ACM Multimedia BioMedia 2019 Grand Challenge Overview

Steven Hicks*
SimulaMet, Norway

Michael Riegler
SimulaMet, Norway

Pia Smedsrud†
Augere Medical AS, Norway

Trine B. Haugen
Oslo Metropolitan University, Norway

Kristin Ranheim Randel
Cancer Registry of Norway

Konstantin Pogorelov
Simula Research Laboratory, Norway

Håkon Kvale Stensland
Simula Research Laboratory, Norway

Duc-Tien Dang-Nguyen
University of Bergen, Norway

Mathias Lux
University of Klagenfurt, Austria

Andreas Petlund
Augere Medical AS, Norway

Thomas de Lange
University of Oslo, Norway

Peter Thelin Schmidt
Karolinska Hospital, Sweden

Pål Halvorsen*
SimulaMet, Norway

ABSTRACT

The BioMedia 2019 ACM Multimedia Grand Challenge is the first in a series of competitions focusing on the use of multimedia for different medical use-cases. In this year's challenge, the participants are asked to develop efficient algorithms which automatically detect a variety of findings commonly identified in the gastrointestinal (GI) tract (a part of the human digestive system). The purpose of this task is to develop methods to aid medical doctors performing routine endoscopy inspections of the GI tract. In this paper, we give a detailed description of the four different tasks of this year's challenge, present the datasets used for training and testing, and discuss how each submission is evaluated both qualitatively and quantitatively.

CCS CONCEPTS

• **Applied computing** → **Consumer health**; **Health informatics**; • **Computing methodologies** → *Supervised learning*; • **Information systems** → Summarization;

KEYWORDS

Computer-aided Diagnosis, Multimedia, Medical

ACM Reference format:
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, Thomas de Lange, Peter Thelin Schmidt, and Pål Halvorsen. 2019. ACM Multimedia BioMedia 2019 Grand Challenge Overview. In *Proceedings of Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, October 21–25, 2019 (MM '19),* 5 pages.
https://doi.org/10.1145/3343031.3356058

*Also affiliated with Oslo Metropolitan University, Norway
†Also affiliated with SimulaMet, Norway

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3356058

1 INTRODUCTION

The BioMedia 2019 ACM Multimedia Grand Challenge[1] is a competition which aims at using multimedia-based approaches to tackle real-world medical challenges. This year, the competition focuses on analyzing images and videos taken from the human digestive tract to automatically detect various findings such as disease, anatomical landmarks, or other relevant findings. This challenge can be seen as an extension of the *Medico: Multimedia for Medicine* task which has previously been held at MediaEval Benchmark [11]. Making the future of GI examinations more efficient and cost-effective would be a substantial achievement, as about 2.8 million new cases of esophagus, stomach, and colorectal cancers are detected yearly with a mortality of about 65% worldwide [7].

Inspection of the digestive tract is done through a procedure called an endoscopy, which is a medical procedure where an endoscope is inserted either orally (gastroscopy) or rectally (colonoscopy) in order to directly examine the GI tract for various diseases. The endoscope is equipped with a tiny camera, for which a doctor can directly view a live feed from the GI tract on an external monitor and evaluate the video in real-time. One limitation with these examinations is that it is mostly dependant on the experience of the doctor operating the endoscope. This variation depends on operator skill, perceptual factors, personality characteristics, knowledge, and attitude [1]. The consequence of this translates into a substantial inter-observer variation in the detection and assessment of mucosal lesions [5, 13], leading to an average polyp miss-rate of 20% in the colon [4]. Seeing as the procedure is primarily based on the visual inspection of videos, this gives us as computer scientists the opportunity to perform some preliminary analysis on frames produced during an endoscopy. Giving the doctors a so-called "third-eye"

[1]https://github.com/kelkalot/biomedia-2019

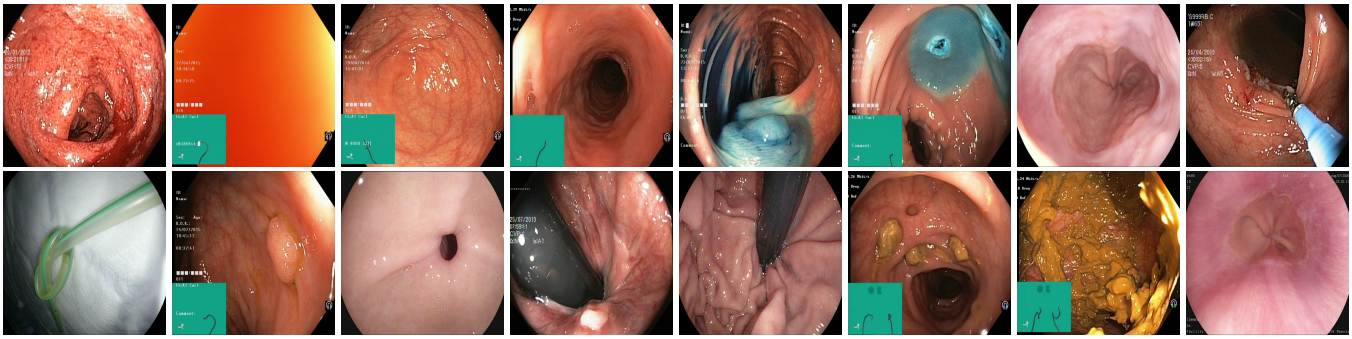

Figure 1: Example images taken from each of the 16 different classes included in the image dataset. From left to right, top to bottom: *ulcerative-colitis, blurry-nothing, normal-cecum, colon-clear, dyed-lifted-polyp, dyed-resection-margins, esophagitis, instruments, out-of-patient, polyps, normal-pylorus, retroflex-rectum, retroflex-stomach, stool-inclusions, stool-plenty,* and *normal-z-line.*

could potentially help quickly identify disease or other findings which may have otherwise been missed.

Recent appearances of several medical multimedia related conferences show that there is much interest in the field of medical multimedia research [3, 12, 14]. Furthermore, applying the existing multimedia research to the field of medicine has the potential of making a significant impact on society as a whole, as the current state of a lot of medical practice lack the modern advancements of computer-based analysis of images and videos. In the case of endoscopies, we see this as a perfect use-case as it includes requirements of real-time analysis and is a medical procedure which is quite common and very important.

The competition proposes four different tasks, each targeting a specific use-case or requirement the GI endoscopy procedure. The first task relates to achieving good classification performance on 16 different findings from the entire GI tract. The second task is quite similar to the first but focuses more on the efficiency of the proposed algorithm. This task asks the participants to measure the time it takes for their algorithm to classify a single image (to measure whether or not the method can be used in real-time). The third task expands on the previous by having the algorithm be evaluated on the same hardware, therefore making a fair comparison between the different submissions. The last task focuses on summarizing the findings in endoscopy videos into a comprehensible report, therefore giving doctors a quick and easy way of interpreting the predictions of the algorithm, and potentially saving time when collecting data for the electronic medical report.

2 DATASET DETAILS

The 2019 BioMedia presents two different datasets, i.e., one containing images and the other containing videos. Each dataset is tailored to a specific task. The detection, efficient detection, and hardware tasks use the image dataset, while the report task uses the video dataset. This section will describe the two datasets in detail.

2.1 Image Dataset

The image-based dataset consists of 14,033 images taken from 16 different classes captured during a typical endoscopy procedure. Each image has been manually annotated and verified by experienced endoscopists to create the ground truth [9, 10]. As it is essential to detect more than just diseases in a routine investigation of the GI tract, the dataset also contains a wide variety of typical findings commonly found in the GI tract.

An example from each class in the challenge can be seen in Figure 1. The classes can be categorized into five different subgroups, namely anatomical landmarks, pathological findings, pre-, while- and post-surgery findings, bowel cleanliness, and images not usable for diagnosis. The classes under **anatomical landmarks** are *normal-z-line*, *normal-pylorus*, *normal-cecum*, *retroflex-rectum*, *retroflex-stomach*. These classes are important to detect as they signal the end of one part of the GI tract (such as the Z-line signaling the end of the esophagus). The **pathological findings** include *esophagitis*, *polyps* and *ulcerative-colitis*. Reporting the pathological findings is very important as it is the main purpose of the endoscopy procedure. The **pre-, while- and post-surgery findings** are the *dyed-lifted-polyps*, the *dyed-resection-margins* and the *instruments*. Documenting the surgeries is essential in writing a complete endoscopy report and to verfiy that the procedure was done correctly. The classes related to the **bowel cleanliness** include *colon-clear*, *stool-inclusions* and *stool-plenty*. These classes relate to normal tissue and is expected to be the majority finding in a healthy individual. Lastly, there are some image classes that are **not usable for diagnosis**, namely *blurry-nothing* and *out-of-patient*. Although not used for diagnosis, it is still useful to classify these images for the purpose of sorting and filtering.

Overall, the dataset is quite unbalanced with the majority class being *stool-plenty* (2,331 images) and the minority class being *out-of-patient* (6 images). This lopsided distribution can be seen as part of the challenge and reflects the real-world data collection at hospitals. The resolution of the included images ranges from 720×576 up to 1920×1072 pixels. For this competition, the dataset was split into a development and test dataset consisting of 5,293

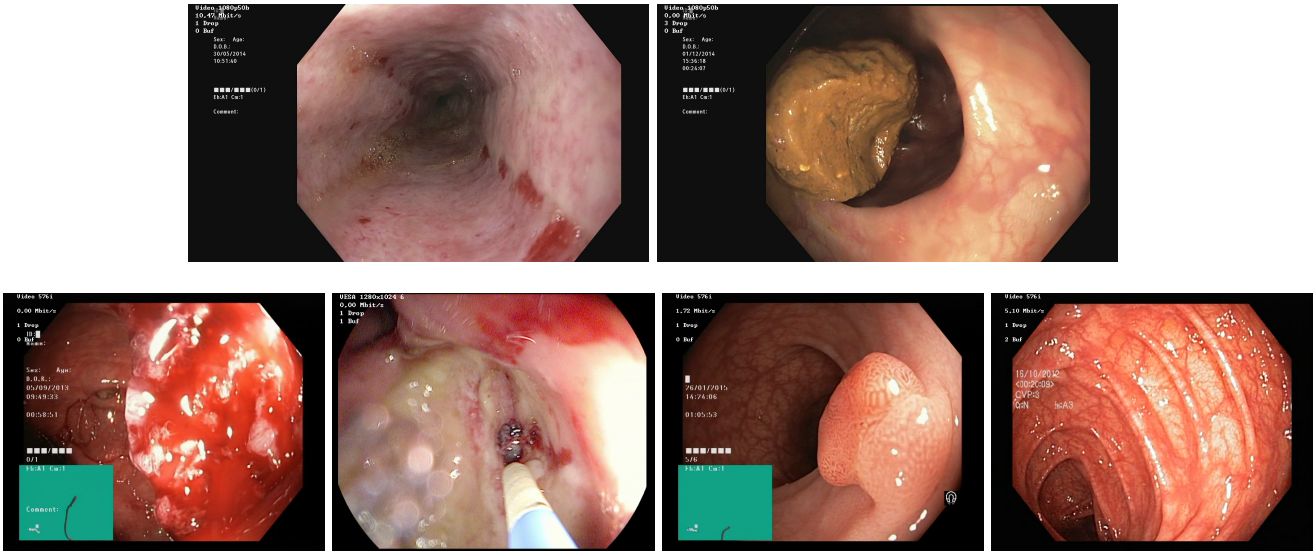

Figure 2: Example frames taken from each of the six videos in the video dataset. Each frame is presented in their original aspect ratio.

Video ID	Expected findings	Length	Resolution
3e3a7ac0-4244-46cc-89a1-44ce84dd1ccf	esophagitis	00:51	1920 × 1072
3f450a06-3397-48ed-ab27-5bd184af2862	stool	00:02	1920 × 1072
4c0cae22-4945-4f17-a60b-23688b485d91	polyp resection, bleeding after	02:00	720 × 576
5c812c3f-33e6-4e1a-b708-637e76ab9244	bleeding ulcer, instrument	01:08	1280 × 1024
17d4f706-b2c0-4d46-9082-92094e6e90fc	polyp, lifting and resection, instrument	05:11	720 × 576
c205ec73-4652-4f91-8711-ddb6a50e1d16	normal colon	00:57	720 × 576

Table 1: An overview of the videos included for the automatic report generation task. The expected findings describe what can be found in each video. No exact time stamps are used for the evaluation because the report assessment is performed manually by two medical doctors who have much experience with colonoscopies.

and 8,740 images, respectively. Both the development dataset and test dataset were available to the participants, but we withheld the ground truth of the test dataset. The images in the development dataset are stored in separate directories reflected by the name of the class each image belongs. The images in the test dataset are all stored in the same directory to hide the ground truth. Every image in both the development and test dataset is compressed using JPEG compression, but the encoding settings may differ between images.

Additionally, we also provide some pre-extracted image features for all images in both the development and test dataset. All features are extracted using the popular image retrieval library LIRE [6], for which the selected image features included in the dataset are JCD, Tamura, ColorLayout, EdgeHistogram, AutoColorCorrelogram, and PHOG. The file structure of the image features mimics that of their image counterparts for both datasets. Each feature file is stored as a text file using the file extension ".features." Within each file, there are eight lines, one for each of the previously described image features. Every line is made up of the name of the feature and a series of floating-point numbers representing the image feature vector.

2.2 Video Dataset

The video dataset consists of 6 different videos ranging from 00 : 02 seconds to 05 : 11 minutes in length with resolutions of up to 1920 × 1072 pixels. The videos contain various findings similar to what can be found in the image dataset. We expect that the participants use the image dataset to train a model, then apply this model to the videos to create the report submission. More details about the 6 videos can be found in Table 1 and an example frame from each video can be seen in Figure 2.

3 EVALUATION METRICS AND TASKS

BioMedia 2019 has four different tasks; the detection task, the efficient detection task, the hardware task, and the report generation task. To participate in the competition, each team must submit at least one submission to the detection task. They may also submit to any other task and may submit as many submissions as they wish. The evaluation script used to evaluate the submissions to the *detection task*, *efficient detection task*, and *hardware task* is available

on GitHub[2]. Two medical doctors will manually evaluate the *report generation task*. In the following few sections, we give a more detailed explanation of each task and report how each submission will be evaluated and ranked.

3.1 Detection Task (required)

The *detection task* aims to satisfy the requirement of the high detection accuracy needed to be viable for deployment in a clinical setting. Participants are asked to submit runs which achieve high classification scores on the 16 different classes previously presented in Section 2. To submit a run to this task, participants must create a ".csv" file which contains one line per image prediction. Each line should start with the name of the predicted file, followed by the predicted label, then end with the model's confidence of that prediction. For the evaluation of detection accuracy, we use several standard metrics commonly used to evaluate classification tasks such as precision, recall/sensitivity, specificity, F1, and matthews correlation coefficient (MCC) for multi-classification (also called R_k statistic). 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 (optional)

The *efficient detection task* addresses the requirement of real-time analysis needed to be able to deliver instant feedback to doctors performing a standard endoscopy procedure. In order to fulfill this requirement, the algorithm must achieve good classification results in addition to being able to classify images as fast as possible, i.e., the frames should be processed at least as fast as the video frame rate. Speed should be measured on the participant's computers on what could be considered consumer-grade PC hardware (no supercomputers or large clusters). Submissions to this task are quite similar to that of the *detection task*, the only difference being that the processing time for each image should be appended to each prediction line of the ".csv" file.

The classification performance of the algorithm will be evaluated in the same way as what was described in Section 3.1 and will be ranked according to the achieved MCC. Speed will be measured based on the average time it takes to classify a single image in milliseconds. The submissions will be ranked based on a combination of the requirement for real-time detection and the overall classification performance of the proposed algorithm. To balance these two requirements, we set a threshold of 85% on specificity and sensitivity [8], which is a standard threshold used in the industry for an automatic detection system for colonoscopies. All submission which reaches this threshold is compared based on their processing time per image. If two teams have the same time, the one with the higher sensitivity and specificity score is taken as the better performing one.

3.3 Hardware Task (optional)

The *hardware task* is quite similar to the *efficient detection* detection task as it will be evaluated based on the efficiency of the proposed algorithm. What differentiates this task from *efficient detection*, is that it requires teams to submit their code in the form of a Docker image so that we can evaluate their submissions on the same hardware (more on the submission format can be found in the submission tutorial available on github[3]). The hardware used to evaluate each submission can be considered as consumer-grade. It contains an Intel Core i7-7700K processor, a single GTX 1080 Ti graphical processing unit (GPU), 16 gigabytes of RAM, and it is running Arch Linux. The submitted Docker image should produce the same submission file as described in the *efficient detection task*, and will, therefore, be evaluated in the same way.

3.4 Report Generation Task (optional)

In the *report generation task*, we ask participants to automatically generate endoscopy reports by analyzing the six videos provided in the video dataset. The report should summarize the videos by detailing the different findings found within each video. The videos do not include a ground truth, so it is expected that the participants train a model using the provided image dataset and use the resulting model to analyze the videos and create the report. Two medical experts perform the evaluation, each experienced with the colonoscopy procedure for more than ten years. The expectation for the generated report is somewhat open beside some minimal requirements (the report should at least detect one finding per video). The visual design and look can be seen as part of the challenge (how to present the results in the best way to the medical experts, for example as shown by Hicks et al. [2]). Overall, the report will be evaluated based on correctness, innovation, and usefulness. The medical experts will also provide information about what could be improved.

4 DISCUSSION AND OUTLOOK

In this paper, we described the BioMedia 2019 competition, which is part of the ACM Multimedia grand challenge 2019 track. The competition focuses on the use-case of analyzing endoscopy videos and images to aid medical doctors to detect and document various disease and other important findings. We presented the four different tasks that are part of this year's competition, which ranges from simple image classification to automatic report generation based on analysis of videos. The competition also focuses on system performance aspects in terms of the time used for making a classification. The BioMedia challenge's primary purpose is to encourage multimedia researchers to explore the field of medical multimedia. In the future, we hope to be able to continue the challenge over the next few years with different medical use-cases each year. The 2020 version would, for example, focus on human reproduction and multimodal analysis of video, sensor data, and person-related data.

REFERENCES

[1] David G Hewett, Charles J Kahi, and Douglas K Rex. 2010. Efficacy and effectiveness of colonoscopy: how do we bridge the gap? *Gastrointestinal Endoscopy Clinics* 20, 4 (2010), 673–684.

[2] Steven Alexander Hicks, Sigrun Eskeland, Mathias Lux, Thomas de Lange, Kristin Ranheim Randel, Mattis Jeppsson, Konstantin Pogorelov, Pål Halvorsen, and Michael Riegler. 2018. Mimir: an automatic reporting and reasoning system for deep learning based analysis in the medical domain. In *Proceedings of the 9th ACM Multimedia Systems Conference (MMSYS)*. ACM, 369–374.

[2] https://github.com/stevenah/biomedia-2019-submission-evaluation

[3] https://www.github.com/stevenah/biomedia-2019-sample-submission

[3] Bogdan Ionescu, Henning Müller, Mauricio Villegas, Helbert Arenas, Giulia Boato, Duc-Tien Dang-Nguyen, Yashin Dicente Cid, Carsten Eickhoff, Alba Garcia Seco de Herrera, Cathal Gurrin, Bayzidul Islam, Vassili Kovalev, Vitali Liauchuk, Josiane Mothe, Luca Piras, Michael Riegler, and Immanuel Schwall. 2017. Overview of ImageCLEF 2017: Information extraction from images. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction 8th International Conference of the CLEF Association, CLEF 2017 (LNCS 10439)*. Springer.

[4] 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. https://doi.org/10.1056/NEJMoa0907667

[5] Si Hyung Lee, Byung Ik Jang, Kyeong Ok Kim, Seong Woo Jeon, Joong Goo Kwon, Eun Young Kim, Jin Tae Jung, Kyung Sik Park, Kwnag Bum Cho, Eun Soo Kim, et al. 2014. 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, 2 (2014), 154.

[6] Mathias Lux and Savvas A Chatzichristofis. 2008. Lire: lucene image retrieval: an extensible java cbir library. In *Proceedings of the 16th ACM international conference on Multimedia (ACM MM)*. ACM, 1085–1088.

[7] World Health Organization et al. 2014. GLOBOCAN 2012: Estimated cancer incidence, mortality and prevalence worldwide in 2012. *Lyon, France: International Agency for Research on Cancer* (2014).

[8] 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 *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE.

[9] 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.

[10] 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, Michael Riegler, and Pål Halvorsen. 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.

[11] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Steven Hicks, Thomas de Lange, Kristin Ranheim Randel, Duc-Tien Dang-Nguyen, Mathias Lux, and Olga Ostroukhova. 2018. Medico Multimedia Task at MediaEval 2018. In *CEUR Workshop Proceedings - Multimedia Benchmark Workshop (MediaEval)*.

[12] Michael Riegler, Mathias Lux, Carsten Gridwodz, Concetto Spampinato, Thomas de Lange, Sigrun L Eskeland, Konstantin Pogorelov, Wallapak Tavanapong, Peter T Schmidt, Cathal Gurrin, Dag Johansen, Håvard Johansen, and Pål Halvorsen. 2016. Multimedia and Medicine: Teammates for better disease detection and survival. In *Proceedings of the 2016 ACM Multimedia Conference (ACM MM)*. ACM, 968–977.

[13] Sascha C Van Doorn, Y Hazewinkel, James E East, Monique E Van Leerdam, Amit Rastogi, Maria Pellisé, Silvia Sanduleanu-Dascalescu, Barbara AJ Bastiaansen, Paul Fockens, and Evelien Dekker. 2015. Polyp morphology: an interobserver evaluation for the Paris classification among international experts. *The American journal of gastroenterology* 110, 1 (2015), 180.

[14] Mauricio Villegas, Henning Müller, Alba García Seco de Herrera, Roger Schaer, Stefano Bromuri, Andrew Gilbert, Luca Piras, Josiah Wang, Fei Yan, Arnau Ramisa, et al. 2016. General overview of imageCLEF at the CLEF 2016 labs. In *Procedings of the International Conference of the Cross-Language Evaluation Forum for European Languages (LNCS 9822)*. Springer, 267–285.