non-instrument image frames are not processed	SGD (Kiefer et al., 1952)
Uniandes	Mask R-CNN (He et al., 2017) (backbone: ResNet-101 (He et al., 2016))	Yes, for data augmentation	Pre-trained on MS-COCO (Lin et al., 2014)	Standard Mask R-CNN loss functions	Applied on the fly on each batch: Random flips (horizontally), propagation of annotation backwards to previous video frames	SGD (Kiefer et al., 1952)
VIE	Mask R-CNN (He et al., 2017) (backbone: ResNet-50 (He et al., 2016))	Yes, calculating the optical flow over 5 frames	No	RPN class loss, MASK R-CNN loss	Applied on the fly on each batch: Image resizing (1024x1024), bounding boxes, label generation	N/A
www[9]	Mask R-CNN (He et al., 2017) (backbone: ResNet-50 (He et al., 2016))	No	Pre-trained[9] on ImageNet (Russakovsky et al., 2015)	Smooth L1 loss, focal loss, binary cross entropy loss	Applied on the fly on each batch: Random flip (horizontally and vertically), rotations of $[0, 10]^\circ$	Adam (Kingma and Ba, 2014)

Team SQUASH*: An ensemble of models, combining image frame classification and multi-instance segmentation*

Team *SQUASH's* hypothesis was that they could increase the robustness and generalizability of all challenge tasks simultaneously by using multiple recognition task training. In training their method from scratch, they assumed that the network capabilities were fully utilized to learn detailed instrument features. Based on a ResNet50 (He et al., 2016), the team used the video data provided and built a classification model in order to predict all instrument frames in a sequence of video frames. On top of this classification model, they built a segmentation model by employing a Mask R-CNN (He et al., 2017) to detect multiple instrument instances in the image frames. The segmentation model was trained by leveraging the preliminary trained classification model on instrument images as a feature extractor to deepen the learning of the task of instrument segmentation. Both models were combined in a two-stage framework to process a sequence of video frames. The team reported that their method had trouble dealing with instrument occlusions, but on the other hand, they were surprised to find that it handled reflections and black borders well.

Team Uniandes*: Instance-based instrument segmentation with temporal information*

Team *Uniandes* based their multi-instance segmentation approach on the Mask R-CNN (He et al., 2017). For training purposes, they created an experimental framework with a training and validation split as well as supplementary metrics in order to identify the best version of their method and gain insight into the performance and limitations. Data augmentation was performed by calculating the optical flow with a pre-trained FlowNet2 (Ilg et al., 2017) and using the flow to map the reference annotation on to the previous frames. However, they did not find significant benefits in using the augmentation technique. The team participated in all three tasks. They produced the same output for the multi-instance segmentation and detection tasks and binarized the output of the multi-instance segmentation for the binary segmentation task. The team observed that their approach was limited in terms of finding all instruments in an image frame, but once an instrument was found it was segmented with a high *DSC* score. Although the team achieved good metric scores they stated that they fell short in segmenting small or partial instruments and instruments covered by smoke.

Team VIE*: Optical flow-based instrument detection and segmentation*

The *VIE* team approached the multi-instance segmentation task with an optical flow-based method. Their hypothesis was that the detection of moving parts in the image enables medical instruments to be detected and segmented. For their approach, they calculated the optical flow over the last five frames of a case by using the OpenCV[9] library and concatenated the optical flow with the raw image as input for a Mask R-CNN (He et al., 2017). The team assumed that this would reduce most of unnecessary clutter segmentation. The team participated in all three tasks. They produced the same output for the multi-instance segmentation and detection tasks and binarized the output of the multi-instance segmentation for the binary segmentation task. The team hypothesized that the temporal data could have been used more effectively.

Team www*: Integration of Mask R-CNN and DAC block*[9]

Team *www* proposed that a framework based on Mask R-CNN (He et al., 2017) to handle the three tasks in the challenge. Based on the observation that the instruments have variable sizes, their idea was to enlarge the receptive field and tune the anchor size for the Mask R-CNN. In addition, the team integrated DAC blocks (Gu et al., 2019) into the framework to collect more information. The team participated in all three tasks. They produced the same output for the multi-instance segmentation and detection tasks and binarized the output of the multi-instance segmentation for the binary segmentation task. The team reported that including temporal information might have helped to improve their performance.[10]

3.2. Individual performance results for participating teams

The teams' individual performances in both segmentation tasks are presented in Fig. 2 and Table 4. The dot- and boxplots show the metric values for each algorithm over all test cases in stage 3.

3.3. Challenge rankings for stage 3

As described in Section 2.3.2, an accuracy and a robustness ranking were computed for both metrics of the segmentation tasks (resulting in 4 rankings for each task). These are shown in Tables 5 and 7. For the multi-instance detection task, the *F1-score* was computed for each participant (see Table 6). The metric computation already included aggregated values, therefore only one ranking was computed for this task.

To provide deeper insight in the ranking variability, ranking heatmaps (see Fig. 3) and blob plots (see Fig. 4) were computed for all rankings of both segmentation tasks. Ranking heatmaps were used to visualize the challenge assessment data (Wiesenfarth et al., 2019b). Blob plots were used to visualize ranking stability based on bootstrap sampling (Wiesenfarth et al., 2019b).

The computed rankings for the remaining stages are given in Appendix D.

3.4. Comparison across all stages

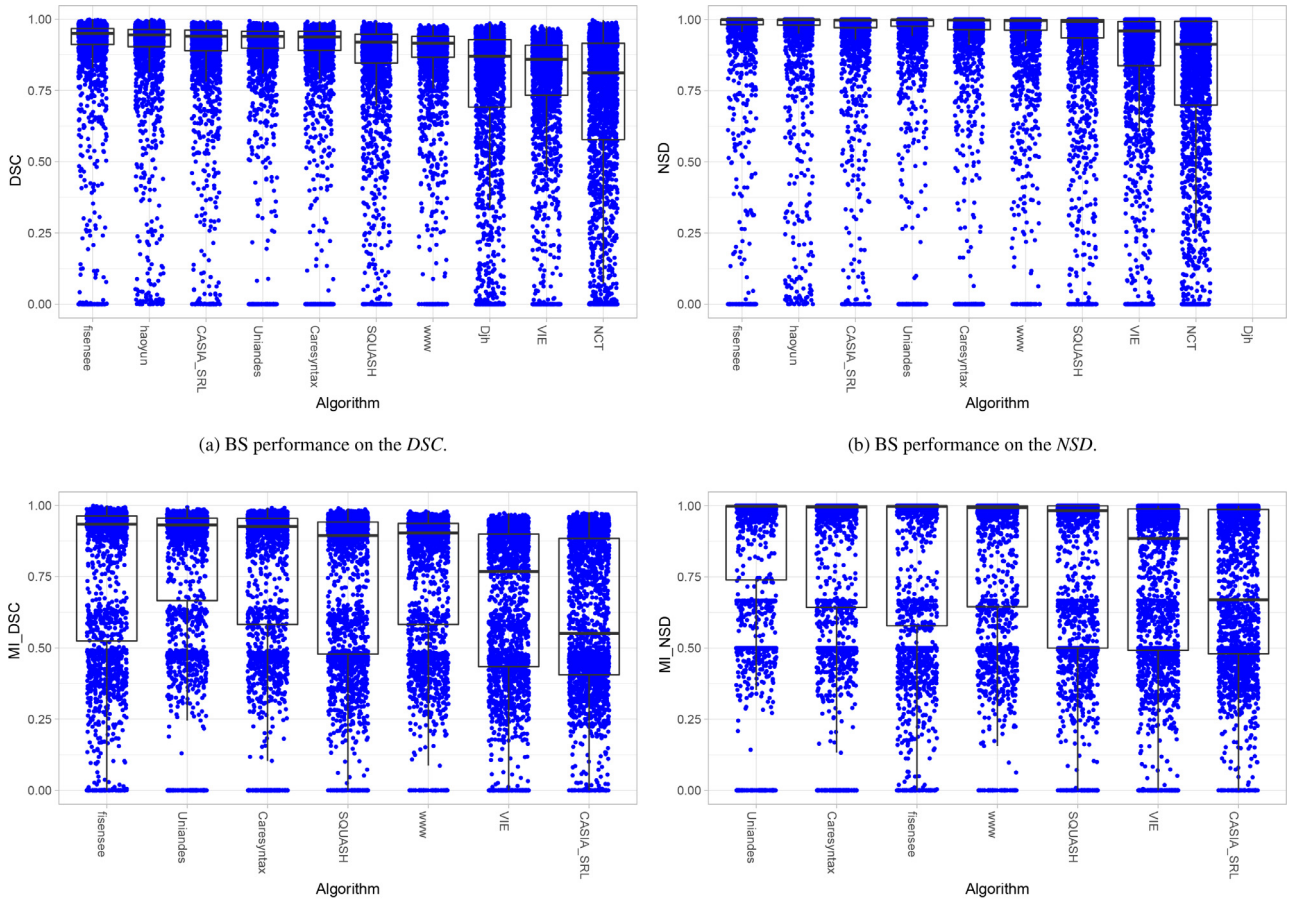
Fig. 5 shows the comparison of the average *(MI_)DSC* performances of the participating algorithms over the three evaluation stages (see Section 2) for both segmentation tasks. For this purpose, boxplots were generated for both tasks over the average metric values per team. A clear performance drop is visible in line with the increasing difficulty of the stages: Average performance produces median values of 0.88 (min: 0.73, max: 0.92) for the binary segmentation task and 0.80 (min: 0.65, max: 0.84) for the multi-instance segmentation task for stage 1. For stage 2, the median metric values decrease to 0.87 (min: 0.76, max: 0.90) and 0.78 (min: 0.64, max: 0.84) and finally, the performance for stage 3 resulted in a median of 0.85 (min: 0.69, max: 0.89) and 0.76 (min: 0.60, max: 0.80).

3.5. Further analysis

Expert baseline Only the rankings of the *(MI_)DSC* metrics were used to compare the algorithms' performances with that of a human annotator, as similar results were obtained for the *(MI_)NSD*. As images in stage 2 contain only a maximum of three instrument instances, the analysis can only show differences for n instances, where $n \in \{1, 2, 3\}$. In both tasks, the expert is the winner for both rankings. Team *fisensee* shares the first rank with the expert in the accuracy rankings for the binary segmentation task and the

[9] https://opencv.org/.

[10] Please note that this team used data from the EndoVis 2017 challenge (Allan et al., 2019) to visually check their performance on a different medical data set. The participation policies (see Appendix A) prohibit the use of other medical data for algorithm training or hyperparameter tuning. The challenge organizers defined this case as a grey zone but noted that the team may have had a competitive advantage in terms of performance generalization.



(a) BS performance on the *DSC*.

(b) BS performance on the *NSD*.

(c) MIS performance on the *MI_DSC*.

(d) MIS performance on the *MI_NSD*.

Fig. 2. Dot- and boxplots showing the individual performances of algorithms on the binary segmentation (BS; top) and multi-instance segmentation (MIS; bottom) tasks. The (multi-instance) Dice Similarity Coefficient (*(MI_)DSC*; left) and the (multi-instance) Normalized Surface Distance (*(MI_)NSD*; right) were used as metrics.

Table 4
Quantitative results of all participating methods for all three stages for the tasks binary and multi-instance segmentation. The metrics are *DSC* for the binary and *MI_DSC* for the multi-instance segmentation task. The table contains the mean, median and the 5th (Q05), 25th (Q25), 75th (Q75) and 95th (Q95) quantile for each metric.

Binary instance segmentation

Team	Stage 1						Stage 2						Stage 3					
	Mean	Median	Q5	Q25	Q75	Q95	Mean	Median	Q05	Q25	Q75	Q95	Mean	Median	Q5	Q25	Q75	Q95
CASIA_SRL	0.90	0.95	0.70	0.91	0.96	**0.98**	0.89	0.95	0.43	0.91	**0.97**	**0.98**	0.88	0.94	0.50	0.89	0.96	**0.98**
caresyntax	0.89	0.94	0.69	0.91	0.96	**0.98**	0.88	0.95	0.36	0.91	0.96	**0.98**	0.85	0.94	0.00	0.89	0.96	0.97
Djh	0.81	0.90	0.08	0.81	0.94	0.96	0.79	0.90	0.03	0.78	0.94	0.97	0.75	0.87	0.00	0.69	0.93	0.96
NCT	0.73	0.87	0.04	0.62	0.94	0.97	0.76	0.86	0.11	0.68	0.94	0.97	0.69	0.81	0.00	0.58	0.92	0.97
SQUASH	0.88	0.93	0.55	0.88	0.95	0.97	0.85	0.93	0.34	0.87	0.95	0.97	0.83	0.92	0.22	0.85	0.95	0.97
Uniandes	0.90	0.94	0.71	0.91	0.96	0.97	0.89	0.95	0.41	0.92	0.96	0.97	0.87	0.94	0.28	0.90	0.96	0.97
VIE	0.79	0.87	0.30	0.76	0.92	0.95	0.77	0.87	0.00	0.74	0.91	0.95	0.76	0.86	0.00	0.73	0.91	0.94
fisensee	**0.92**	**0.96**	**0.76**	**0.93**	**0.97**	**0.98**	**0.90**	**0.96**	**0.54**	**0.93**	**0.97**	**0.98**	0.88	**0.95**	0.34	**0.91**	**0.97**	**0.98**
haoyun	0.90	0.95	0.64	0.91	0.96	**0.98**	0.89	0.95	0.42	0.91	**0.97**	**0.98**	**0.89**	0.94	**0.52**	0.90	0.96	**0.98**
www	0.88	0.92	0.68	0.88	0.94	0.96	0.86	0.92	0.37	0.88	0.94	0.95	0.85	0.91	**0.52**	0.86	0.94	0.95
expert	-	-	-	-	-	-	***0.91***	***0.96***	***0.73***	***0.93***	***0.97***	***0.98***	-	-	-	-	-	-

Multi-instance segmentation

Team	*Stage 1*						*Stage 2*						*Stage 3*					
	Mean	Median	Q5	Q25	Q75	Q95	Mean	Median	Q05	Q25	Q75	Q95	Mean	Median	Q5	Q25	Q75	Q95
CASIA_SRL	0.65	0.69	0.24	0.44	0.91	0.96	0.64	0.68	0.18	0.43	0.91	0.96	0.60	0.55	0.19	0.41	0.88	0.95
caresyntax	0.82	0.93	0.32	0.83	0.96	0.97	**0.80**	**0.94**	0.32	0.68	0.96	**0.98**	0.77	**0.93**	0.00	0.58	0.95	0.97
SQUASH	0.78	0.90	0.32	0.60	0.94	0.97	0.75	0.91	0.26	0.48	0.95	0.97	0.73	0.89	0.22	0.48	0.94	0.97
Uniandes	**0.84**	**0.94**	**0.40**	**0.88**	0.96	0.97	0.84	**0.94**	**0.39**	**0.88**	0.96	0.97	**0.80**	**0.93**	0.26	**0.67**	0.95	0.97
VIE	0.67	0.81	0.16	0.45	0.90	0.95	0.65	0.77	0.00	0.43	0.90	0.95	0.65	0.77	0.00	0.43	0.90	0.94
fisensee	0.80	**0.94**	0.32	0.62	**0.97**	**0.98**	**0.80**	**0.94**	0.28	0.61	**0.97**	**0.98**	0.76	**0.93**	0.17	0.52	**0.96**	**0.98**
www[10]	0.81	0.90	0.37	0.79	0.94	0.96	0.78	0.91	0.30	0.63	0.94	0.96	0.76	0.89	**0.31**	0.58	0.93	0.95
expert	-	-	-	-	-	-	***0.88***	***0.95***	***0.47***	***0.91***	***0.97***	**0.98**	-	-	-	-	-	-

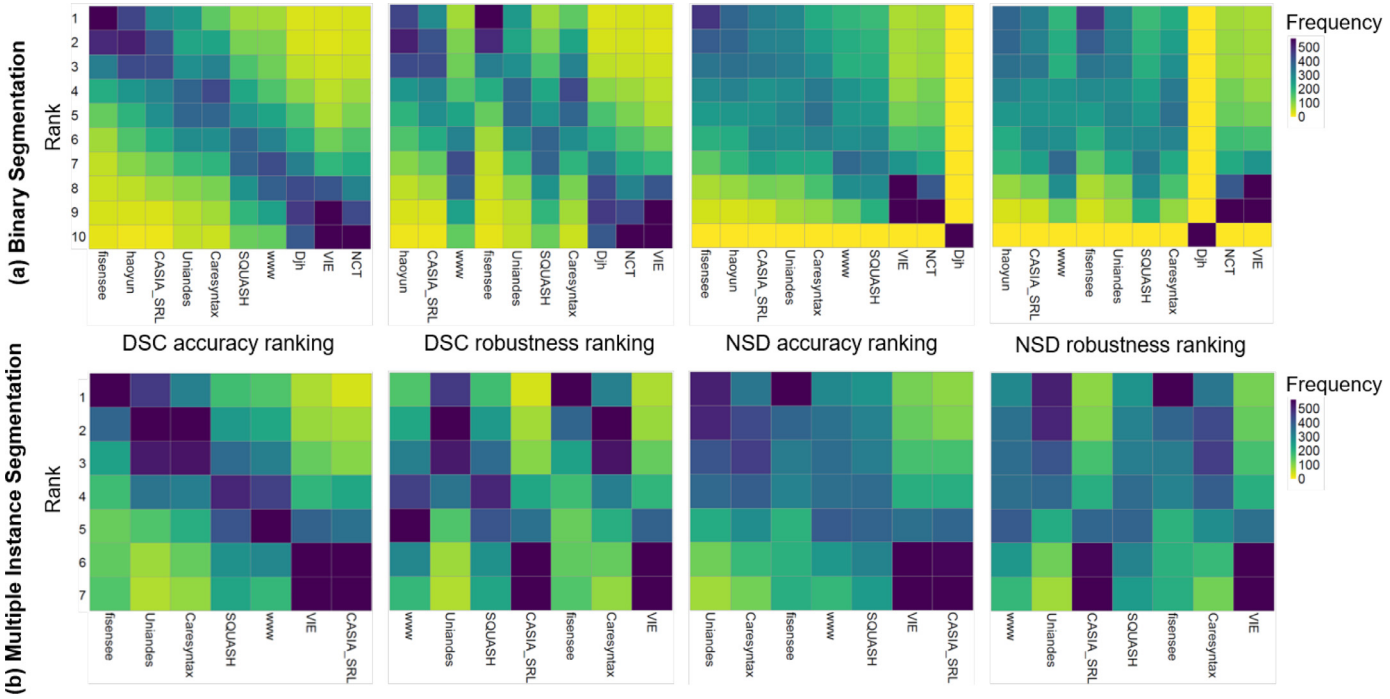


Fig. 3. Ranking heatmaps for the four rankings in the binary segmentation and multi-instance segmentation tasks. Each cell (i, A_j) shows the absolute frequency of test cases in which algorithm A_j achieved rank i. The plots were generated using the package challengeR (Wiesenfarth et al., 2019b; 2019a).

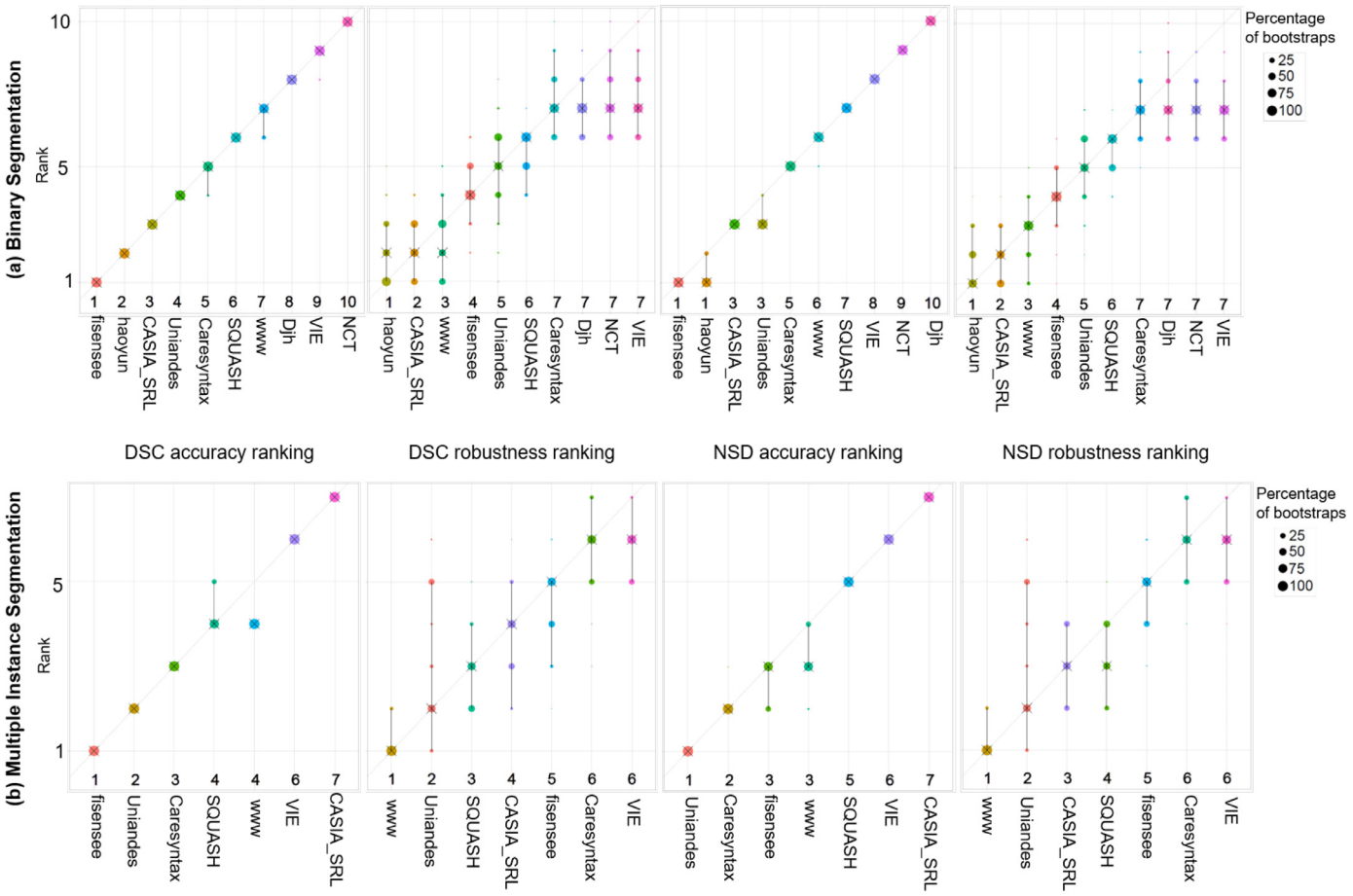


Fig. 4. Blob plots for the four rankings in the binary segmentation and multi-instance segmentation tasks. Blob plots are used to visualize ranking stability based on bootstrap sampling. Algorithms are color-coded, and the area of each blob at position (A_i, rank j) is proportional to the relative frequency A_i of the achieved rank j across $b = 1000$ bootstrap samples. The median rank for each algorithm is indicated by a black cross. 95% bootstrap intervals across bootstrap samples are indicated by black lines. The plots were generated using the package challengeR (Wiesenfarth et al., 2019b; 2019a).

Table 5
Binary segmentation: Rankings for stage 3 of the challenge. The upper part of the table shows the Dice Similarity Coefficient (*DSC*) rankings and the lower part shows the Normalized Surface Distance (*NSD*) rankings (accuracy rankings on the left, robustness rankings on the right). Each ranking contains a team identifier, either a proportion of significant tests divided by the number of algorithms (prop. sign.) for the accuracy ranking or an aggregated *DSC*/*NSD* value (aggr.*DSC*/*NSD* value) and a rank.

DSC: ACCURACY RANKING			DSC: ROBUSTNESS RANKING		
Team identifier	Prop. Sign.	Rank	Team identifier	Aggr. *DSC* Value	Rank
fisensee	**1.00**	**1**	***haoyun***	**0.52**	**1**
haoyun	0.89	2	*CASIA_SRL*	0.50	2
CASIA_SRL	0.78	3	*www*[10]	0.49	3
Uniandes	0.67	4	*fisensee*	0.34	4
caresyntax	0.56	5	*Uniandes*	0.28	5
SQUASH	0.44	6	*SQUASH*	0.22	6
www[10]	0.33	7	*caresyntax*	0.00	7
Djh	0.22	8	*Djh*	0.00	7
VIE	0.11	9	*NCT*	0.00	7
NCT	0.00	10	*VIE*	0.00	7
NSD: ACCURACY RANKING			*NSD*: ROBUSTNESS RANKING		
Team identifier	Prop. Sign.	Rank	Team identifier	Aggr. *NSD*Value	Rank
haoyun	**0.89**	**1**	***haoyun***	**0.63**	**1**
fisensee	**0.89**	**1**	*CASIA_SRL*	0.62	2
CASIA_SRL	0.67	3	*www*[10]	0.57	3
Uniandes	0.67	3	*fisensee*	0.45	4
caresyntax	0.56	5	*Uniandes*	0.32	5
www[10]	0.44	6	*SQUASH*	0.26	6
SQUASH	0.33	7	*caresyntax*	0.00	7
VIE	0.22	8	*Djh*	0.00	7
NCT	0.11	9	*NCT*	0.00	7
Djh	0.00	10	*VIE*	0.00	7

Table 6
Multi-instance detection: Ranking for the mean average precision (*mAP*) in stage 3 of the challenge.

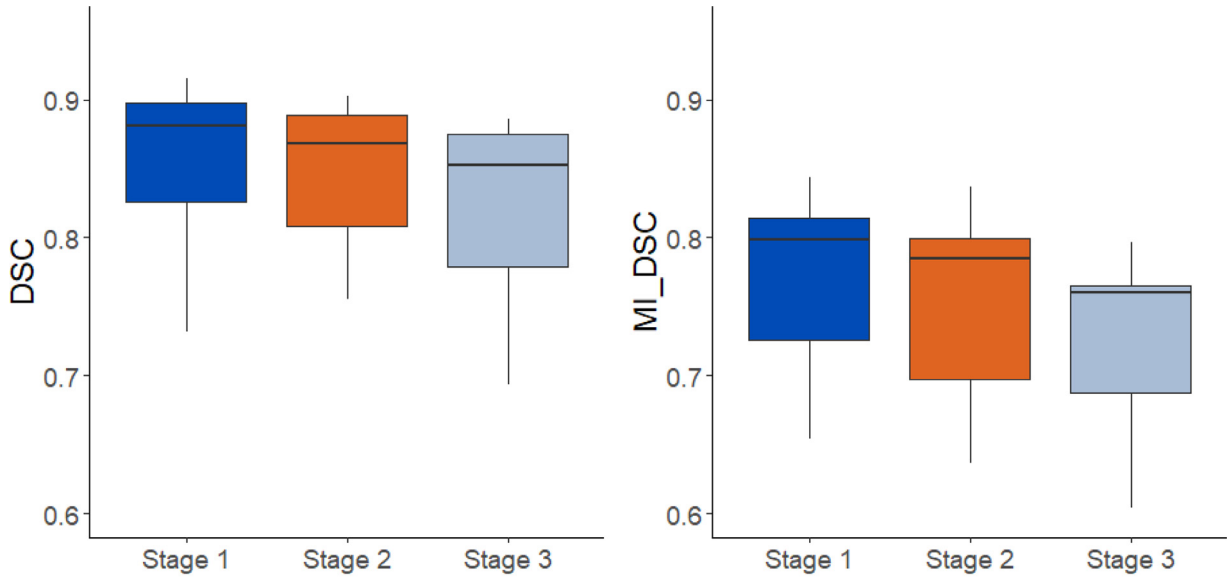
Team identifier	F1-score	Rank
Uniandes	0.91	**1**
www[10]	0.90	2
caresyntax	0.89	3
SQUASH	0.86	4
fisensee	0.86	5
VIE[9]	0.82	6

multi-instance segmentation task for frames with 1 instrument. The mean segmentation accuracy per instrument instance can be seen in Fig. 6.

Worst case analysis For further analyses, we investigated the image frames that produced the 100 best or worst metric values of participating teams. This investigation revealed the strengths and weaknesses of the proposed methods. In general, algorithm performance drops with the number of instruments in the image as illustrated in Fig. 7. The algorithms succeeded in images containing reflections, blood, different illuminations and in finding the

Table 7
Multi-instance segmentation: Rankings for stage 3 of the challenge. The upper part of the table shows the multi-instance Dice Similarity Coefficient (*MI_DSC*) rankings and the lower part shows the multi-instance Normalized Surface Distance (*MI_NSD*) rankings (accuracy rankings on the left, robustness rankings on the right). Each ranking contains a team identifier, either a proportion of significant tests divided by the number of algorithms (prop. sign.) for the accuracy ranking or an aggregated *MI_DSC*/*MI_NSD* value (aggr. *MI_DSC*/*MI_NSD* value) and a rank.

MI_DSC: ACCURACY RANKING			MI_DSC: ROBUSTNESS RANKING		
Team identifier	Prop. Sign.	Rank	Team identifier	Aggr. *MI_DSC*Value	Rank
fisensee	**1.00**	**1**	***www***[10]	**0.31**	**1**
Uniandes	0.83	2	*Uniandes*	0.26	2
caresyntax	0.67	3	*SQUASH*	0.22	3
SQUASH	0.33	4	*CASIA_SRL*	0.19	4
www[9]	0.33	4	*fisensee*	0.17	5
VIE	0.17	6	*caresyntax*	0.00	6
CASIA_SRL	0.00	7	*VIE*	0.00	6
MI_NSD: ACCURACY RANKING			*MI_NSD*: ROBUSTNESS RANKING		
Team identifier	Prop. Sign.	Rank	Team identifier	Aggr. *MI_NSD*Value	Rank
Uniandes	**1.00**	**1**	***www***[10]	**0.35**	**1**
caresyntax	0.67	2	*Uniandes*	0.29	2
fisensee	0.50	3	*CASIA_SRL*	0.27	3
www[9]	0.50	3	*SQUASH*	0.26	4
SQUASH	0.33	5	*fisensee*	0.16	5
VIE	0.17	6	*caresyntax*	0.00	6
CASIA_SRL	0.00	7	*VIE*	0.00	6



(a) BS (b) MIS

Fig. 5. Boxplots of the variance across all test images for the (a) binary segmentation task with the Dice Similarity Coefficient (*DSC*) and (b) the multi-instance segmentation task with the Multi-instance Dice Similarity Coefficient (*(MI_)DSC*) for stages 1 to 3. The boxplots show the average algorithm performances (mean over all participant predictions per image) per image.

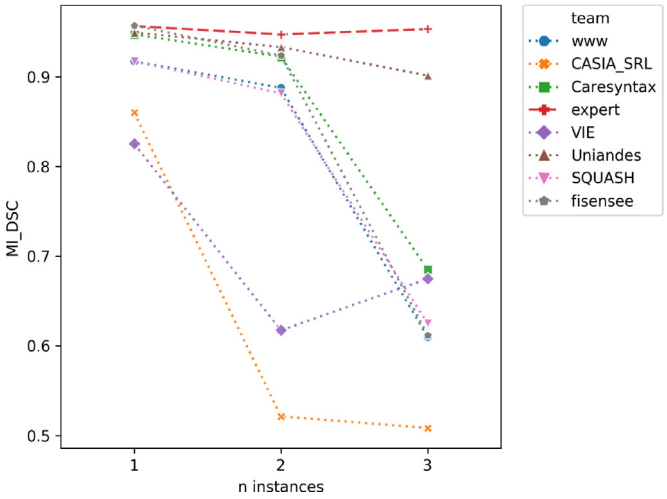


Fig. 6. *Median MI_DSC* as a function of the number of instruments in the image for stage 2 of the test data for the multi-instance segmentation task. It shows the performance of all algorithms in comparison to the human expert. Clearly, all algorithms' performance drops with the number of visible instruments in the image while the experts performance stays constant..

inside of the trocar (see Fig. 8). Problems still arose in image frames which contained small and transparent instruments. False positives (mainly objects that were not defined as instruments) turned out to be a problem for all tasks. Furthermore, algorithm performance was poor for images with instruments, close to another as well as crossing, partially hidden or moving instruments, instruments close to the image border and images containing smoke (see Fig. 9 and 10).

4. Discussion

We organized the first challenge in the field of surgical data science that (1) included tasks on multi-instance detection/tracking and (2) placed particular emphasis on the robustness and generalization capabilities of the algorithms. The key insights are:

1. Competing methods: These state-of-the-art methods are exclusively based on deep learning with a specific focus on U-Nets (Ronneberger et al., 2015) (binary segmentation) and Mask R-CNNs (He et al., 2017) (multi-instance detection and segmentation). For binary segmentation, the U-Net and the new DeepLabV3 architecture yielded an equally strong performance. For the multi-instance segmentation, a U-Net in combination with a connected component analysis was a strong baseline, but a Mask R-CNN approach was more promising overall, especially in terms of robustness.
2. Performance:
   (a) Binary segmentation: The mean performances of the winning algorithms for the accuracy ranking (*DSC* of 0.88) and the robustness ranking (*DSC* of 0.89) were similar to that of the previous winners of binary segmentation challenges (winner of the EndoVis Instrument Segmentation and Tracking Challenge 2015[11]: *DSC* of 0.84; winner of the EndoVis 2017 Robotic Instrument Segmentation Challenge (Allan et al., 2019): *DSC* of 0.88). Given the high complexity of ROBUST-MIS' data in comparison to previously released data sets, we attribute the fact that the performances are similar to the high amount of training data.
   (b) Multi-instance detection: The top three algorithms achieved *F1-score* $\geq$ 0.89 for stage 3. The winning algorithms featured very high accuracy, robustness and generalization capabilities. The few failure cases were related to the detection of small instruments, instruments close to another or instruments close to the image border.
   (c) Multi-instance segmentation: The mean *MI_DSC* scores for the winning algorithm of the accuracy ranking were
   - 0.82 for cases with one instrument instance,
   - 0.71 for cases with two instrument instances,
   - 0.62 for cases with three instrument instances,
   - 0.45 for cases with more than three instrument instances.

[11] https://endovissub-instrument.grand-challenge.org/.

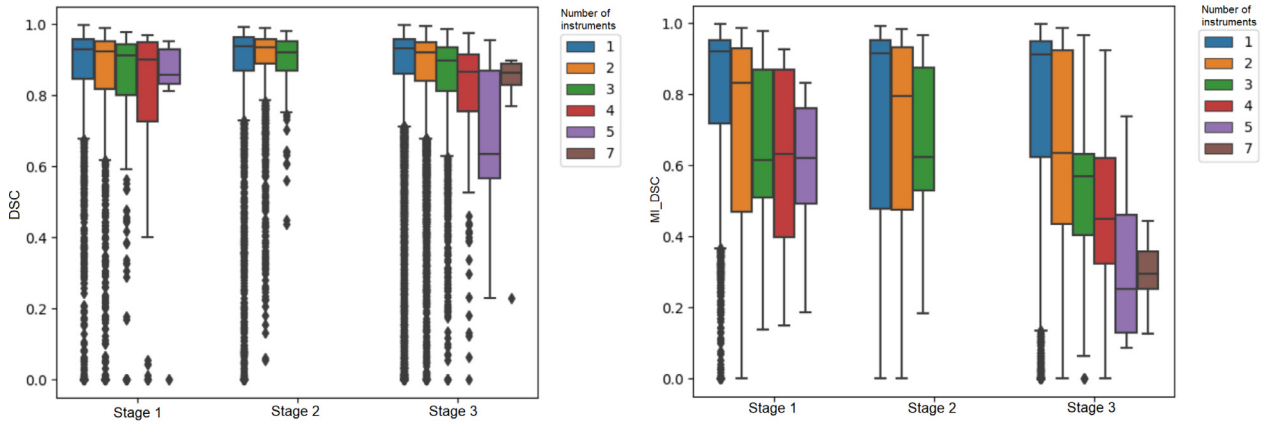


Fig. 7. Boxplots of mean (multi-instance) Dice Similaritiy Coefficient (*(MI_)DSC*) values of participating algorithms for the binary and multi-instance segmentation tasks for stages 1 to 3 stratified by the number of instruments in the video frames.

Multi-instance segmentation in endoscopic video data, therefore cannot be regarded as a solved problem.

3. Generalization: All participating methods for the binary segmentation tasks had a satisfying generalization capability over all three stages, with a median drop from 0.88 (stage 1) to 0.85 (stage 3; 3%). The generalization capabilities for the multi-instance segmentation were slightly worse, with a median drop form 0.80 (stage 1) to 0.76 (stage 3; 5%).
4. Robustness: The most successful algorithms are robust to reflections, blood and smoke. The segmentation of small, close positioned, transparent, moving, overlapping and crossing instruments, however, remains a great challenge that needs to be addressed.

The following sections provide a detailed discussion on the challenge infrastructure (Section 4.1.1), challenge data (Section 4.1.3), challenge methods (Section 4.2.1) and challenge results (Section 4.2.2).

4.1. Challenge design

In this section, we discuss the infrastucture and the data of our challenge.

4.1.1. Challenge infrastructure

We decided to use Synapse[12] as our challenge platform as it is the underlying platform of the well-known and DREAM challenges[13], and, as such, provides a complete and easy to use environment for both challenge participants and organizers. Furthermore, in addition to helping organizers monitor on how a challenge should be structured, it also helps them to follow current best practices by relying on docker submissions. However, while the overall experience with Synapse was very good, downloading the data was a problem due to slow download rates, which were dependent on the global download location and the size of the data set (about 400 GB). Unlike the data download, the docker upload was very quick and easy to follow.

The submission of docker containers and complete evaluation is already in common usage in other disciplines (e.g. CARLA[14]). However, most of the very recent challenges in the biomedical image analysis community still use plain results submissions (e.g. BraTS[15], KiTS2019[16], PAIP 2019[17]). We believe that using dockers for the evaluation is the best way as it can help (1) to avoid test data set overfitting and (2) to prevent potential instances of fraud such as manually labeling the test data (Reinke et al., 2018). However, using docker containers also means more work for the individual participants (in creating of the docker containers) and for the organizers. In addition to providing the Computing Processing Unit (CPU) and Graphics Processing Unit (GPU) resources, they have to provide support for docker related questions and must have a strategy for dealing with invalid submissions (e.g. allowing re-submission). In our challenge for example, submitted dockers were run on a small proportion of the training set to check whether the submissions worked. For five participants, the first submission failed. They were allowed to re-submit but we manually checked whether the network parameters had changed.

4.1.2. Metrics and ranking

Following recommendations of the Medical Segmentation Decathlon (Cardoso, 2018), we decided to use two metrics for the segmentation task; an overlap measure (*DSC*) and a distance measure (*NSD*). We used a non-global *DSC* for the multi-instance segmentation, meaning that the *DSC* values of instrument instances were first averaged to get an image-based score before taking the mean over all images. Another option would have been to use a global *DSC* measure, which would compute the *DSC* score globally over the complete data set and all instrument instances. However, we decided to use the non-global metric to give higher weight to small instruments.

To put a particular focus on the robustness of the methods, we decided to compute a dedicated ranking for the 5% percentile performance of the methods, as summarized in Section 2.3.2. Given our previous work on ranking stability (Maier-Hein et al., 2018), it can be assumed that a ranking based on the 5% percentile would naturally lead to less robust rankings compared to an aggregation with the mean or the median. This is one possible explanation for the fact that the ranking stability for the robustness ranking was worse compared to that of the accuracy ranking, as shown in Fig. 4.

Initially, during the challenge event at the MICCAI conference, the mean average precision (*mAP*) (Everingham, Van Gool, Williams, Winn, Zisserman, 2010) metric was used (results are provided in Table D.1) to determine the best performing algorithm. However, due to an error in the implementation and missing con-

[12] https://www.synapse.org/.
[13] http://dreamchallenges.org/.
[14] https://carlachallenge.org/.
[15] http://braintumorsegmentation.org/.
[16] https://kits19.grand-challenge.org/rules/.
[17] https://paip2019.grand-challenge.org/.

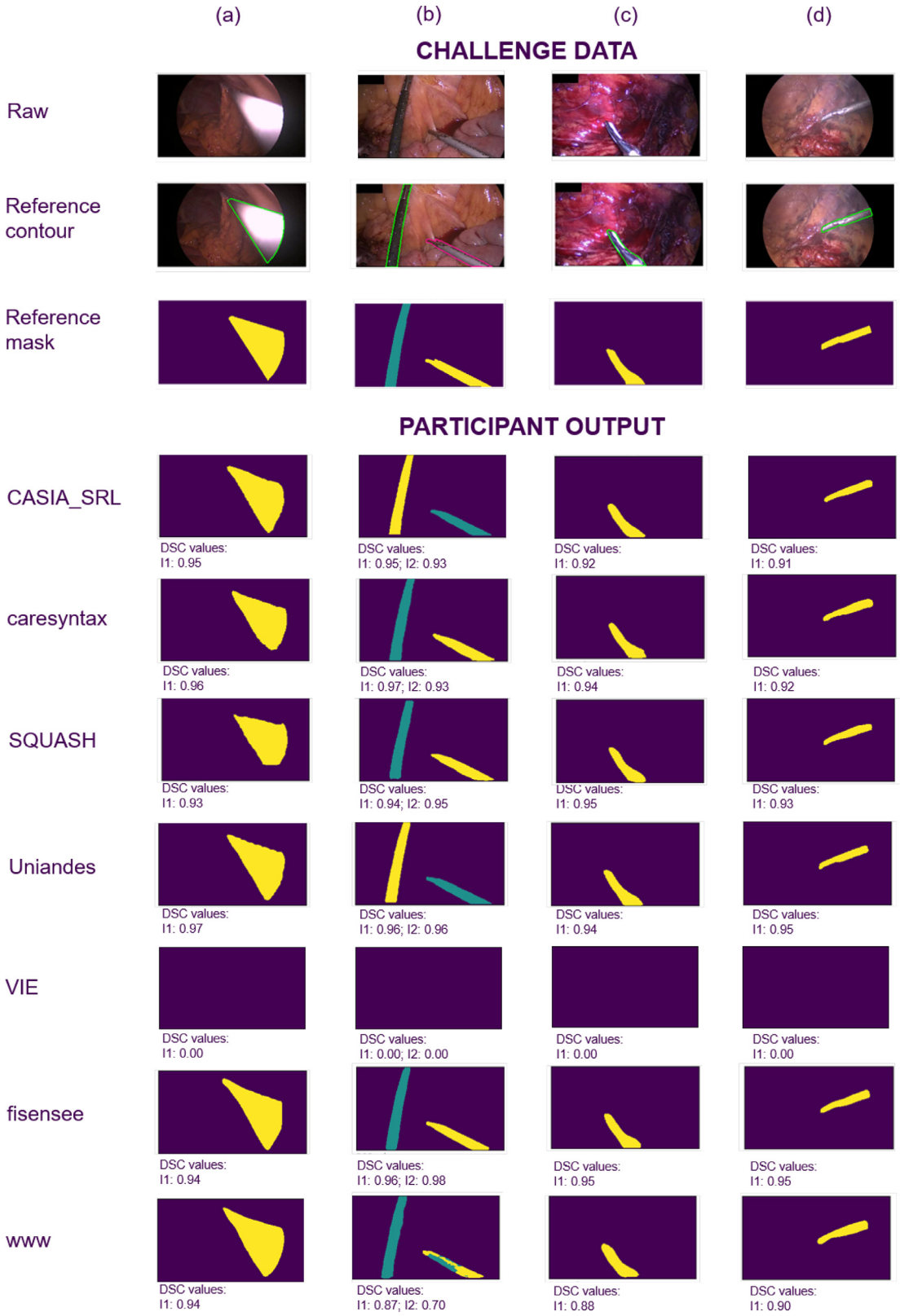


Fig. 8. Test cases with high corresponding algorithm performances. Each row shows the raw frame, the reference contours and mask as well as the algorithm output of the participating teams of the multi-instance segmentation (MIS) task for one representative frame. The columns represent image frames with (a) overexposure, (b) clearly separated instruments, (c) blood and reflections, (d) smoke, respectively. For participants, the Dice Similarity Coefficient (DSC) values are provided per instrument instance (I_i).

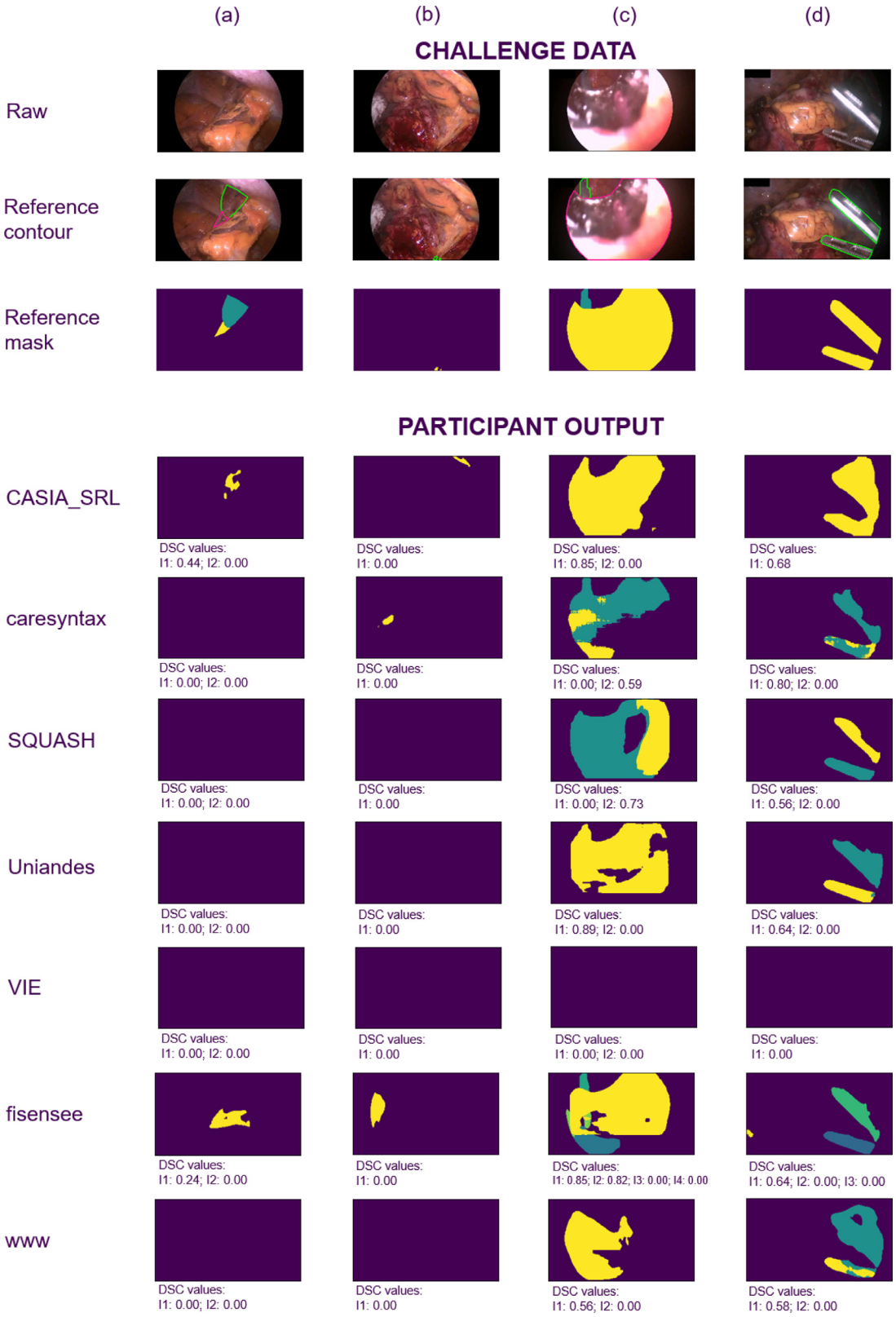


Fig. 9. Test cases with low corresponding algorithm performances. Each row shows the raw frame, the reference contours and mask as well as the algorithm output of the participating teams of the multi-instance segmentation (MIS) task for one representative frame. The columns represent image frames with (a) transparency, (b) small instruments, (c) overlapping instruments, (d) instruments near the border, respectively. For participants, the Dice Similarity Coefficient (DSC) values are provided per instrument instance (I_i).

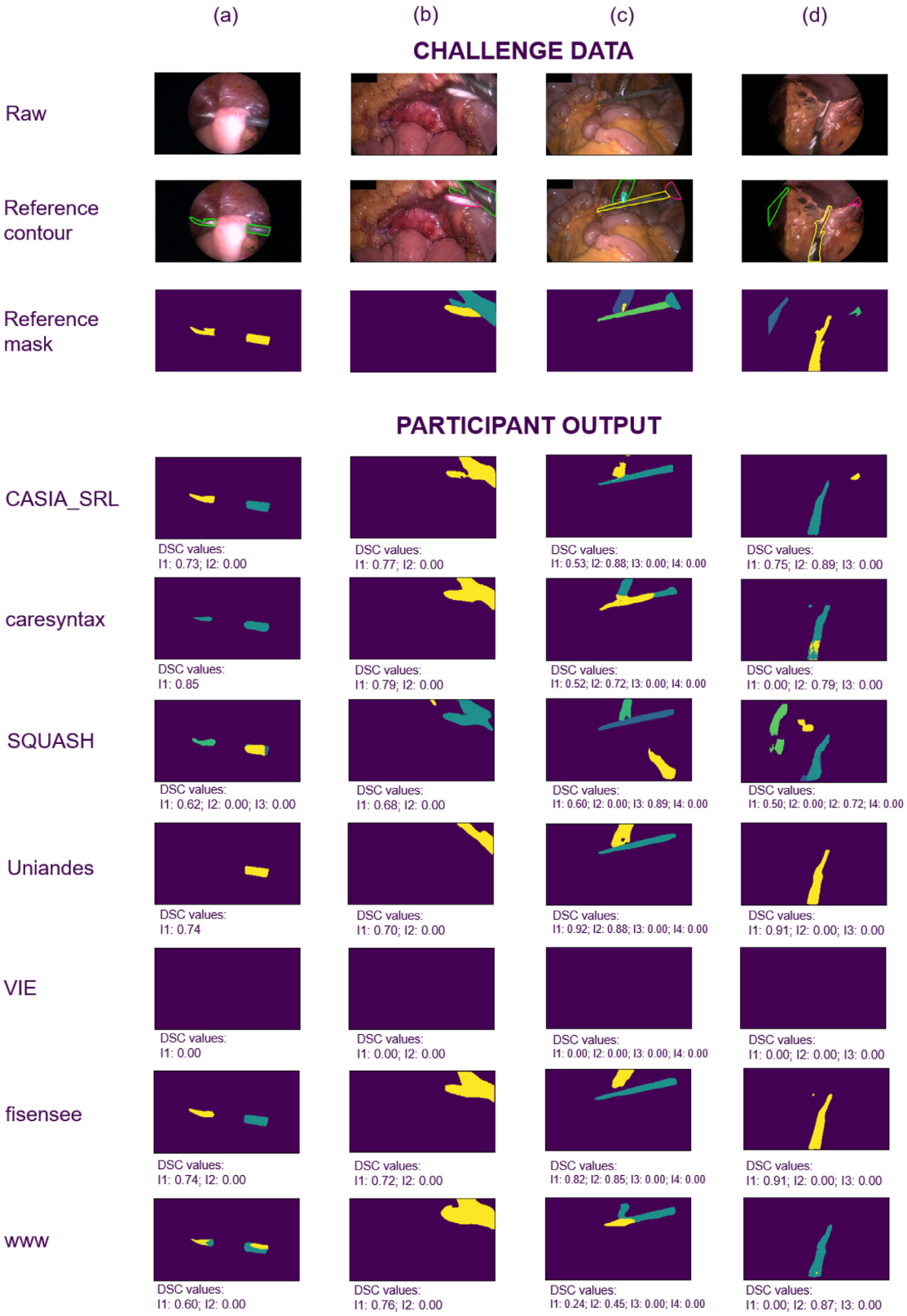


Fig. 10. Test cases with low corresponding algorithm performances. Each row shows the raw frame, the reference contours and mask as well as the algorithm output of the participating teams of the multi-instance segmentation (MIS) task for one representative frame. The columns represent image frames with (a) an instrument overlain by tissue, (b) motion, (c) multiple instruments, (d) underexposure and multiple instruments, respectively. For participants, the Dice Similarity Coefficient (DSC) values are provided per instrument instance (I_i).

fidence scores from the algorithms, we decided to update the ranking with the F1-score.

4.1.3. Challenge data

In general, we observed many inconsistencies in the initial data annotation, which is why we introduced a structured multi-stage annotation process involving medical experts and following a pre-defined annotation protocol (see Appendix B). We recommend challenge organizers to generate such a protocol from the outset of their challenge.

It should be noted that three different surgical procedures were used for the challenge, yet, these three procedures are all colorectal surgeries that share similarities. A rectal resection incorporates parts of a sigmoid resection, for example. It is possible that performance drops will be more radical when analyzing a wider variety of procedures such as biliopancreatic or upper gastrointestinal surgeries.

In the future, we will also prevent the potential side effects which resulting from pre-processing. The fact that we downsampled our video images may have harmed performance. However, due to the fact that (1) all participants had the same starting conditions, (2) the applied CNNs methods had to fit to GPUs and (3) all participants reduced the resolution further, we think that these effects are only minor.

4.2. Challenge outcome

4.2.1. Methods

The variability of all of the methods, submitted for the binary segmentation was vast and ranged from 2D U-Net versions (TernausNet, multi scale U-Net) to different implementations of the Mask R-CNN with a ResNet backbone to the latest DeepLabV3 network architecture. For the multi-instance detection and multi-instance segmentation tasks, however, the range of the underlying architecture was much narrower, with multiple Mask R-CNN variations and one combination of a U-Net, a classical approach and the principal component analysis (see Table 3).

The most successful participating team (*haoyun*) in the binary segmentation task implemented a DeepLabV3+ architecture which gave them the top rank in three out of the four rankings for the binary segmentation task. A relatively simple approach based on the combination of a U-Net with a connected component analysis by the *fisensee* team turned out to be a strong baseline and won accuracy rankings in both the binary segmentation task and the *DSC* accuracy ranking for the multi-instance segmentation task. It was, however, less successful in terms of robustness.

An increasingly relevant problem in reporting challenge results is the fact that it is often hard to understand which specific design choice for a certain algorithm make this algorithm better than the competing methods (Maier-Hein et al., 2018). Based on our challenge analysis, we hypothesize that data augmentation and the specifics of the training process are the key to a winning result. In other words, we believe that focusing on one architecture and performing a broad hyperparameter search in combination with an extensive data augmentation technique and a well-thought-out training procedure will create more benefit than testing many different network architectures without optimizing the training process. This is in line with recent findings in the field of radiological data science (Isensee et al., 2018).

4.2.2. Results

The key insights have already been summarized at the beginning of the discussion. Methods that tackle the multi-instance segmentation performed worse compared to the binary segmentation task. In fact, when multiple instrument instances were visible in one image, the algorithm performance decreased dramatically from over 0.8 for one instance to less than 0.6 for more than three instances (see Fig. 7). This is also reflected in Fig. 2 (c) and (d), which show clusters in the boxplots at specific metric values. These clusters correspond to the performance with respect to different numbers of instrument instances. For a single instrument, metric values are high, for multiple instruments the metric values are grouped around lower values. We thus conclude that detection of multiple instances remains an unsolved problem.

Although the described winning methods produce median *MI_DSC* results above 0.9 (see Fig. 6), most of them could not outperform the expert baseline in the multi-instance segmentation task, especially if more than one instrument was present in the image frames. In fact, only the teams *fisensee* (binary segmentation) and *Uniandes* (multi-instance segmentation) produced similar performances to the human annotator in stage 2 of the challenge. It should be noted that for pragmatic reasons, the additional labeling was performed only on a subset of images and with only one additional medical expert. The discrepancy in performance between algorithms and experts may differ based on the data and the annotator.

Generally, the expert accuracy is independent of the number of visible instances, while the performance of the algorithm drops with an increasing number. However, to our surprise, the expert also achieved comparatively low values in the robustness rankings (aggregated values of 0.43 or 0.47 for $n = 1$ and $n = 2$ instruments). We found this mainly to be caused by missing or wrong instrument instances (see Fig. E.1). However, where the expert did detect an instance, the segmentation quality of this instance is almost always good (*MI_DSC*= 0.9 is on the 10th percentile and *MI_DSC*= 0.95 on the 37th percentile), which is not the case for the algorithms as shown in see Fig. E.1 (Team *Uniandes* with *MI_DSC*=0.9 on the 14th percentile and *MI_DSC*= 0.95 on the 48th percentile; Team *fisensee* with *MI_DSC*= 0.9 on the 14th percentile and *MI_DSC*= 0.95 on the 37th percentile).

By analyzing the worst 100 cases across all of the methods, we found that all methods generally had issues with small, transparent or fast moving instruments. In addition, instruments close to other instruments or the image border, as well as partially hidden or crossing instruments were difficult to detect and segment (see Fig. 9 and 10). We also observed that classic challenges (Bodenstedt et al., 2018) such as reflections, blood, different illumination conditions did not pose any great problems. Images acquired when the lens of the endoscope was inside of a trocar were not particularly difficult to process.

It should be noted that only three of the ten methods incorporated the temporal video information provided with the frames to be annotated. One method used the video information to predict the likelihood of instrument presence in a multi-task setting while two approaches used the videos to calculate the optical flow. However, based on the team reports and on the challenge results, none of the teams where able taking a benefit from using the video data, neither for the binary segmentation task, nor for the multi-instance detection/segmentation tasks. Given the way in which medical and technical experts annotated the data, this is surprising, and we speculate that much of the potential of temporal context remains to be discovered.

Finally, it should be noted that an evaluation of the inference time of methods was not included in this paper because a respective metric had not been announced to the challenge participants. Although we assume that the participating teams had not optimized their methods for performance, we performed a preliminary analysis of the docker submissions to approximate computation times. This yielded runtimes between 0.07 and 7.3 seconds per image frame (mean: 1.09 seconds per image frame). Given the need for real-time inference, we recommend using a runtime-based metric in future challenges of this kind.

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 and conflicts of interest

This challenge has been funded by the Surgical Oncology Program of the National Center for Tumor Diseases (NCT) Heidelberg and the project "OP4.1", funded by the German Federal Ministry of Economic Affairs and Energy (grant number BMWI 01MT17001C). It was further supported by UNDERSTAND.AI[18], NVIDIA GmbH[19] and Digital Surgery[20]. The challenge was further supported by the Helmholtz Association under the joint research school HIDSS4Health (Helmholtz Information and Data Science School for Health). Furthermore, the authors wish to thank Tim Adler, Janek Gröhl, Alexander Seitel and Minu Dietlinde Tizabi for proofreading the paper.

L.M.-H., T.R., A.R., S.B. and S.S. worked with device manufacturer Karl Storz GmbH & Co. KG in the joint research project "OP4.1", funded by the German Federal Ministry of Economic Affairs and Energy (grant number BMWI 01MT17001C). M.W., H.G.K. and P.P.M. worked with device manufacturer Karl Storz GmbH & Co. KG in the joint research project "InnOPlan", funded by the German Federal Ministry of Economic Affairs and Energy (grant number BMWI 01MD15002E).

G.-B.B., Z.-G.H., Z.-L.N., Y.-J.Z. and H.-B.C. were supported by the National Key Research and Development Program of China (Grant 2017YFB1302704).

D.G., G.W. and L.W. were funded by National Natural Science Fundation of China (81771921, 61901084).

P.H., M.A.R. and D.J. were funded by Research Council of Norway projects number 263,248 (Privaton).

S.K. and K.S. were funded by the FWF Austrian Science Fund under grant P 32010-N38.

All challenge organizers and some members of their institute had access to training and test cases and were therefore not eligible for awards.

Appendix A. Challenge organization

The "Robust Medical Instrument Segmentation Challenge 2019 (ROBUST-MIS 2019)" was organized as a sub-challenge of the Endoscopic Vision Challenge 2019 at the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) in Shenzhen, China. It was organized by T. Roß, A. Reinke, M. Wagner, H. Kenngott, B. Müller, A. Kopp-Schneider and L. Maier-Hein. See Section A.1 for detailed description. The challenge was intended as a one-time event with a fixed submission deadline. The platforms grand-challenge.org (Roß et al., 2019a) and synapse.org (Roß et al., 2019b) served as websites for the challenge. Synapse served as data providing platform which was further used to upload the challenge participants' submissions.

The participation policies for the challenge allowed only fully automatic algorithms to be submitted. Although it was possible to use publicly available data released outside the field of medicine to train the methods or to tune hyperparameters, it was forbidden to use any medical data, besides the training data offered by the challenge. For members of the organizers' departments it was possible to participate in the challenge but they were not eligible for awards and their participation would have been highlighted in the leaderboards. The challenge was funded by the company Digital Surgery with a total monetary award of 10,000€. As the challenge comprised 9 rankings in total (see Section 2.3.2), each winning team was awarded 1,000€ and each runner-up team 125€. Moreover, the top three performing methods for each ranking were announced publicly. The remaining teams could decide whether or not their identity was revealed. One team decided not to be mentioned in the rankings. Finally, for this publication, each participating team could nominate members of their team as co-authors. The method description submitted by the authors was used in the publication (see Section 3.1). Personal data of the authors include their names, affiliations and contact addresses. References used in the method description were published as well. Participating teams are allowed to publish their results separately with explicit permission from the challenge organizers once this paper has been accepted for publication.

The submission instructions for the participating methods are published on the Synapse website and consist of a detailed description of the submission of docker containers which were used to evaluate the results. The complete submission instructions are provided in Appendix C. Algorithms were only evaluated on the test data set, so no leaderboard was published before the final result submission. The initial training data set was released on 1st July 2019, the final training data set on 5th August 2019. Participants could register for the challenge until 14th September 2019. The docker submission took place between 15th September and 28th September 2019. There where two deadlines, the 21th September for participants, whose methods would require more than 3h of runtime and the 28th September for participants, whose dockers needs less than 3h runtime. Participating teams had to submit a method description in addition to the docker containers.

The data sets of the challenge were fully anonymized (see Section 2.2) and could therefore be used without any ethics approval (Recital26, 2016). By registering in the challenge, each team agreed (1) to use the data provided only in the scope of the challenge and (2) to neither pass it on to a third party nor use it for any publication or for commercial use. The data will be made publicly available for non-commercial use.

The evaluation code for the challenge was made publicly available (Roß and Reinke, 2019) and participants were encouraged to release their methods in open source.

A1. Author contributions

All authors read the paper and agreed to publish it.

- T. Roß and A. Reinke organized the challenge, performed the evaluation and statistical analyses and wrote the manuscript
- P.M. Full, H. Hempe, D. Mindroc-Filimon, P. Scholz, T.N. Tran and P. Bruno reviewed and labeled the challenge data set
- M. Wagner, H. Kenngott, B.P. Müller-Stich organized the challenge and performed the medical expert review of the challenge data set
- M. Apitz performed the medical expert review of the challenge data set
- K. Kirtac, J. Lindström Bolmgrem, M. Stenzel, I. Twick and E. Hosgor participated in the challenge as team *caresyntax* in all three tasks
- Z.-L. Ni, H.-B. Chen, Y.-J. Zhou, G.-B. Bian and Z.-G. Hou participated in the challenge as team *CASIA_SRL* in the binary and multi-instance segmentation tasks
- D. Jha, M.A. Riegler and P. Halvorsen participated in the challenge as team *Djh* in the binary segmentation task
- F. Isensee and K. Maier-Hein participated in the challenge as team *fisensee* in all three tasks

[18] https://understand.ai.
[19] https://www.nvidia.com.
[20] https://digitalsurgery.co.

- L. Wang, D. Guo and G. Wang participated in the challenge as team *haoyun* in the binary segmentation task
- S. Leger, S. Bodenstedt and S. Speidel participated in the challenge as team *NCT* in the binary segmentation task
- S. Kletz and K. Schoeffmann participated in the challenge as team *SQUASH* in all three tasks
- L. Bravo, C. González and P. Arbeláez participated in the challenge as team *Uniandes* in all three tasks
- R. Shi, Z. Li, T. Jiang participated in the challenge as team *VIE* in all three tasks
- J. Wang, Y. Zhang, Y. Jin, L. Zhu, L. Wang and P.-A. Heng participated in the challenge as team *www* in all three tasks
- A. Kopp-Schneider and M. Wiesenfarth performed statistical analyses
- L. Maier-Hein organized the challenge, wrote the manuscript and supervised the project

Appendix B. Annotation instructions

B1. Terminology

- Matter: Anything that has mass, takes up space and can be clearly identified.
- Examples: tissue, surgical tools, blood
- Counterexamples: reflections, digital overlays, movement artifacts, smoke

Medical instrument to be detected and segmented: Elongated rigid object introduced into the patient and manipulated directly from outside the patient.

- Examples: grasper, scalpel, (transparent) trocar, clip applicator, hooks, stapling device, suction
- Counterexamples: non-rigid tubes, bandage, compress, needle (not directly manipulated from outside but manipulated with an instrument), coagulation sponges, metal clips

B2. Tasks

Participating teams may enter competitions related to the following tasks:

Binary segmentation:

- Input: 250 consecutive frames (10sec) of a laparoscopic video with the last frame containing at least one medical instrument.
- Output: A binary image, in which "0" indicates the absence of a medical instrument and a number ">0" represents the presence of a medical instrument.

Multi-instance detection and segmentation:

- Input: 250 consecutive frames (10sec) of a laparoscopic video with the last frame containing at least one medical instrument.
- Output: An image, in which "0" indicates the absence of a medical instrument and numbers "1", "2",... represent different instances of medical instruments.

For all three tasks, the entire corresponding video of the surgery is provided along with the training data as context information. In the test phase, only the test image along with the preceding 250 frames is provided. See Supplementary file S1.

Appendix C. Submission instructions

The following section provides the instruction document that challenge participants obtained. See Supplementary file S3.

Appendix D. Rankings for all stages

The ranking schemes described in Section 2.3.2 were also computed for stages 1 and 2. To compare the performance of participating teams across stages, stacked frequency plots of the observed ranks, separated by the algorithms, for each ranking of the binary and multi-instance segmentation tasks are displayed in Fig. D.1 to D.8. Observed ranks across bootstrap samples are presented over the three stages the stages. The metric values for the multi-instance detection task are displayed in Table D.1

Table D.1
Results over all stages for the multi-instance detection task.

Team identifier	mAP		
	Stage 1	Stage 2	Stage 3
Uniandes	1.000	0.833	1.000
VIE	0.750	0.778	0.978
caresyntax	0.944	0.833	0.972
SQUASH	0.967	1.000	0.966
fisensee	1.000	1.000	0.964
www	0.900	0.833	0.944

Table D.2
Results over all stages for the multi-instance detection task as reported during the challenge event. Those values have to be interpreted with care due to an implementation error in the validation.

Team identifier	F1-score		
	Stage 1	Stage 2	Stage 3
Uniandes	0.94	0.93	0.91
www	0.92	0.90	0.90
caresyntax	0.92	0.91	0.89
SQUASH	0.90	0.86	0.86
fisensee	0.89	0.89	0.86
VIE	0.84	0.82	0.82

.

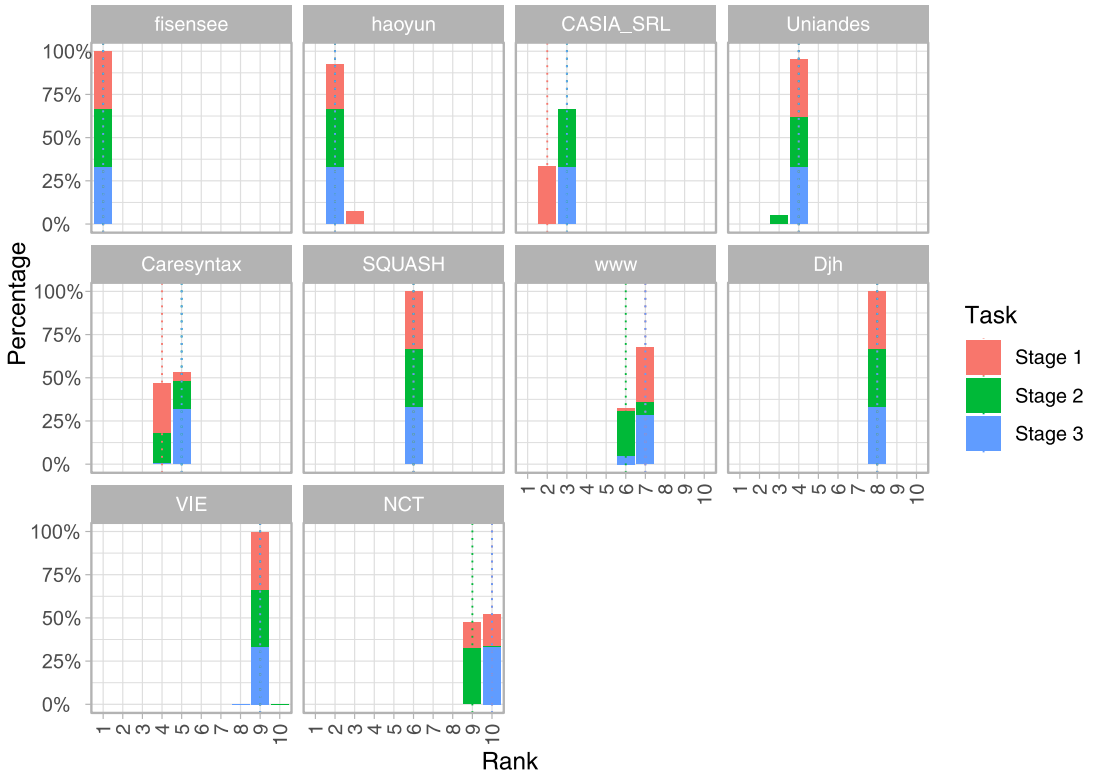


Fig. D.1. Stacked frequency plot for stages 1 to 3 with the Dice Similarity Coefficient (*DSC*) accuracy ranking of the binary segmentation task. The plots were generated using the package challengeR (Wiesenfarth et al., 2019b; 2019a).

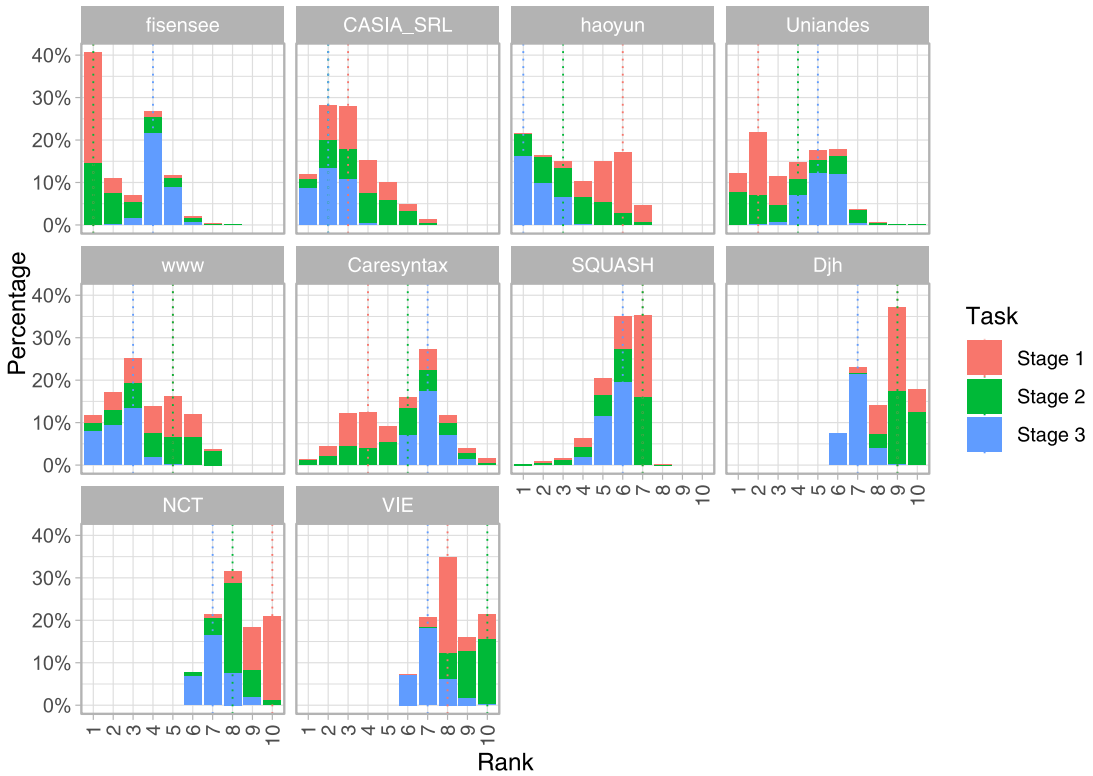


Fig. D.2. Stacked frequency plot for stages 1 to 3 with the Dice Similarity Coefficient (*DSC*) robustness ranking of the binary segmentation task. The plots were generated using the package challengeR (Wiesenfarth et al., 2019b; 2019a).

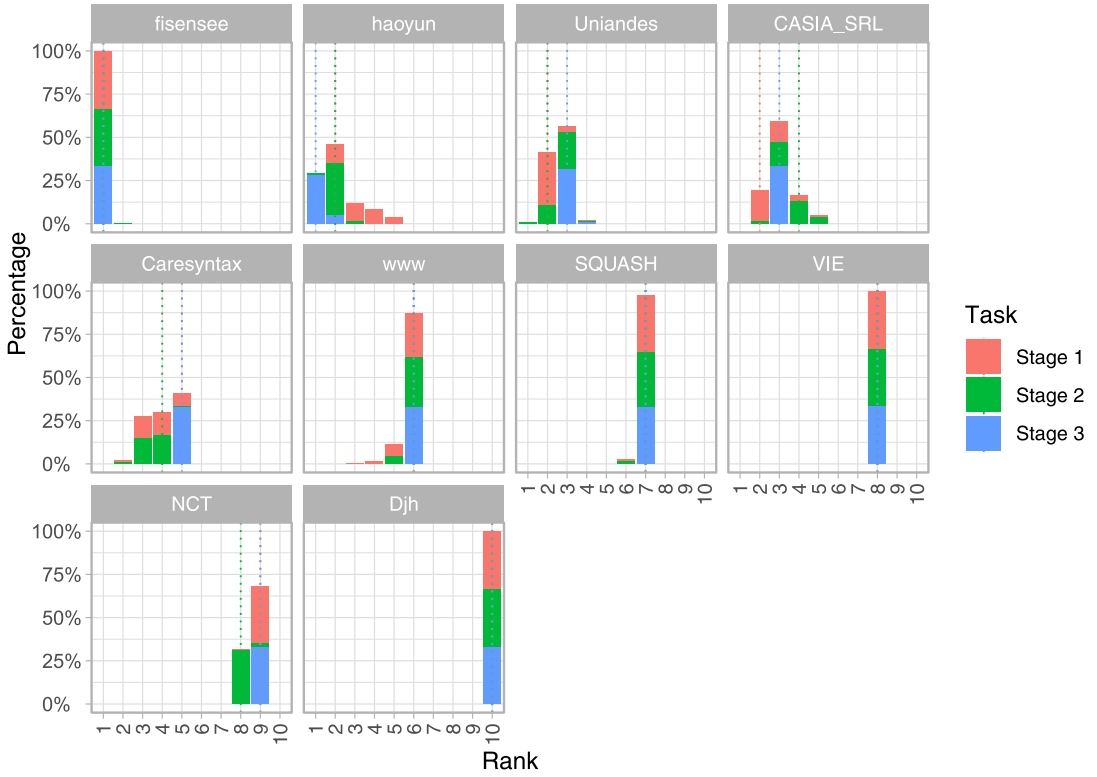


Fig. D.3. Stacked frequency plot for stages 1 to 3 with the Normalized Surface Distance (*NSD*) accuracy ranking of the binary segmentation task. The plots were generated using the package challengeR (Wiesenfarth et al., 2019b; 2019a).

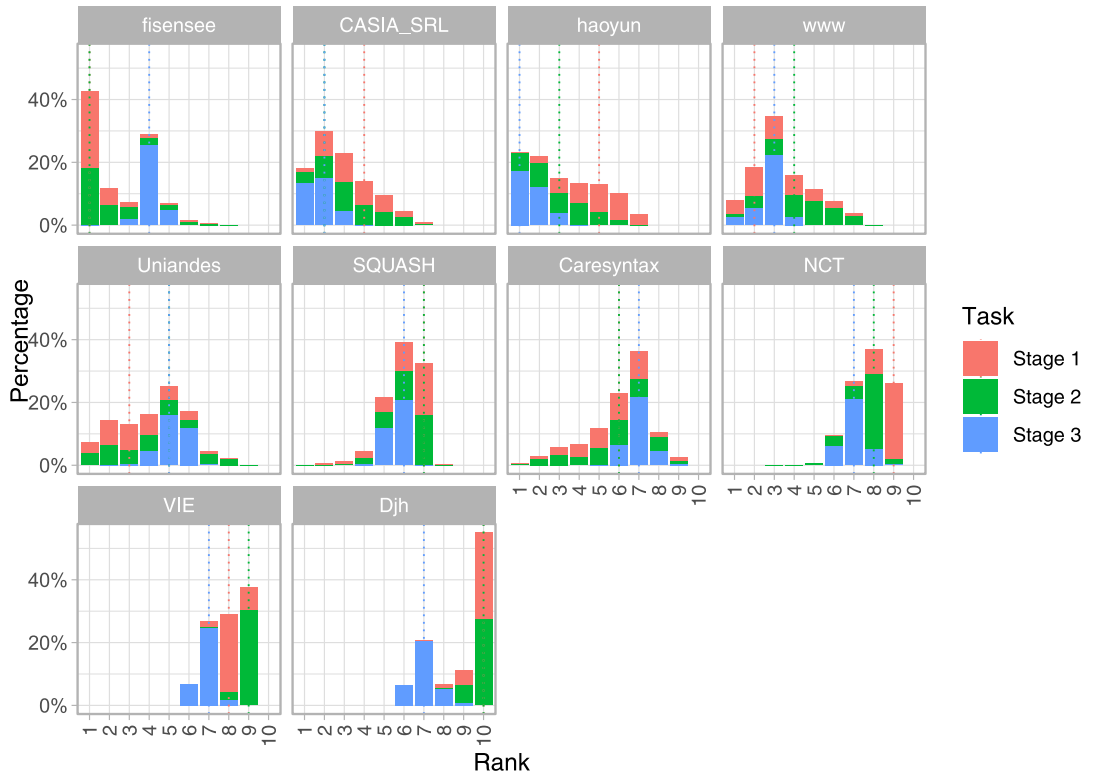


Fig. D.4. Stacked frequency plot for stages 1 to 3 with the Normalized Surface Distance (*NSD*) robustness ranking of the binary segmentation task.

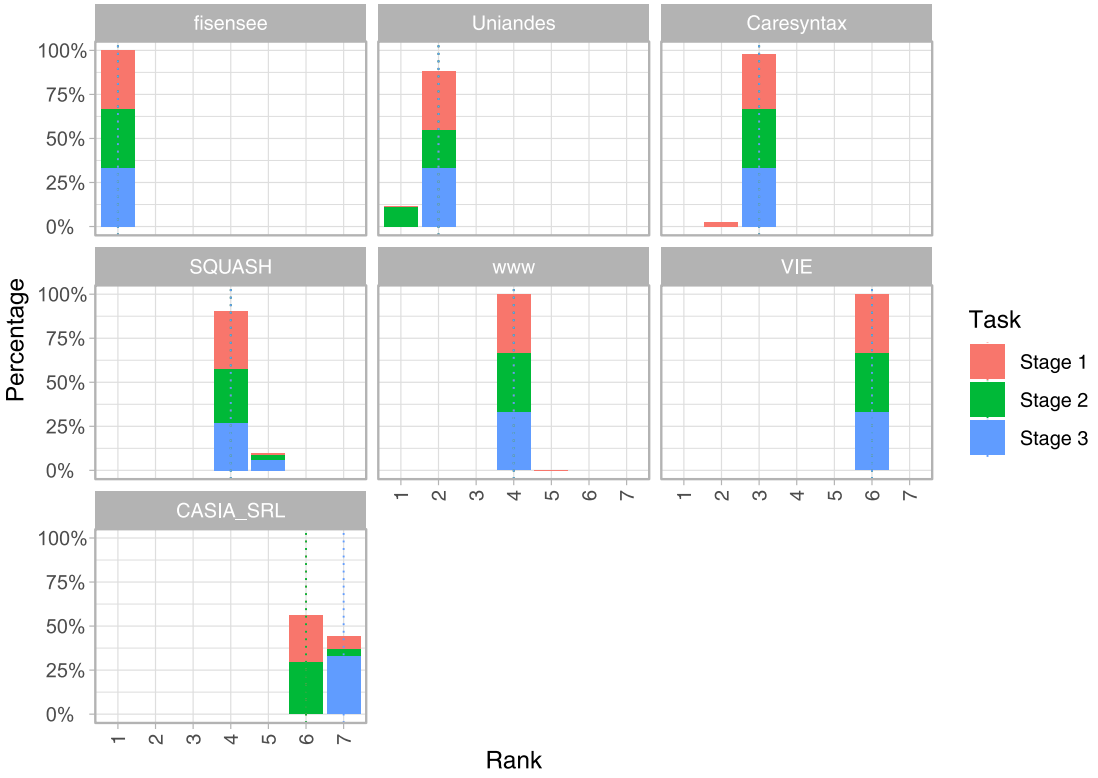


Fig. D.5. Stacked frequency plot for stages 1 to 3 with (multi-instance) Dice Similarity Coefficient (*(MI_)DSC*) accuracy ranking of the multi-instance segmentation task. The plots were generated using the package challengeR (Wiesenfarth et al., 2019b; 2019a).

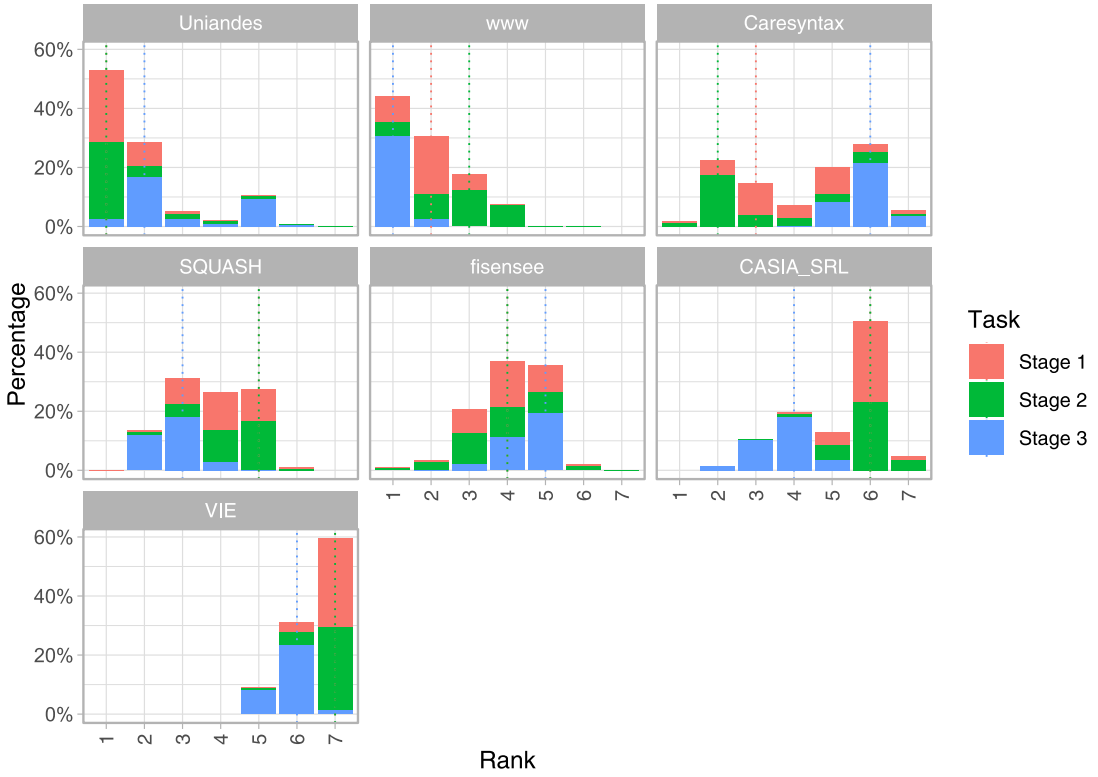


Fig. D.6. Stacked frequency plot for stages 1 to 3 with the (multi-instance) Dice Similarity Coefficient (*(MI_)DSC*) robustness ranking of the multi-instance segmentation task. The plots were generated using the package challengeR (Wiesenfarth et al., 2019b; 2019a).

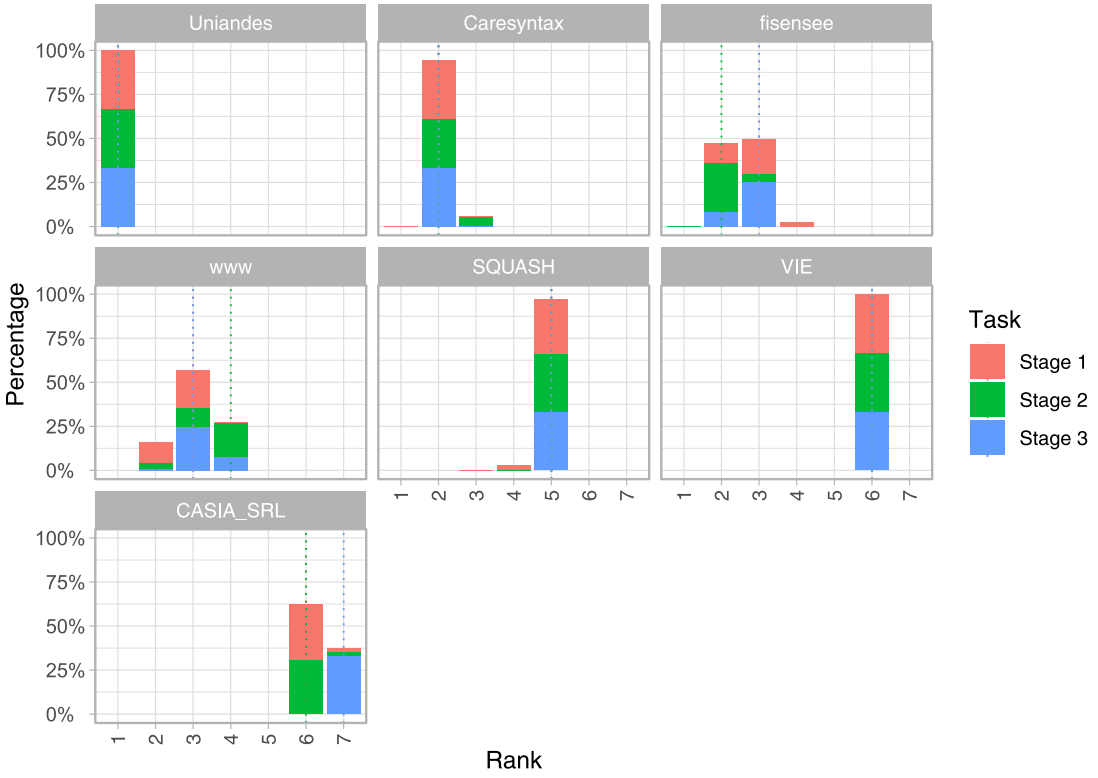


Fig. D.7. Stacked frequency plot for stages 1 to 3 with the (multi-instance) Normalized Surface Distance (*(MI_)NSD*) accuracy ranking of the multi-instance segmentation task. The plots were generated using the package challengeR (Wiesenfarth et al., 2019b; 2019a).

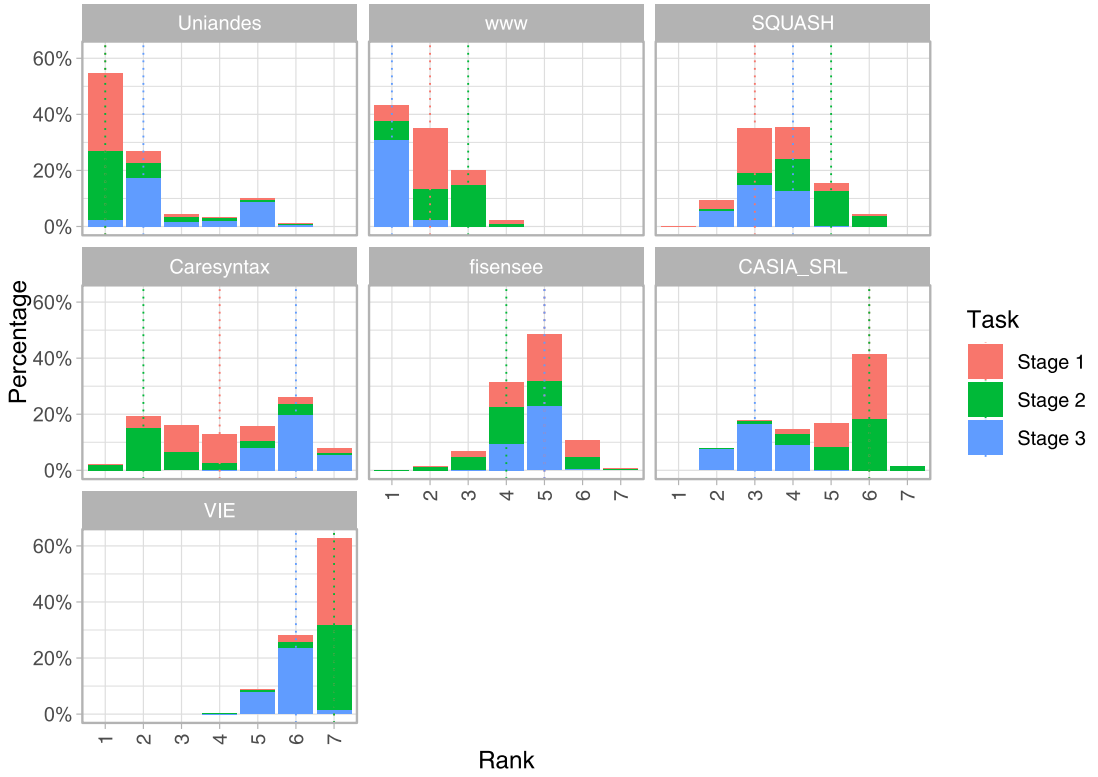


Fig. D.8. Stacked frequency plot for stages 1 to 3 with the (multi-instance) Normalized Surface Distance (*(MI_)NSD*) robustness ranking of the multi-instance segmentation task. The plots were generated using the package challengeR (Wiesenfarth et al., 2019b; 2019a).

Appendix E. Results for stage 2 including expert baseline

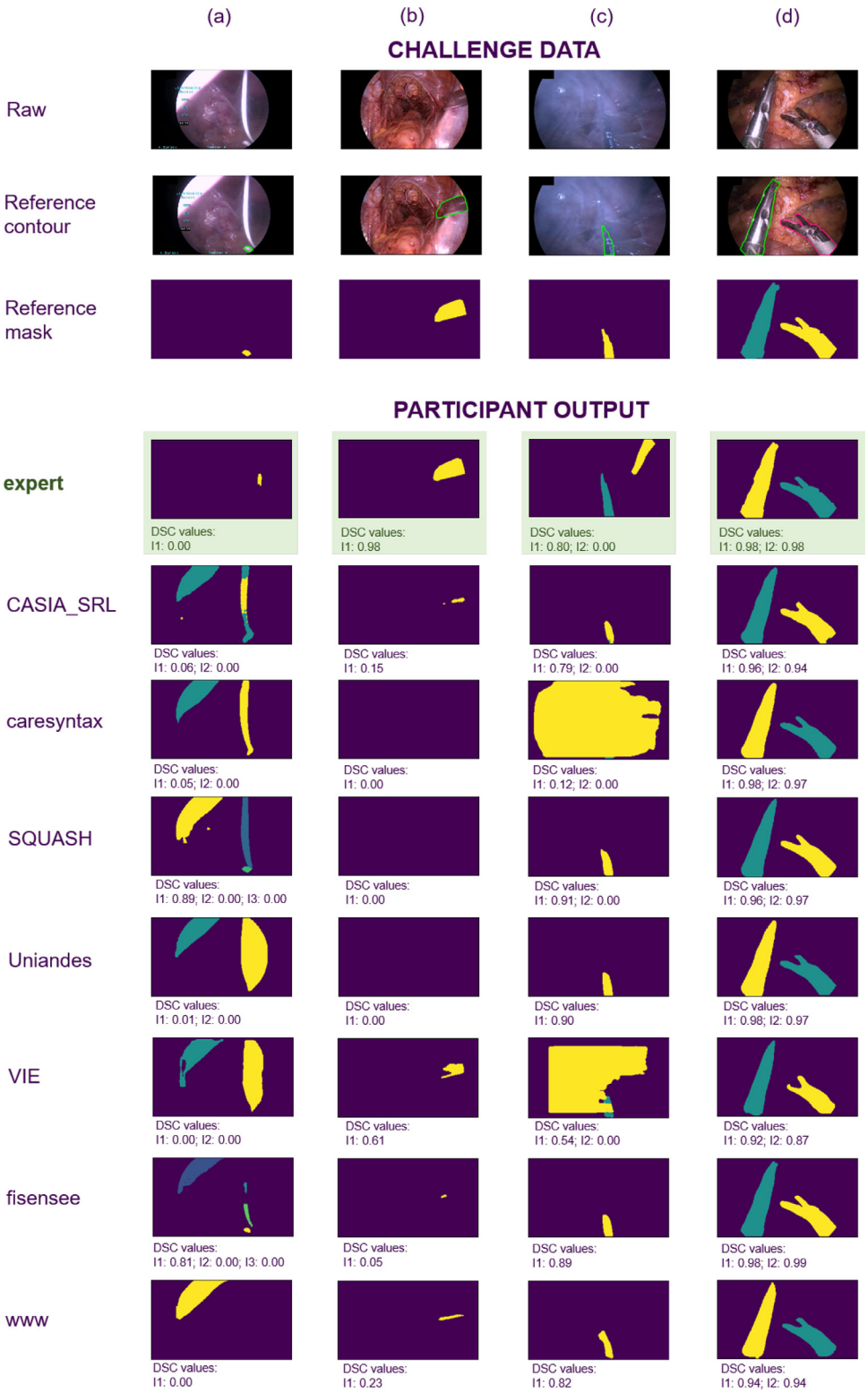


Fig. E.1. Example frames from stage 2 with corresponding participant and expert performances. Each row shows the raw frame, the reference contours and mask as well as the (algorithm) output of the participating teams/expert of the multi-instance segmentation (MIS) task for one representative frame. The columns represent image frames with (a) low expert, low algorithm performances, (b) high expert, low algorithm performances, (c) low expert, high performances, (d) high expert, high algorithm performances, respectively. For participants, the Dice Similarity Coefficient (DSC) values are provided per instrument instance (I_i).

Appendix F. Challenge design document

See Supplementary file S2..

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2020.101920.

References

Allan, Max, Kondo, Satoshi, Bodenstedt, Sebastian, Leger, Stefan, Kadkhodamohammadi, Rahim, Luengo, Imanol, Fuentes, Felix, Flouty, Evangello, Mohammed, Ahmed, Pedersen, Marius, et al., 2020. 2018 Robotic Scene Segmentation Challenge. arXiv:2001.11190.

Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.-H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., et al., 2019. 2017 Robotic instrument segmentation challenge. arXiv:1902.06426.

Amini Khoiy, K., Mirbagheri, A., Farahmand, F., 2016. Automatic tracking of laparoscopic instruments for autonomous control of a cameraman robot. Minimally Invasive Therapy & Allied Technologies 25 (3), 121–128.

Armstrong, T.G., Moffat, A., Webber, W., Zobel, J., 2009. Improvements that don't add up: ad-hoc retrieval results since 1998. In: Proceedings of the 18th ACM conference on Information and knowledge management, pp. 601–610.

Bianchi, F., Masaracchia, A., Shojaei Barjuei, E., Menciassi, A., Arezzo, A., Koulaouzidis, A., Stoyanov, D., Dario, P., Ciuti, G., 2019. Localization strategies for robotic endoscopic capsules: a review. Expert Rev. Med. Devices 16 (5), 381–403.

Bodenstedt, S., Allan, M., Agustinos, A., Du, X., Garcia-Peraza-Herrera, L., Kenngott, H., Kurmann, T., Müller-Stich, B., Ourselin, S., Pakhomov, D., et al., 2018. Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. arXiv:1805.02475.

Burström, G., Nachabe, R., Persson, O., Edström, E., Terander, A.E., 2019. Augmented and virtual reality instrument tracking for minimally invasive spine surgery: a feasibility and accuracy study. Spine 44 (15), 1097–1104.

Cardoso, M.J., 2018. Medical segmentation decathlon. http://medicaldecathlon.com/. Accessed: 2019-10-29.

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

Clevert, D.-A., Unterthiner, T., Hochreiter, S., 2015. Fast and accurate deep network learning by exponential linear units (elus). arXiv:1511.07289.

De Paolis, L.T., De Luca, V., 2019. Augmented visualization with depth perception cues to improve the surgeon's performance in minimally invasive surgery. Medical & biological engineering & computing 57 (5), 995–1013.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., Dean, J., 2019. A guide to deep learning in healthcare. Nat. Med. 25 (1), 24–29.

Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2015. The pascal visual object classes challenge: aretrospective. Int. J. Comput. Vis. 111 (1), 98–136.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. 88 (2), 303–338.

Fawaz, H.I., Forestier, G., Weber, J., Idoumghar, L., Muller, P.-A., 2019. Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. Int. J. Comput. Assist. Radiol. Surg. 14 (9), 1611–1617.

García-Peraza-Herrera, L.C., Li, W., Gruijthuijsen, C., Devreker, A., Attilakos, G., Deprest, J., Vander Poorten, E., Stoyanov, D., Vercauteren, T., Ourselin, S., 2016. Real-time segmentation of non-rigid surgical tools based on deep learning and tracking. In: International Workshop on Computer-Assisted and Robotic Endoscopy. Springer, pp. 84–95.

Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., Zhang, T., Gao, S., Liu, J., 2019. Ce-net: context encoder network for 2d medical image segmentation. IEEE Trans. Med. Imaging 38 (10), 2281–2292.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp. 2961–2969.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Iglovikov, V., Shvets, A., 2018. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. arXiv:1801.05746.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T., 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2462–2470.

Ioannidis, J.P., 2005. Why most published research findings are false. PLos med 2 (8), e124.

Isensee, F., Maier-Hein, K.H., 2020. Or-unet: an optimized robust residual u-net for instrument segmentation in endoscopic images. arXiv:2004.12668.

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al., 2018. Nnu-net: self-adapting framework for u-net-based medical image segmentation. arXiv:1809.10486.

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

Kiefer, J., Wolfowitz, J., et al., 1952. Stochastic estimation of the maximum of a regression function. The Annals of Mathematical Statistics 23 (3), 462–466.

Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv:1412.6980.

Kuhn, H.W., 1955. The hungarian method for the assignment problem. Naval research logistics quarterly 2 (1–2), 83–97.

Kurmann, Thomas, Neila, Pablo Marquez, Du, Xiaofei, Fua, Pascal, Stoyanov, Danail, Wolf, Sebastian, Sznitman, Raphael, 2017. Simultaneous recognition and pose estimation of instruments in minimally invasive surgery. International Conference on Medical Image Computing and Computer-Assisted Intervention 505–513.

Laina, I., Rieke, N., Rupprecht, C., Vizcaíno, J.P., Eslami, A., Tombari, F., Navab, N., 2017. Concurrent segmentation and localization for tracking of surgical instruments. In: International conference on medical image computing and computer-assisted intervention. Springer, pp. 664–672.

Law, H., Ghani, K., Deng, J., 2017. Surgeon technical skill assessment using computer vision based analysis. In: Machine learning for healthcare conference, pp. 88–99.

Lin, S., Qin, F., Bly, R.A., Moe, K.S., Hannaford, B., 2019. Automatic sinus surgery skill assessment based on instrument segmentation and tracking in endoscopic video. In: International Workshop on Multiscale Multimodal Medical Imaging. Springer, pp. 93–100.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: European conference on computer vision. Springer, pp. 740–755.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A.P., Carass, A., et al., 2018. Why rankings of biomedical image analysis competitions should be interpreted with care. Nat. Commun. 9 (1), 5217.

Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A.L., Arbel, T., Eisenmann, M., Hanbuary, A., Jannin, P., Müller, H., Onogur, S., et al., 2019. Bias: transparent reporting of biomedical image analysis challenges. arXiv:1910.04071.

Maier-Hein, L., Vedula, S.S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., et al., 2017. Surgical data science for next-generation interventions. Nat. Biomed. Eng. 1 (9), 691–696.

Nguyen, X.A., Ljuhar, D., Pacilli, M., Nataraja, R.M., Chauhan, S., 2019. Surgical skill levels: classification and analysis using deep neural network model and motion signals. Comput. Methods Programs Biomed. 177, 1–8.

Ni, Z.-L., Bian, G.-B., Wang, G.-A., Zhou, X.-H., Hou, Z.-G., Xie, X.-L., Li, Z., Wang, Y.-H., 2020. Barnet: bilinear attention network with adaptive receptive field for surgical instrument segmentation. arXiv:2001.07093.

Ni, Z.-L., Bian, G.-B., Xie, X.-L., Hou, Z.-G., Zhou, X.-H., Zhou, Y.-J., 2019. Rasnet: Segmentation for tracking surgical instruments in surgical videos using refined attention segmentation network. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, pp. 5735–5738.

Nikolov, S., Blackwell, S., Mendes, R., De Fauw, J., Meyer, C., Hughes, C., Askham, H., Romera-Paredes, B., Karthikesalingam, A., Chu, C., et al., 2018. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. arXiv:1809.04430.

Pakhomov, Daniil, Premachandran, Vittal, Allan, Max, Azizian, Mahdi, Navab, Nassir, 2019. Deep residual learning for instrument segmentation in robotic surgery. International Workshop on Machine Learning in Medical Imaging 566–573.

Panch, T., Mattie, H., Celi, L.A., 2019. The 'inconvenient truth' about ai in healthcare. Npj Digital Medicine 2 (1), 1–3.

Qayyum, A., Qadir, J., Bilal, M., Al-Fuqaha, A., 2020. Secure and robust machine learning for healthcare: asurvey. arXiv:2001.08103.

Recital26, 2016. General data protection regulation of the european union. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679_d1e1374-1-1. Accessed: 2019-10-29.

Reinke, A., Eisenmann, M., Onogur, S., Stankovic, M., Scholz, P., Full, P.M., Bogunovic, H., Landman, B.A., Maier, O., Menze, B., et al., 2018. How to exploit weaknesses in biomedical challenge design and organization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 388–395.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234–241.

Roß, T., Reinke, A., 2019. Robustmis2019. https://phabricator.mitk.org/source/rmis2019/. Accessed: 2019-10-29.

Roß, T., Reinke, A., Maier-Hein, L., 2019a. Robust medical instrument segmentation (ROBUST-MIS) challenge (grand-challenge.org). https://robustmis2019.grand-challenge.org/. Accessed: 2019-10-29.

Roß, T., Reinke, A., Maier-Hein, L., 2019b. Robust medical instrument segmentation (ROBUST-MIS) challenge (synapse.org). https://www.synapse.org/_!Synapse:syn18779624/wiki/. Accessed: 2019-10-29.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252.

Shankar, V., Roelofs, R., Mania, H., Fang, A., Recht, B., Schmidt, L., 2020. Evaluating machine accuracy on imagenet. In: International Conference on Machine Learning (ICML).
Shapiro, L.G., 1996. Connected Component Labeling and Adjacency Graph Construction. In: Machine Intelligence and Pattern Recognition, 19. Elsevier, pp. 1–30.
Siddaiah-Subramanya, M., Tiang, K.W., Nyandowe, M., 2017. A new era of minimally invasive surgery: progress and development of major technical innovations in general surgery over the last decade. The Surgery Journal 3 (04), e163–e166.
Su, Y.-H., Huang, K., Hannaford, B., 2018. Real-time vision-based surgical tool segmentation with robot kinematics prior. In: 2018 International Symposium on Medical Robotics (ISMR). IEEE, pp. 1–6.
Wang, R., Zhang, M., Meng, X., Geng, Z., Wang, F.-Y., 2017. 3-D tracking for augmented reality using combined region and dense cues in endoscopic surgery. IEEE J. Biomed. Health Inform. 22 (5), 1540–1551.
Wiesenfarth, M., Reinke, A., Landman, B.A., Cardoso, M.J., Maier-Hein, L., Kopp-Schneider, A., 2019a. challengeR: Methods and open-source toolkit for analyzing and visualizing challenge results. https://github.com/wiesenfa/challengeR. Accessed: 2019-10-29.
Wiesenfarth, M., Reinke, A., Landman, B.A., Cardoso, M.J., Maier-Hein, L., Kopp-Schneider, A., 2019. Methods and open-source toolkit for analyzing and visualizing challenge results. arXiv:1910.05121.
Zhang, J., Gao, X., 2020. Object extraction via deep learning-based marker-free tracking framework of surgical instruments for laparoscope-holder robots. Int. J. Comput. Assist. Radiol. Surg. 15 (8), 1335–1345.
Zhao, Z., Chen, Z., Voros, S., Cheng, X., 2019. Real-time tracking of surgical instruments based on spatio-temporal context and deep learning. Computer Assisted Surgery 24 (sup1), 20–29.
Zhou, S.K., Rueckert, D., Fichtinger, G., 2019. Handbook of medical image computing and computer assisted intervention. Academic Press.

A.21 Paper XXI : Progressively Normalized Self - Attention Network for Video Polyp Segmentation

Authors: Ge-Peng Ji, Yu-Cheng Chou, Deng-Ping Fan, Geng Chen, Huazhu Fu, Debesh Jha, and Ling Shao

Abstract: Existing video polyp segmentation (VPS) models typically employ convolutional neural networks (CNNs) to extract features. However, due to their limited receptive fields, CNNs cannot fully exploit the global temporal and spatial information in successive video frames, resulting in false positive segmentation results. In this paper, we propose the novel PNS-Net (Progressively Normalized Self-attention Net- work), which can efficiently learn representations from polyp videos with real-time speed ($\sim$ 140fps) on a single RTX 2080 GPU and no post-processing. Our PNS-Net is based solely on a basic normalized self-attention block, equipping with recurrence and CNNs entirely. Experiments on challenging VPS datasets demonstrate that the proposed PNS-Net achieves state-of-the-art performance. We also conduct extensive experiments to study the effectiveness of the channel split, soft-attention, and progressive learning strategy. We find that our PNS-Net works well under different settings, making it a promising solution to the VPS task.

Published: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2021.

Candidate contributions: D. Jha experimented and provided the results of the baseline experiment, ResUNet++ and participated in the revision of the manuscript.

Thesis objectives: Objective III

Progressively Normalized Self-Attention Network for Video Polyp Segmentation

Ge-Peng Ji[1,2], Yu-Cheng Chou[2], Deng-Ping Fan[1,✉]
Geng Chen[1], Huazhu Fu[1], Debesh Jha[3], and Ling Shao[1]

[1] Inception Institute of AI (IIAI) [2] Wuhan University [3] SimulaMet
dengpfan@gmail.com

Abstract. Existing video polyp segmentation (VPS) models typically employ convolutional neural networks (CNNs) to extract features. However, due to their limited receptive fields, CNNs cannot fully exploit the global temporal and spatial information in successive video frames, resulting in false positive segmentation results. In this paper, we propose the novel ***PNS-Net*** (Progressively Normalized Self-attention Network), which can efficiently learn representations from polyp videos with real-time speed (∼**140fps**) on a single RTX 2080 GPU and no post-processing. Our *PNS-Net* is based solely on a basic normalized self-attention block, equipping with recurrence and CNNs entirely. Experiments on challenging VPS datasets demonstrate that the proposed *PNS-Net* achieves state-of-the-art performance. We also conduct extensive experiments to study the effectiveness of the channel split, soft-attention, and progressive learning strategy. We find that our *PNS-Net* works well under different settings, making it a promising solution to the VPS task.

Keywords: Normalized self-attention · Polyp segmentation · Colonoscopy

1 Introduction

Early diagnosis of colorectal cancer (CRC) plays a vital role in improving the survival rate of CRC patients. In fact, the survival rate in the first stage of CRC is over 95%, decreasing to below 35% in the fourth and fifth stages [4]. Currently, colonoscopy is widely adopted in clinical practice and has become a standard method for screening CRC. During the colonoscopy, physicians visually inspect the bowel with an endoscope to identify polyps, which can develop into CRC if left untreated. In practice, colonoscopy is highly dependent on the physicians' level of experience and suffers from a high polyp miss rate [18]. These limitations can be resolved with automatic polyp segmentation techniques, which segment polyps from colonoscopy images/videos without intervention from physicians. However, accurate and real-time polyp segmentation is a challenging task due to the low boundary contrast between a polyp and its surroundings and the large shape variation of polyps [8].

G.-P. Ji and Y.-C. Chou contributed equally. Code: http://dpfan.net/pranet/

Significant efforts have been dedicated to overcoming these challenges. In early studies, learning-based methods turned to handcrafted features [16,20], such as color, shape, texture, appearance, or some combination. These methods train a classifier to separate the polyps from the background. However, they usually suffer from low accuracy due to the limited representation capability of handcrafted features in depicting heterogeneous polyps, as well as the close resemblance between polyps and hard mimics [24]. In more recent studies, deep learning methods have been used for polyp segmentation [24,26]. Although these methods have made some progress, they only use bounding boxes to detect polyps, and therefore cannot accurately locate the boundaries. To solve this, Brandao *et al.* [5] adopted a fully convolutional networks (FCN) with a pre-trained model to recognize and segment polyps. Later, Akbari *et al.* [1] introduced a modified FCN to increase the accuracy of polyp segmentation. Inspired by the success of UNet [19] in biomedical image segmentation, UNet++ [28] and ResUNet [13] are employed for polyp segmentation and achieved good results. Some methods also focus on area-boundary constraints. For instance, PsiNet [17] makes use of polyp boundary and area information simultaneously. Fang *et al.* [9] introduced a three-step selective feature aggregation network. ACSNet [25] utilized an adaptive context selection based encoder-decoder framework. Zhong *et al.* [27] propose a context-aware network based on adaptive scale and global semantic context. Introduced more recently, the current golden standard for image polyp segmentation, PraNet [8], applies area and boundary cues in a reverse attention module, achieving the cutting-edge performance. However, these methods have only been trained and evaluated on still images and focus on static information, ignoring the temporal information in endoscopic videos which can be exploited for better results. To this end, Puyal *et al.* [18] propose a hybrid 2D/3D CNN architecture. Their model aggregates spatial and temporal correlations and achieves better segmentation results. However, the spatial correlation between frames is restricted by the size of the kernel, preventing the accurate segmentation of fast videos.

Recently, the self-attention network [22] has shown superior performance in computer vision tasks such as video object segmentation [10], image super-resolution [23], and others. Inspired by this, in this paper, we propose a novel self-attention framework, called the **P**rogressively **N**ormalized **S**elf-attention **Net**work (***PNS-Net***), for the video polyp segmentation (VPS) task. Our contributions are as follows:

- Different from existing CNN-based models, the proposed *PNS-Net* framework is a self-attention model for VPS, introducing a new perspective for addressing this task.
- To fully utilize the temporal and spatial cues, we propose a simple normalized self-attention (NS) block. The NS block is flexible (backbone-free) and efficient, enabling it to easily be embedded into current CNN-based encoder-decoder architectures for better performance.
- We evaluate the proposed *PNS-Net* on challenging VPS datasets and compare it with two classical methods (*i.e.*, UNet [19] and UNet++ [28]) and

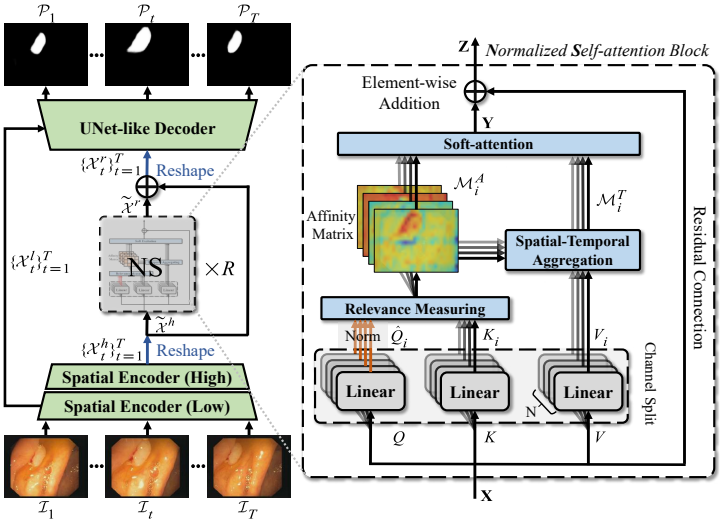


Fig. 1: Pipeline of the proposed *PNS-Net*, including the normalized self-attention block (see § 2.1) with a stacked (×R) learning strategy (see § 2.2).

three cutting-edge models (*i.e.*, ResUNet [13], ACSNet [25], and PraNet [8]). Experimental results show that *PNS-Net* achieves state-of-the-art performance with real-time speed. All the training data, models, results, and evaluation tools will be released to advance the development of this field.

2 Method

2.1 Normalized Self-attention (NS)

Motivation. Recently, the self-attention mechanism [22] has been widely exploited in many popular computer vision tasks. However, in our initial studies, we found that introducing the original self-attention mechanism to the VPS task does not achieve satisfactory results (*i.e.*, high accuracy and speed).

Analysis. For the VPS task, multi-scale polyps move at various speeds. Thus, dynamically updating the receptive field of the network is important. Further, the self-attention, such as the non-local network [22], incurs a high computational and memory cost, which limits the inference speed for our fast and dense prediction task. Motivated by the recent video salient object detection model [10], we utilize the **channel split**, **query-dependent**, and **normalization** rules to reduce the computational cost and improve the accuracy, respectively.

Channel Split Rule. Specifically, given an input feature (*i.e.*, $\mathbf{X} \in \mathbb{R}^{T\times H\times W\times C}$) extracted from T video frames with a size of $H \times W$ and C channels, we first utilize three linear embedding functions $\theta(\cdot)$, $\phi(\cdot)$, and $g(\cdot)$ to generate the corresponding attention features, which are implemented by a $1 \times 1 \times 1$ convolutional

layer [22]. This can be expressed as:

$$Q = \theta(\mathbf{X}), K = \phi(\mathbf{X}), V = g(\mathbf{X}). \tag{1}$$

Then we split each attention feature $\{Q, K, V\} \in \mathbb{R}^{T\times H\times W\times C}$ into N groups along the channel dimension and generate query, key, and value features, *i.e.*, $\{Q_i, K_i, V_i\} \in \mathbb{R}^{T\times H\times W\times \frac{C}{N}}$, where $i = \{1, 2, \cdots, N\}$.

Query-Dependent Rule. To extract the spatial-temporal relationship between successive video frames, we need to measure the similarity between query features Q_i and key features K_i. Inspired by [10], we introduce N relevance measuring (*i.e.*, query-dependent rule) blocks to compute the spatial-temporal affinity matrix for the *constrained neighborhood* of the target pixel. Rather than computing the response between a query position and the feature at all positions, as done in [22], the relevance measuring block can capture more relevance regarding the target object within T frames. More specifically, given a sliding window with fixed kernel size k and dilation rate $d_i = 2i - 1$, we get the corresponding constrained neighborhood in K_i for query pixel $\mathbf{X}^q$ of Q_i in position (x, y, z), which can be obtained by a sampling function $\mathcal{F}^S$. This is computed by:

$$\mathcal{F}^S\langle \mathbf{X}^q, K_i\rangle \in \mathbb{R}^{T(2k+1)^2\times \frac{C}{N}} = \Sigma_{m=x-kd_i}^{x+kd_i}\Sigma_{n=y-kd_i}^{y+kd_i}\Sigma_{t=1}^{T} K_i(m, n, t), \tag{2}$$

where $1 \le x \le H$, $1 \le y \le W$, and $1 \le z \le T$. Thus, the size of the constrained neighborhood depends on the various spatial-temporal receptive fields with different kernel size k, dilation rate d_i, and frame number T, respectively.

Normalization Rule. However, the internal covariate shift problem [11] exists in the feed-forward of input Q_i, incurring that the layer parameters cannot dynamically adapt the next mini-batch. Therefore, we maintain a fixed distribution for Q_i via:

$$\hat{Q}_i = \texttt{Norm}(Q_i), \tag{3}$$

where `Norm` is implemented by layer normalization [2] *along temporal dimension.*

Relevance Measuring. Finally, the affinity matrix is computed as:

$$\mathcal{M}_i^A \in \mathbb{R}^{THW\times T(2k+1)^2} = \texttt{Softmax}\left(\frac{\hat{Q}_i\mathcal{F}^S\langle \hat{\mathbf{X}}^q, K_i\rangle^{\mathbf{T}}}{\sqrt{C/N}}\right), \text{ when } \hat{\mathbf{X}}^q \in \hat{Q}_i, \tag{4}$$

where $\sqrt{C/N}$ is a scaling factor to balance the multi-head attention [21].

Spatial-Temporal Aggregation. Similar to relevance measuring, we also compute the spatial-temporally aggregated features $\mathcal{M}_i^T$ within the constrained neighborhood during temporal aggregation. This can be formulated as:

$$\mathcal{M}_i^T \in \mathbb{R}^{THW\times \frac{C}{N}} = \mathcal{M}_i^A\mathcal{F}^S\langle \mathbf{X}^a, V_i\rangle, \text{ when } \mathbf{X}^a \in \mathcal{M}_i^A, \tag{5}$$

Soft-Attention. We use a soft-attention block to synthesize features from the group of affinity matrices $\mathcal{M}_i^A$ and aggregated features $\mathcal{M}_i^T$. During the synthesis process, relevant spatial-temporal patterns should be enhanced while less relevant ones should be suppressed. We first concatenate a group of affinity

matrices $\mathcal{M}_i^A$ along the channel dimension to generate $\mathcal{M}^A$. Thus, the soft-attention map $\mathcal{M}^S$ is computed by:

$$\mathcal{M}^S \in \mathbb{R}^{THW\times 1} \leftarrow \texttt{max}\mathcal{M}^A, \text{ when } \mathcal{M}^A \in \mathbb{R}^{THW\times T(2k+1)^2N}, \tag{6}$$

where the `max` function computes the channel-wise maximum value. Then we concatenate a group of the spatial-temporally aggregated features $\mathcal{M}_i^T$ along the channel dimension to generate $\mathcal{M}^T$.

Normalized Self-attention. Finally, our NS block can be computed as:

$$\mathbf{Z} \in \mathbb{R}^{T\times H\times W\times C} = \mathbf{X} + \mathbf{Y} = \mathbf{X} + (\mathcal{M}^T\mathbf{W}_T) \circledast \mathcal{M}^S, \tag{7}$$

where $\mathbf{W}_T$ is the learnable weight and $\circledast$ is the channel-wise Hadamard product.

2.2 Progressive Learning Strategy

Encoder. For fair comparison, we use the same backbone (*i.e.*, Res2Net-50) as in PraNet [8]. Given a polyp video clip with T frames as input (*i.e.*, $\{\mathcal{I}\}_{t=1}^T \in \mathbb{R}^{H'\times W'\times 3}$), we first feed it into a spatial encoder to extract two spatial features from the conv3_4 and conv4_6 layers, respectively. To alleviate the computational burden, we adopt an RFB-like [15] module to reduce the feature channel. Thus, we generate two spatial features, including low-level (*i.e.*, $\{\mathcal{X}_t^l\}_{t=1}^T \in \mathbb{R}^{H^l\times W^l\times C^l}$) and high-level (*i.e.*, $\{\mathcal{X}_t^h\}_{t=1}^T \in \mathbb{R}^{H^h\times W^h\times C^h}$)[1].

Progressively Normalized Self-attention (PNS). Most attention strategies aim to refine candidate features, such as first-order [8] and second-order [22,21] functions. As such, the strong semantic information in high-level features might be diffused gradually during the forward pass of the network. To alleviate this, we introduce a progressive residual learning strategy in our NS block. Specifically, we first reshape the corresponding high-level features $\{\mathcal{X}_t^h\}_{t=1}^T$ of consecutive input frames into a temporal feature, which can be viewed as a four-dimensional tensor (*i.e.*, $\widetilde{\mathcal{X}}^h \in \mathbb{R}^{T\times H^h\times W^h\times C^h}$). Then we refine $\widetilde{\mathcal{X}}^h$ via stacked normalized self-attention in a progressive manner:

$$\widetilde{\mathcal{X}}^r \in \mathbb{R}^{T\times H^h\times W^h\times C^h} = \texttt{NS}^{\times R}(\widetilde{\mathcal{X}}^h) = \texttt{NS}^{\times R}(\mathcal{F}^R(\{\mathcal{X}_t^h\}_{t=1}^T)), \tag{8}$$

where $\texttt{NS}^{\times R}$ means that R normalized self-attention blocks are stacked in the refinement process. $\mathcal{F}^R$ is the reshaping function for the temporal dimension. To allow this block to easily be plugged into pre-trained networks, the commonly adopted solution is to add a residual learning process. Finally, the refined spatial-temporal feature is generated by:

$$\{\mathcal{X}_t^r\}_{t=1}^T \in \mathbb{R}^{H^h\times W^h\times C^h} = \mathcal{F}^R(\widetilde{\mathcal{X}}^h + \widetilde{\mathcal{X}}^r). \tag{9}$$

Decoder and Learning Strategy. We combine the low-level feature $\{\mathcal{X}_t^l\}_{t=1}^T$ from the spatial decoder and the spatial-temporal feature $\{\mathcal{X}_t^r\}_{t=1}^T$ from the PNS block via a two-stage UNet-like decoder $\mathcal{F}^D$. Thus, the output of our method is computed by $\{\mathcal{P}_t\}_{t=1}^T = \mathcal{F}^D(\{\mathcal{X}_t^l\}_{t=1}^T, \{\mathcal{X}_t^r\}_{t=1}^T)$. We adopt the standard *cross-entropy* loss function in the learning process.

[1] We set $H^l = \frac{H'}{4}$, $W^l = \frac{W'}{4}$, $C^l = 24$, $H^h = \frac{H'}{8}$, $W^h = \frac{W'}{8}$, and $C^h = 32$.

3 Experiments

3.1 Implementation Details

Datasets. We adopt four widely used polyp datasets in our experiments, including image-based (*i.e.*, Kvasir [12]) and video-based (*i.e.*, CVC-300 [4], CVC-612 [3], and ASU-Mayo [20]) ones. Kvasir is a large-scale and challenging dataset, which consists of 1,000 polyp images with fully annotated pixel-level ground truths (GTs). The whole Kvasir is used for training. ASU-Mayo contains 10 negative video samples from normal subjects and 10 positive samples from patients. We only adopt the positive part for training. Following the same protocol as [4,3], we split the videos from CVC-300 (12 clips) and CVC-612 (29 clips) into 60% for training, 20% for validation, and 20% for testing.
Training. Due to the limited video training data, we try to fully utilize large-scale image data to capture more appearances of the polyp and scene. Thus, we train our model in two steps: *i) Pre-training phase.* We remove the normalized self-attention (NS) block from *PNS-Net* and pre-train the static backbone using an image-based polyp dataset (*i.e.*, Kvasir [12]) and the training set of video-based polyp datasets (*i.e.*, CVC-300 [4], CVC-612 [3], and ASU-Mayo [20]). The initial learning rate of the Adam algorithm and the weight decay are both 1e-4. The static part of our *PNS-Net* convergences after 100 epochs. *ii) Fine-tuning phase.* We plug the NS block into our *PNS-Net* and fine-tune the whole network using the video polyp datasets, including the ASU-Mayo and the training sets of CVC-300 and CVC-612. We set the number of attention groups $N = 4$ and the number of stacked normalized self-attention blocks $R = 2$, along with a kernel size of $k = 3$. The initial learning is set to 1e-4, and the whole model is fine-tuned over one epoch. In this way, although the densely labeled VPS data is scarce, our *PNS-Net* still achieves good generalization performance.
Testing and Runtime. To test the performance of our *PNS-Net*, we validate it on challenging datasets, including the test set of CVC-612 (*i.e.*, CVC-612-T), the validation set of CVC-612 (*i.e.*, CVC-612-V), and the test/validation set of CVC-300 (*i.e.*, CVC-300-TV). During inference, we sample T=5 frames from a polyp clip and resize them to 256×448 as the input. For final prediction, we use the output $\mathcal{P}_t$ of the network followed by a *sigmoid* function. Our *PNS-Net* achieves a speed of ∼140fps on a single RTX 2080 GPU without any post-processing (*e.g.*, CRF [14]). The speeds of the compared methods are listed in Tab. 1.

3.2 Evaluation on Video Polyp Segmentation

Baselines. We re-train five cutting-edge polyp segmentation baselines (*i.e.*, UNet [19], UNet++ [28], ResUNet [13], ACSNet [25], and PraNet [8]) with the same data used by our *PNS-Net*, under their default settings, for fair comparison.
Metrics. The metrics used included: (1) maximum Dice (maxDice), which measures the similarity between two sets of data; (2) maximum specificity (maxSpe), which refers to the percentage of the samples that are negative and are judged as such; (3) maximum IoU (maxIoU), which measures the overlap between two

Table 1: Quantitative results on different datasets.

		2018∼2019			2020		2021
	Metrics	UNet MICCAI [19]	UNet++ TMI [28]	ResUNet ISM [13]	ACSNet MICCAI [25]	PraNet MICCAI [8]	***PNS-Net* (OUR)**
	Speed	108fps	45fps	20fps	35fps	97fps	**140fps**
CVC-300-TV	maxDice↑	0.639	0.649	0.535	0.738	0.739	**0.840**
	maxSpe↑	0.963	0.944	0.852	0.987	0.993	**0.996**
	maxIoU↑	0.525	0.539	0.412	0.632	0.645	**0.745**
	S_α ↑	0.793	0.796	0.703	0.837	0.833	**0.909**
	E_ϕ ↑	0.826	0.831	0.718	0.871	0.852	**0.921**
	M ↓	0.027	0.024	0.052	0.016	0.016	**0.013**
CVC-612-V	maxDice↑	0.725	0.684	0.752	0.804	0.869	**0.873**
	maxSpe↑	0.971	0.952	0.939	0.929	0.983	**0.991**
	maxIoU↑	0.610	0.570	0.648	0.712	0.799	**0.800**
	S_α ↑	0.826	0.805	0.829	0.847	0.915	**0.923**
	E_ϕ ↑	0.855	0.830	0.877	0.887	0.936	**0.944**
	M ↓	0.023	0.025	0.023	0.054	0.013	**0.012**
CVC-612-T	maxDice↑	0.729	0.740	0.617	0.782	0.852	**0.860**
	maxSpe↑	0.971	0.975	0.950	0.975	0.986	**0.992**
	maxIoU↑	0.635	0.635	0.514	0.700	0.786	**0.795**
	S_α ↑	0.810	0.800	0.727	0.838	0.886	**0.903**
	E_ϕ ↑	0.836	0.817	0.758	0.864	**0.904**	0.903
	M ↓	0.058	0.059	0.084	0.053	0.038	**0.038**

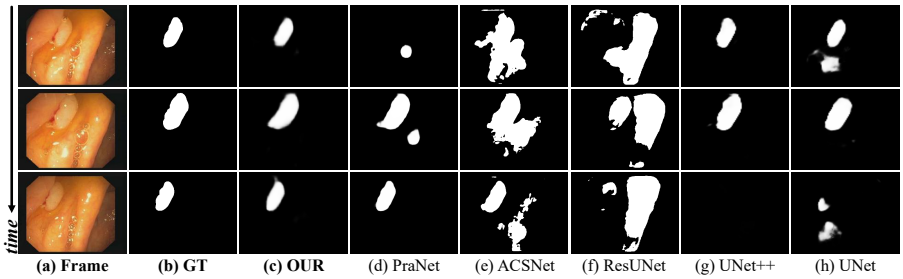


Fig. 2: Qualitative results on CVC-612-T [3]. For more visualization results please refer to the supplementary material (*i.e.*, PDF file and videos).

masks; (4) S-measure [6] (S_α), which evaluates region- and object-aware structural similarity; (5) enhanced-alignment measure [7] (E_ϕ), which measures pixel-level matching and image-level statistics; and (6) mean absolute error (M), which measures the pixel-level error between the prediction and GT.

Qualitative Comparison. In Fig. 2, We provide the polyp segmentation results of our *PNS-Net* on CVC-612-T. Our model can accurately locate and segment polyps in many difficult situations, such as different sizes, homogeneous areas, different textures, *etc.*

Quantitative Comparison. Quantitative comparison results are summarized in Tab. 1. We conduct three experiments on test datasets to verify the model's performance. CVC-300-TV consists of both validation set and test set, which include six videos in total. CVC-612-V and CVC-612-T each contain five videos. On CVC-300, where all the baseline methods perform poorly, our *PNS-Net* achieves remarkable performance in all metrics and outperforms all SOTA methods by

Table 2: Ablation studies. See § 3.3 for more details.

No.	Variants					CVC-300-TV				CVC-612-T			
	Base	N	Soft	Norm	R	maxDice↑	maxIoU↑	$S_\alpha\uparrow$	$E_\phi\uparrow$	maxDice↑	maxIoU↑	$S_\alpha\uparrow$	$E_\phi\uparrow$
#1	✓					0.778	0.665	0.850	0.858	0.850	0.778	0.896	0.885
#2	✓	1			1	0.755	0.650	0.865	0.844	0.850	0.779	0.896	0.891
#3	✓	2			1	0.790	0.679	0.876	0.872	0.825	0.746	0.870	0.856
#4	✓	4			1	0.809	0.709	0.893	0.884	0.834	0.760	0.881	0.867
#5	✓	8			1	0.763	0.663	0.867	0.842	0.787	0.702	0.841	0.829
#6	✓	4	✓		1	0.829	0.729	0.896	0.903	0.852	0.784	0.895	0.897
#7	✓	4	✓	✓	1	0.827	0.732	0.897	0.898	0.856	0.792	0.898	0.896
#8	✓	4	✓	✓	2	**0.840**	**0.745**	**0.909**	**0.921**	**0.860**	**0.795**	**0.903**	**0.903**
#9	✓	4	✓	✓	3	0.737	0.609	0.793	0.751	0.732	0.613	0.776	0.728

a large margin (max Dice: ∼10%). On CVC-612-V and CVC-612-T, our *PNS-Net* consistently outperforms other SOTAs.

3.3 Ablation Study

Effectiveness of Channel Split. We investigate the contribution of channel split rule under different scales. The results are listed in rows #2 to #5 in Tab. 2. We observe that #4 (N=4) outperforms other settings (*i.e.*, #2, #3, and #5) on CVC-300-TV, in all metrics. This improvement shows that an improper receptive field (RF) harms the ability to excavate temporal information, since a large RF will pay more attention to the global environment rather than local motion information. On the other hand, when the split number is too small, the model fails to capture multi-scale polyps moving at various speeds.

Effectiveness of Soft-attention. We further investigate the contribution of the soft-attention mechanism. As shown in Tab. 2, #6 is generally better than #4 with the soft-attention block on CVC-612-T. This improvement suggests that introducing the soft-attention block to synthesize the aggregation feature and affinity matrix is necessary for increasing performance.

Effectiveness of the Number of NS Blocks. To access the number of normalized self-attention blocks under different settings, we derive three variants as #7, #8, and #9. We observe that #8 (*PNS-Net* setting) is significantly better than #7 and #9, with $R = 2$, in all metrics on CVC-300-TV and CVC-612-T. This improvement illustrates that too many iterations of NS blocks may cause overfitting on small datasets (#9). In contrast, the model fails to alleviate the diffusion issue of high-level features with a single residual block. Empirically, we recommend increasing the number of NS blocks when training on larger datasets.

4 Conclusion

We have proposed a self-attention based framework, *PNS-Net*, to accurately segment polyps from colonoscopy videos with super high speed (∼140fps). Our basic normalized self-attention blocks can be easily plugged into existing CNN-based architectures. We experimentally show that our *PNS-Net* achieves the best performance on all existing publicly available datasets under six metrics. Further, extensive ablation studies demonstrate that the core components in our

PNS-Net are all effective. We hope that the proposed *PNS-Net* can serve as a catalyst for progressing both in VPS as well as other closely related video-based medical segmentation tasks. Exploring the performance of *PNS-Net* on a larger VPS dataset will be left to our future work.

References

1. Akbari, M., Mohrekesh, M., Nasr-Esfahani, E., Soroushmehr, S.R., Karimi, N., Samavi, S., Najarian, K.: Polyp segmentation in colonoscopy images using fully convolutional network. In: IEEE EMBC. pp. 69–72 (2018)
2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
3. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. CMIG **43**, 99–111 (2015)
4. Bernal, J., Sánchez, J., Vilarino, F.: Towards automatic polyp detection with a polyp appearance model. PR **45**(9), 3166–3182 (2012)
5. Brandao, P., Mazomenos, E., Ciuti, G., Caliò, R., Bianchi, F., Menciassi, A., Dario, P., Koulaouzidis, A., Arezzo, A., Stoyanov, D.: Fully convolutional neural networks for polyp segmentation in colonoscopy. In: MICAD. vol. 10134, p. 101340F (2017)
6. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: IEEE ICCV. pp. 4548–4557 (2017)
7. Fan, D.P., Ji, G.P., Qin, X., Cheng, M.M.: Cognitive vision inspired object segmentation metric and loss function. SSI (2020)
8. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: MICCAI. pp. 263—-273 (2020)
9. Fang, Y., Chen, C., Yuan, Y., Tong, K.y.: Selective feature aggregation network with area-boundary constraints for polyp segmentation. In: MICCAI. pp. 302–310. Springer (2019)
10. Gu, Y., Wang, L., Wang, Z., Liu, Y., Cheng, M.M., Lu, S.P.: Pyramid constrained self-attention network for fast video salient object detection. In: AAAI. vol. 34, pp. 10869–10876 (2020)
11. Guo, L., Liu, J., Zhu, X., Yao, P., Lu, S., Lu, H.: Normalized and geometry-aware self-attention network for image captioning. In: IEEE CVPR. pp. 10327–10336 (2020)
12. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: MMM. pp. 451–462. Springer (2020)
13. Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P., Johansen, H.D.: Resunet++: An advanced architecture for medical image segmentation. In: IEEE ISM. pp. 225–2255 (2019)
14. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. NIPS **24**, 109–117 (2011)
15. Liu, S., Huang, D., et al.: Receptive field block net for accurate and fast object detection. In: ECCV. pp. 385–400 (2018)
16. Mamonov, A.V., Figueiredo, I.N., Figueiredo, P.N., Tsai, Y.H.R.: Automated polyp detection in colon capsule endoscopy. IEEE TMI **33**(7), 1488–1502 (2014)

17. Murugesan, B., Sarveswaran, K., Shankaranarayana, S.M., Ram, K., Joseph, J., Sivaprakasam, M.: Psi-Net: Shape and boundary aware joint multi-task deep network for medical image segmentation. In: IEEE EMBC. pp. 7223–7226 (2019)
18. Puyal, J.G.B., Bhatia, K.K., Brandao, P., Ahmad, O.F., Toth, D., Kader, R., Lovat, L., Mountney, P., Stoyanov, D.: Endoscopic polyp segmentation using a hybrid 2d/3d cnn. In: MICCAI. pp. 295–305. Springer (2020)
19. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
20. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE TMI **35**(2), 630–644 (2015)
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS (2017)
22. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: IEEE CVPR. pp. 7794–7803 (2018)
23. Yang, F., Yang, H., Fu, J., Lu, H., Guo, B.: Learning texture transformer network for image super-resolution. In: IEEE CVPR. pp. 5791–5800 (2020)
24. Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A.: Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. IEEE JBHI **21**(1), 65–75 (2016)
25. Zhang, R., Li, G., Li, Z., Cui, S., Qian, D., Yu, Y.: Adaptive context selection for polyp segmentation. In: MICCAI. pp. 253–262. Springer (2020)
26. Zhang, R., Zheng, Y., Poon, C.C., Shen, D., Lau, J.Y.: Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker. PR **83**, 209–219 (2018)
27. Zhong, J., Wang, W., Wu, H., Wen, Z., Qin, J.: Polypseg: An efficient context-aware network for polyp segmentation from colonoscopy videos. In: MICCAI. pp. 285–294. Springer (2020)
28. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. IEEE TMI pp. 3–11 (2019)

A.22 Additional papers

In addition to the papers related to the PhD topic listed above, the following papers have been published by the PhD candidate:

- V. Thambawita, S. A. Hicks, H. Borgli, H. Stensland, **D. Jha**, M. K. Svensen, S. A. Pettersen, D. Johansen, H. D. Johansen, S. D. Pettersen, S. Nordvang, S. Pedersen, A. T. Gjerdrum, T. M. Grønli, P. M. Fredriksen, R. Eg, K. S. Hansen, S. Fagernes, C. Claudi, A. Biørn-Hansen, D. T. Nguyen, T. Kupka, H. L. Hammer, R. Jain, M. A. Riegler, P. Halvorsen, "PMData: a sports logging dataset," ***Proceedings of the ACM Multimedia Systems (MMSYS)***, 2020.

- E. Garcia-Ceja, V. Thambawita, S. A. Hicks, **D. Jha**, P. Jakobsen, H. L. Hammer, P. Halvorsen, M. A. Riegler, "HTAD: A Home-Tasks Activities Dataset with Wrist-accelerometer and Audio Features," ***Proceedings of the international conference on multimedia modeling (MMM)***, 2021.

Appendix B

Scientific Activities

Challenges and Workshop Organized

Medico Automatic Polyp Segmentation Challenge (MediaEval 2020)
EndoTect Challenge (ICPR 2020)
3rd International Endoscopy Computer Vision Challenge and Workshop ((EndoCV2021)-part of ISBI 2021)
MedAI: Transparency in Medical Image Segmentation
Medico: Transparency in Medical Image Segmentation

TPC Member

ACM International Conference on Multimedia 2019 (ACMMM 2019)
26th International Conference on MultiMedia Modeling (MMM 2020)
25th International Conference on MultiMedia Modeling (MMM 2019)
IEEE Special Track on Deep Learning Applications in Medical Care 2019
ACM Multimedia Systems Conference 2018
31st IEEE Symposium on Content Based Medical System (CBMS 2018)
MediaEval Benchmarking Initiative for Multimedia Evaluation (MediaEval2018)
27th ACM International Conference on Multimedia
IEEE 33rd International Symposium on Computer Based Medical Systems (CBMS)

Reviewer

IEEE Transactions on Medical Imaging

Pattern Recognition

IEEE Journal of Biomedical and Health Informatics

Neurocomputing

IEEE/ACM Transactions on Computational Biology and Bioinformatics

Artificial Intelligence In Medicine

Computers in Biology and Medicine

World Journal of Gastrointestinal Endoscopy

Signal Processing: Image Communication

Mathematical Biosciences and Engineering

Medical Physics

ACM Multimedia Conference 2020

IET Image Processing

Electronics Letter

Digital Signal Processing

Frontiers in Artificial Intelligence

PLOS One

13th International Symposium on Medical Information and Communication Technology

IEEE International Conference on Multimedia and Expo (ICME 2020)

IEEE International Symposium on Computer Based Medical Systems (CBMS 2020)

IEEE International Conference on Multimedia and Expo (ICME) 2021

Master Thesis Supervision

Rabindra Khadga, Meta-Learning for Medical Image Segmentation, master thesis, 2021.

Supervision of Summer Intern

Krister Emanuelsen