KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection

Konstantin Pogorelov
Simula Research Laboratory, Norway
University of Oslo, Norway

Kristin Ranheim Randel
Cancer Registry of Norway
University of Oslo, Norway

Carsten Griwodz
Simula Research Laboratory, Norway
University of Oslo, Norway

Sigrun Losada Eskeland
Bærum Hospital, Norway

Thomas de Lange
Bærum Hospital, Norway
Cancer Registry of Norway

Dag Johansen
UiT-The Arctic University of Norway

Concetto Spampinato
University of Catania, Italy

Duc-Tien Dang-Nguyen
Dublin City University, Ireland

Mathias Lux
University of Klagenfurt, Austria

Peter Thelin Schmidt
Karolinska Institutet, Solna, Sweden
Karolinska Hospital, Sweden

Michael Riegler
Simula Research Laboratory, Norway
University of Oslo, Norway

Pål Halvorsen
Simula Research Laboratory, Norway
University of Oslo, Norway

ABSTRACT

Automatic detection of diseases by use of computers is an important, but still unexplored field of research. Such innovations may improve medical practice and refine health care systems all over the world. However, datasets containing medical images are hardly available, making reproducibility and comparison of approaches almost impossible. In this paper, we present KVASIR, a dataset containing images from inside the gastrointestinal (GI) tract. The collection of images are classified into three important anatomical landmarks and three clinically significant findings. In addition, it contains two categories of images related to endoscopic polyp removal. Sorting and annotation of the dataset is performed by medical doctors (experienced endoscopists). In this respect, KVASIR is important for research on both single- and multi-disease computer aided detection. By providing it, we invite and enable multimedia researcher into the medical domain of detection and retrieval.

ACM Reference format:
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 MMSys '17, Taipei, Taiwan, June 20–23, 2017,* 6 pages.
https://doi.org/http://dx.doi.org/10.1145/3083187.3083212

1 INTRODUCTION

The human digestive system may be affected by several diseases. As an example, three of the eight most common cancers overall are located in the gastrointestinal (GI) tract (figure 1). Altogether esophageal, stomach and colorectal cancer accounts for about 2.8 million new cases and 1.8 million deaths per year [40]. Endoscopic examinations (figures 2(a) and 2(b)) are the gold standards for investigation of the GI tract. Gastroscopy is an examination of the upper GI tract including esophagus, stomach and first part of small bowel, while colonoscopy covers the large bowel (colon) and rectum. Both these examinations are real-time video examinations of the inside of the GI tract by use of digital high definition endoscopes (figures 2(c)). Endoscopic examinations are resource demanding and requires both expensive technical equipment and trained personnel.

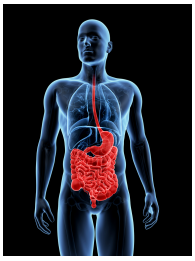

Figure 1: GI tract (shutterstock).

For colorectal cancer prevention, endoscopic detection and removal of possible precancerous lesions are essential. Adenoma detection is therefore considered to be an important quality indicator in colorectal cancer screening. However, the ability to detect adenomas varies between doctors, and this may eventually affect the individuals' risk of getting colorectal cancer [19].

Endoscopic assessment of severity and sub-classification of different findings may also vary from one doctor to another. Accurate grading of diseases are important since it may influence decision-making on treatment and follow-up [4, 11, 16]. For example, the degree of inflammation directly affects the choice of therapy in inflammatory bowel diseases (IBD) [37]. An objective and automated scoring system would therefore be highly welcomed.

Automatic detection, recognition and assessment of pathological findings will probably contribute to reduce inequalities, improve quality and optimize use of scarce medical resources. Furthermore, since endoscopic examinations are real-time investigations, both normal and abnormal findings have to be recorded and documented within written reports. Automatic report generation would probably contribute to reduce doctors' time required for paperwork and thereby increase time to patient care. Reliable and careful documentation with the use of minimal standard terminology (MST) [1]

This work is founded by the Norwegian FRINATEK project "EONS" (#231687).
Contact author: Konstantin Pogorelov, email: konstantin@simula.no .

MMSys '17, June 20–23, 2017, Taipei, Taiwan

ACM ISBN 978-1-4503-5002-0/17/06.
https://doi.org/http://dx.doi.org/10.1145/3083187.3083212


(a) Colonoscopy


(b) Gastroscopy


(c) A colonoscope

Figure 2: Various types of endoscopy examinations.

may also contribute to improved patient follow-up and treatment. To our knowledge, a standardized and automatic reporting system that ensure high quality endoscopy reports does not exist.

In order to make the health care system more scalable and cost effective, basic research in the intersection between computer science and medicine must go beyond traditional medical imaging by combining this area with multimedia data analysis and retrieval, artificial intelligence, and distributed processing. Next-generation medical big-data applications are a frontier for innovation, competition and productivity, where there are currently large initiatives both in the EU [26] and the US [24]. In the area of multimedia research, people are starting to see the synergies between multimedia and medical systems [31]. For automatic algorithmic detection of abnormalities in the GI tract, there have been many proposals from various research communities. For example, many systems present promising results for polyp detection [3, 5, 9, 18, 21, 23, 32, 38, 39, 41] reaching high precision and recall scores. However, the results are hard to reproduce due to lack of available medical data, i.e., the work listed above all use different data sets ranging from 35 to 1.8 million images/video frames.

In our earlier work [27, 32, 33], we have used the two only usable, publicly available GI tract datasets: the ASU-Mayo Clinic polyp database [35] and the CVC-ColonDB colonoscopy video database [7]. The ASU-Mayo dataset consists of training and test sets of images and videos with corresponding ground truth showing the exact polyp location areas. This is currently the biggest available dataset consisting of 20 videos from standard colonoscopies with a total of 18, 781 frames and different resolution up to full HD. However, the images in this dataset are very similar raising the challenge of overfitting, and currently, the use of the dataset is restricted. The CVC-ColonDB dataset consists of images and videos partially covered by corresponding ground truth showing the exact polyp location areas. This is currently the second biggest available dataset consisting of 15 small videos from standard colonoscopies with a total of 1, 200 frames and 300 frames with the region of interest marked. The resolution is 500x574 pixels. Furthermore, both these datasets contain only one endoscopic finding (polyps). In this paper, we therefore publish Kvasir our multi-class dataset[1] from the Vestre Viken Health Trust (Norway) containing not only polyps, but also two other findings, two classes related to polyp removal and three anatomical landmarks in the GI tract.

2 DATA COLLECTION

The data is collected using equipment as shown in figure 2(c) at Vestre Viken Health Trust (VV) in Norway. The VV consists of 4 hospitals and provides health care to 470.000 people. One of these hospitals (the Bærum Hospital) has a large gastroenterology department from where training data have been collected and will be provided, making the dataset larger in the future. Furthermore, the images are carefully annotated by one or more medical experts from VV and the Cancer Registry of Norway (CRN). The CRN provides new knowledge about cancer through research on cancer. It is part of South-Eastern Norway Regional Health Authority and is organized as an independent institution under Oslo University Hospital Trust. CRN is responsible for the national cancer screening programmes with the goal to prevent cancer death by discovering cancers or pre-cancerous lesions as early as possible.

3 DATASET DETAILS

The initial Kvasir dataset consists of 4, 000 images, annotated and verified by medical doctors (experienced endoscopists), including 8 classes showing anatomical landmarks, phatological findings or endoscopic procedures in the GI tract, i.e., 500 images for each class. The number of images is sufficient to be used for different tasks, e.g., image retrieval, machine learning, deep learning and transfer learning, etc. [2, 12, 28]. The anatomical landmarks are Z-line, pylorus and cecum, while the pathological finding includes esophagitis, polyps and ulcerative colitis. In addition, we provide two set of images related to removal of polyps, the "dyed and lifted polyp" and the "dyed resection margins". The dataset consist of the images with different resolution from 720x576 up to 1920x1072 pixels and organized in a way where they are sorted in separate folders named accordingly to the content. Some of the included classes of images have a green picture in picture illustrating the position and configuration of the endoscope inside the bowel, by use of an electromagnetic imaging system (ScopeGuide, Olympus Europe) that may support the interpretation of the image. This type of information may be important for later investigations (thus included), but must be handled with care for the detection of the endoscopic findings.

3.1 Anatomical Landmarks

An anatomical landmark is a recognizable feature within the GI tract that is easily visible through the endoscope. They are essential for navigating and as a reference point to describe the location of a given finding. The landmarks may also be typical sites for pathology like ulcers or inflammation. A complete endoscopic rapport should preferably contain both brief descriptions and image documentation of the most important anatomical landmarks [30].

3.1.1 Z-line. The Z-line marks the transition site between the esophagus and the stomach. Endoscopically, it is visible as a clear border where the white mucosa in the esophagus meets the red gastric mucosa. An example of the Z-line is shown in figure 3. Recognition and assessment of the Z-line is important in order to determine whether disease is present or not. For example, this is the area where signs of gastro-esophageal reflux may appear. The Z-line is

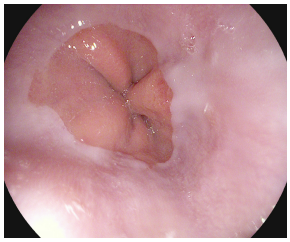
Figure 3: Z-line

[1] http://datasets.simula.no/kvasir

also useful as a reference point when describing pathology in the esophagus.

3.1.2 Pylorus. The pylorus is defined as the area around the opening from the stomach into the first part of the small bowel (duodenum). The opening contains circumferential muscles that regulates the movement of food from the stomach. The identification of pylorus is necessary for endoscopic instrumentation to the duodenum, one of the challenging maneuvers within gastroscopy. A complete gastroscopy includes inspection on both sides of the pyloric opening to reveal findings like ulcerations, erosions or stenosis. Figure 4 shows an endoscopic image of a normal pylorus viewed from inside the stomach. Here, the smooth, round opening is visible as a dark circle surrounded by homogeneous pink stomach mucosa.

Figure 4: Pylorus

3.1.3 Cecum. The cecum the most proximal part of the large bowel. Reaching cecum is the proof for a complete colonoscopy and completion rate has shown to be a valid quality indicator for colonoscopy [6]. Therefore, recognition and documentation of the cecum is important. One of the characteristics hallmarks of cecum is the appendiceal orifice. This combined with a typical configuration on the electromagnetic scope tracking system may be used as proof for cecum intubation when named or photo documented in the reports [29, 36]. Figure 5 shows an example of the appendiceal orifice visible as a crescent shaped slit, and the green picture in picture shows the scope configuration for cecal position.

Figure 5: Cecum

3.2 Phatological findings

A pathological finding in this context is an abnormal feature within the gastrointestinal tract. Endoscopically, it is visible as a damage or change in the normal mucosa. The finding may be signs of an ongoing disease or a precursor to for example cancer. Detection and classification of pathology is important in order to initiate correct treatment and/or follow-up of the patient.

3.2.1 Esophagitis. Esophagitis is an inflammation of the esophagus visible as a break in the esophageal mucosa in relation to the Z-line. Figure 6 shows an example with red mucosal tongues projecting up in the white esophageal lining. The grade of inflammation is defined by length of the mucosal breaks and proportion of the circumference involved. This is most commonly caused by conditions where gastric acid flows back into the esophagus as gastroesophageal reflux, vomiting or hernia. Clinically, detection is necessary for treatment initiation to relieve symptoms and prevent further development of possible complications. Computer detection would be of special value in assessing severity and for automatic reporting.

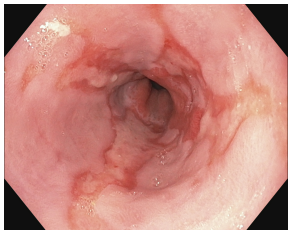

Figure 6: Esophagitis

3.2.2 Polyps. Polyps are lesions within the bowel detectable as mucosal outgrows. An example of a typical polyp is shown in figure 7. The polyps are either flat, elevated or pedunculated, and can be distinguished from normal mucosa by color and surface pattern. Most bowel polyps are harmless, but some have the potential to grow into cancer. Detection and removal of polyps are therefore important to prevent development of colorectal cancer. Since polyps may be overlooked by the doctors, automatic detection would most likely improve examination quality. The green boxes within the image shows an illustration of the endoscope configuration. In live endoscopy, this helps to determine the current localisation of the endoscope-tip (and thereby also the polyp site) within the length of the bowel. Automatic computer aided detection of polyps would be valuable both for diagnosis, assessment and reporting.

Figure 7: Polyp

3.2.3 Ulcerative colitis. Ulcerative colitis is a chronic inflammatory disease affecting the large bowel. The disease may have a large impact on quality of life, and diagnosis is mainly based on colonoscopic findings. The degree of inflammation varies from none, mild, moderate and severe, all with different endoscopic aspects. For example, in a mild disease, the mucosa appears swollen and red, while in moderate cases, ulcerations are prominent. Figure 8 shows an example of ulcerative colitis with bleeding, swelling and ulceration of the mucosa. The white coating visible in the illustration is fibrin covering the wounds. As mentioned earlier, an automatic computer aided assessment system will contribute to more accurate grading of the disease severity.

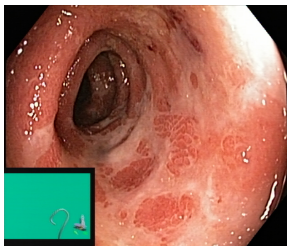

Figure 8: Ulcerative colitis

3.3 Polyp removal

Polyps in the large bowel may be precursors of cancer and are therefore removed during endoscopy if possible. One of the polyp removal techniques is called endoscopic mucosal resection (EMR). This includes injection of a liquid underneath the polyp, lifting the polyp from the underlying tissue. The polyp is then captured and removed by use of a snare. The lifting minimizes risk of mechanical or electrocautery damage to the deeper layers of the GI wall. Staining dye (i.e., diluted indigo carmine) is added to facilitate accurate identification of the polyp margins [17]. Computer detection of dyed polyps and the site of resection would be important in order to generate computer aided reporting systems for the future.

3.3.1 Dyed and Lifted Polyps. Figure 9 shows an example of a polyp lifted by injection of saline and indigocarmine. The light blue polyp margins are clearly visible against the darker normal mucosa. Additional valuable information related to automatic reporting may involve successfulness of the lifting and eventual presence of non-lifted areas that might indicate malignancy.

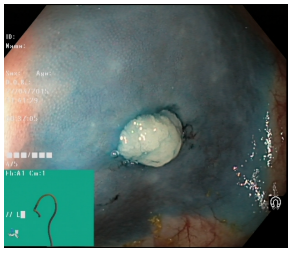

Figure 9: Dyed and Lifted Polyp

3.3.2 Dyed Resection Margins. The resection margins are important in order to evaluate whether the polyp is completely removed or not. Residual polyp tissue may lead to continued growth and in worst case malignancy development. Figure 10 illustrates the resection site after removal of a polyp. Automatic recognition of the site of polyp removals are of value for automatic reporting systems and for computer aided assessment on completeness of the polyp removal.

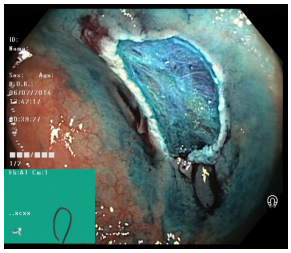

Figure 10: Dyed Resection Margin

4 APPLICATIONS OF THE DATASET

Our vision is that the available data may eventually help researchers to develop systems that improve the health-care system in the context of disease detection in videos of the GI tract. Such a system may automate video analysis and endoscopic findings detection in the esophagus, stomach, bowel and rectum. Important results will include higher detection accuracies, reduced manual labor for medical personnel, reduced average cost, less patient discomfort and possibly increased willingness to undertake the examination. In the end, the improved screening will probably significantly reduce mortality and number of luminal GI disease incidents.

With respect to direct use in the multimedia research areas, the main application area of Kvasir is automatic detection, classification and localization of endoscopic pathological findings in an image captured in the GI tract. Thus, the provided dataset can be used in several scenarios where the aim is to develop and evaluate algoritmic analysis of images. Using the same collection of data, researchers can easier compare approaches and experimental results, and results can easier be reproduced. In particular, in the area of image retrieval and object detection, Kvasir will play an important initial role where the image collection can be divided into training and test sets for developments of and experiments for various image retrieval and object localization methods including search-based systems, neural-networks, video analysis, information retrieval, machine learning, object detection, deep learning, computer vision, data fusion and big data processing.

In our work [27, 32, 33], we have for example conducted a leave-one-out cross-validation to evaluate our system. This is a method that assesses the generalization of a predictive model where the training and testing datasets are rotated, i.e., leaving out a single different non-overlapping item or portion for testing, and using the remaining items for training. This process is repeated until every item or portion has been used for testing exactly once [13]. Being one of the first medical multi-class datasets available to the multimedia community, we hereby invite and enable multimedia researcher into the medical domain of detection and retrieval.

5 SUGGESTED METRICS

Looking at the list of related work in this area, there are a lot of different metrics used, with potentially different names when used in the medical area and the computer science (information retrieval) area. Here, we provide a small list of the most important metrics. For future research, in addition to describing the dataset with respect to total number of images, total number of images in each class and total number of positives, it might be good to provide as many of the metrics below as possible in order to enable an indirect comparison with older work:

True positive (TP): The number of correctly identified samples. The number of frames with an endoscopic finding which correctly is identified as a frame with an endoscopic finding.

True negative (TN): The number of correctly identified negative samples, i.e., frames without an endoscopic finding which correctly is identified as a frame without an endoscopic finding.

False positive (FP): The number of wrongly identified samples, i.e., a commonly called a "false alarm". Frames without an endoscopic finding which is erroneously identified as a frame with an endoscopic finding.

False negative (FN): The number of wrongly identified negative samples. Frames without an endoscopic finding which erroneously is identified as a frame with an endoscopic finding.

Recall (REC): 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:

$$recall = \frac{TP}{\#\ of\ all\ positives} = \frac{TP}{TP + FN}$$

Precision (PREC): 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):

$$precision = \frac{TP}{\#\ of\ all\ returned\ samples} = \frac{TP}{TP + FP}$$

Specificity (SPEC): This metric is frequently called the *true negative rate*, and shows the ratio of negatives that are correctly identified as such (e.g., the fraction of frames without an endoscopic finding are correctly identified as a negative result):

$$specificity = \frac{TN}{\#\ of\ all\ negatives} = \frac{TN}{FP + TN}$$

Accuracy (ACC): The percentage of correctly identified true and false samples:

$$accuracy = \frac{TP + TN}{\#\ of\ samples\ in\ total}$$

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:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

F1 score (F1): A measure of a test's accuracy by calculating the harmonic mean of the precision and recall:

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall} = \frac{2TP}{2TP + FP + FN}$$

In addition to the above metrics, system performance metrics processing speed and resource consumption are of interest. In our work, we have used the achieved frame-rate (FPS) as a metric as real-time feedback is important.

6 BASELINE PERFORMANCE

We have here performed an initial multi-class detection experiment on Kvasir as a baseline for future experiments. We have experimented using various configurations of three different main approaches, i.e., classification using global features (GF), deep learning convolutional neural networks (CNN) and transfer learning in deep learning (TFL).

For the GF approaches, we extracted several image features for classification using the latest version of the Lire open source software [22], i.e., the extracted features are JCD, Tamura, Color Layout, Edge Histogram, Auto Color Correlogram and Pyramid Histrogram of Oriented Gradients. For the 2 GF run, we combined JCD and Tamura resulting in a feature vector of 187. For the 6 GF run, we combined all extracted features resulting in a feature vector of 1186. We decided for these combinations based on our previous findings and experiments in [32]. We performed a simple early fusion of the features, and all extracted features are included in the dataset in the *arff* file format for reuse and reproducibility. We used the Random Forrest (RF) and Logistic Model Tree (LMT) classifiers provided in the Weka machine learning library [15].

For all deep learning implementations, we used Keras [10] with Google Tensorflow [2] as backend. For the two CNN runs, we trained two different CNNs from scratch, i.e., one with three convolution layers and one with six. As activation function, we used the rectified linear unit (ReLU) [14] and for pooling maxpooling. In all layers, we also included a 0.5 dropout, and the final classification step was performed using two dense layers with first ReLU and then Sigmoid as activation functions. Both networks were trained for 200 epochs using the Adam optimizer [20].

The TFL run is based on transfer learning [8] by re-training and fine-tuning the pre-trained Inception v3 model [34]. For the re-training, we followed a similar approach as presented in [12]. Firstly, we locked all the basic convolutional layers of the network and only retrained the two top dense classification layers. The dense layers were retrained for 1,000 epochs using the RMSprop optimizer that allows an adaptive learning rate during the training process. After that, fine-tuning of a subset of the convolutional layers was performed. We decided to apply the fine-tuning on the two top convolutional layers of the re-trained model. For this training step, we used the SGD optimizer with a low learning rate (to achieve the best effect in terms of speed and accuracy) [25].

Table 1: Classification performance in terms of weighted average (2-folded) using the metrics described above.

Method	PREC	REC	SPEC	ACC	MCC	F1	FPS
6 Layer CNN	0.661	0.640	0.953	0.914	0.602	0.651	43
3 Layer CNN	0.589	0.408	0.890	0.959	0.430	0.453	45
Inception v3 TFL	0.698	0.689	0.957	0.924	0.649	0.693	66
2 GF Random Forrest	0.713	0.715	0.959	0.928	0.672	0.711	333
2 GF Logistic Model Tree	0.706	0.707	0.958	0.926	0.664	0.705	210
6 GF Random Forrest	0.732	0.732	0.962	0.933	0.692	0.727	105
6 GF Logistic Model Tree	0.748	0.748	0.964	0.937	0.711	0.747	80
Baseline (JCD Random Forrest)	0.708	0.710	0.958	0.927	0.666	0.706	370
Baseline (Random/Majority)	0.016	0.125	0.000	0.016	0.666	0.000	-

Table 2: Confusion matrix for both cross validated folds for the 6 GF LMT experiment in table 1. The classes are Esophagitis (A), Dyed and Lifted Polyps (B), Dyed Resection Margins (C), Cecum (D), Pylorus (E), Z-line (F), Polyps (G) and Ulcerative colitis (H). The test set in each fold contains 250 images for each class.

		Detected class							
		A	B	C	D	E	F	G	H
Actual class	A	**198/177**	0/0	0/0	0/0	3/8	49/64	0/1	0/0
	B	0/0	**139/149**	104/92	4/0	0/0	1/0	1/7	1/2
	C	0/0	90/100	**154/148**	2/0	0/0	1/0	2/1	1/1
	D	0/0	0/1	0/0	**214/223**	0/0	0/0	30/18	6/8
	E	5/3	0/0	0/0	0/0	**235/227**	2/8	5/12	3/0
	F	64/33	0/0	0/0	0/0	6/6	**180/210**	0/0	0/1
	G	0/0	0/0	4/1	24/26	10/2	2/2	**169/178**	41/41
	H	1/0	2/0	1/0	18/8	3/1	1/1	32/44	**192/196**

The exact configurations of the CNN and TFL approaches are included in the dataset. We did not perform any data augmentation, such as cropping, for any of the approaches for this work. For the experiments, we split the dataset randomly in two equally sized subsets (training and testing) containing 250 images per class each. We also performed two-folded cross-validation by switching the training and testing and calculated the average. As baselines, we provide one using the RF classifier with the JCD feature and one based on the random/majority class.

Table 1 gives and overview of the results, and table 2 contains the confusion matrix for the best performing approach (6 GF with LMT) for a more detailed insight into the performance. We can see that all approaches would outperform the random and majority class baseline, which is presented in the last row. Our own baseline in the second last row is only outperformed by three approaches. The best performing approach is a combination of six global features and the LMT classifier with an overall F1 score of 0.747 and 80 FPS. The 6 layer CNN outperforms the 3 layer CNN in terms of detection performance but not in terms of speed. The TFL approach outperforms the two other deep learning based approaches, which we expected since our CNN parameters are not optimized and we trained over a rather small number of epochs. Nevertheless, even if we use very basic methods, the here presented results can be a good starting point for other researchers and used as baselines to benchmark other methods applied to the dataset. In short, we see that multi-class detection is much more challenging than single detection, and that some findings are harder to detect than others, indicating that there are great potentials for improvements and innovations in future medical multimedia research.

7 CONCLUSION

To enable (reproducible) research in the intersection between multimedia and medicine, on analysis of images and videos of the human

GI tract in particular, we have presented the Kvasir dataset. The dataset has been collected during real endoscopy examinations and sorted and analyzed by medical experts. Initially, it contains 8 classes of images of important lesions and landmarks found in the GI tract, but it will be continuously updated. Medical datasets are hard to find, and such a dataset enables multi-disciplinary retrieval and detection research in order to improve health care systems all over the world.

REFERENCES

[1] Lars Aabakken, Alan N Barkun, Peter B Cotton, Evgeny Fedorov, Masayuki A Fujino, katerina Ivanova, Shin ei Kudo, Konstantin Kuznetzov, Thomas de Lange, Koji Matsuda, Olivier Moine, BjÃűrn Rembacken, Jean-Francois Rey, Joseph Romagnuolo, Thomas Rösch, Mandeep Sawhney, Kenshi Yao, and Jerome D Waye. 2014. Standardized endoscopic reporting. *J. of Gastroenterology and Hepatology* 29, 2 (2014), 234–240.

[2] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, and others. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).

[3] Luís A Alexandre, Joao Casteleiro, and Nuno Nobreinst. 2007. Polyp detection in endoscopic video using SVMs. In *Proc. of PKDD*. 358–365.

[4] Y. Amano, N. Ishimura, K. Furuta, K. Okita, M. Masaharu, T. Azumi, T. Ose, K. Koshino, S. Ishihara, K. Adachi, and Y. Kinoshita. 2006. Interobserver Agreement on Classifying Endoscopic Diagnoses of Nonerosive Esophagitis. *Endoscopy* 38, 10 (2006), 1032–1035.

[5] Stefan Ameling, Stephan Wirth, Dietrich Paulus, Gerard Lacey, and Fernando Vilarino. 2009. Texture-based polyp detection in colonoscopy. In *Bildverarbeitung für die Medizin*. Springer, 346–350.

[6] Nancy N. Baxter, Rinku Sutradhar, Shawn S. Forbes, Lawrence F. Paszat, Refik Saskin, and Linda Rabeneck. 2011. Analysis of administrative data finds endoscopist quality measures associated with postcolonoscopy colorectal cancer. *Gastroenterology* 140, 1 (2011), 65–72.

[7] Jorge Bernal, F. Javier Sanchez, and Fernando Vilarino. 2012. Towards Automatic Polyp Detection with a Polyp Appearance Model. *Pattern Recognition* 45, 9 (2012), 3166–3182.

[8] Souad Chaabouni, Jenny Benois-Pineau, and Chokri Ben Amar. 2016. Transfer learning with deep networks for saliency prediction in natural video. In *Proc. of ICIP*. 1604–1608.

[9] Da-Chuan Cheng, Wen-Chien Ting, Yung-Fu Chen, Qin Pu, and Xiaoyi Jiang. 2008. Colorectal polyps detection using texture features and support vector machine. In *Advances in Mass Data Analysis of Images and Signals in Medicine, Biotechnology, Chemistry and Food Industry*. 62–72.

[10] François Chollet. 2015. Keras: Deep learning library for theano and tensorflow. (2015). https://keras.io/ Accessed: 2017-04-19.

[11] Thomas de Lange, Stig Larsen, and Lars Aabakken. 2004. Inter-observer agreement in the assessment of endoscopic findings in ulcerative colitis. *BMC Gastroenterology* 4, 9 (2004).

[12] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition.. In *Proc. of ICML*, Vol. 32. 647–655.

[13] Bradley Efron and Robert Tibshirani. 1997. Improvements on Cross-Validation: The .632+ Bootstrap Method. *J. Amer. Statist. Assoc.* 92, 438 (1997), 548–560.

[14] Richard HR Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung. 2000. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature* 405, 6789 (2000), 947–951.

[15] 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.

[16] Lorenza A Herrero, Wouter L Curvers, Frederike G van Vilsteren, Herbert Wolfsen, Krish Ragunath, Louis-Michel W Song, Rosalie C Mallant-Hent, Arnoud van Oijen, Pieter Scholten, Erik J Schoon, Ed B Schenk, Bas L Weusten, and Jacques G Bergman. 2013. Validation of the Prague C&M classification of Barrett's esophagus in clinical practice. *Endoscopy* 45, 11 (2013), 876–882.

[17] Joo Ha Hwang, Vani Konda, Barham K Abu Dayyeh, Shailendra S Chauhan, Brintha K Enestvedt, Larissa L Fujii-Lau, Sri Komanduri, John T Maple, Faris M Murad, Rahul Pannala, Nirav C. Thosani, and Subhas Banerjee. 2015. Endoscopic mucosal resection. *Gastrointestinal Endoscopy* 82, 2 (2015), 215–226.

[18] Sae Hwang, JungHwan Oh, W. Tavanapong, J. Wong, and P.C. de Groen. 2007. Polyp Detection in Colonoscopy Video using Elliptical Shape Feature. In *Proc. of ICIP*. 465–468.

[19] M. F. Kaminski, J. Regula, E. Kraszewska, M. Polkowski, U. Wojciechowska, J. Didkowska, M. Zwierko, M. Rupinski, M. P. Nowacki, and E. Butruk. 2010. Quality indicators for colonoscopy and the risk of interval cancer. *New England Journal of Medicine* 362, 19 (2010), 1795–1803.

[20] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[21] Baopu Li and M.Q.-H. Meng. 2012. Tumor Recognition in Wireless Capsule Endoscopy Images Using Textural Features and SVM-Based Feature Selection. *IEEE Trans. Information Technology in Biomedicine* 16, 3 (May 2012), 323–329.

[22] Mathias Lux, Michael Riegler, Pål Halvorsen, Konstantin Pogorelov, and Nektarios Anagnostopoulos. 2016. LIRE: open source visual information retrieval. In *Proc. of MMSys*. Article no. 30.

[23] A.V. Mamonov, I.N. Figueiredo, P.N. Figueiredo, and Y.-H.R. Tsai. 2014. Automated Polyp Detection in Colon Capsule Endoscopy. *IEEE Transactions on Medical Imaging* 33, 7 (July 2014), 1488–1502.

[24] McKinsey Global Institute. 2013. The big-data revolution in US health care: Accelerating value and innovation. (2013). https://goo.gl/SqS5Dl Accessed: 2017-04-19.

[25] Jiquan Ngiam, Adam Coates, Ahbik Lahiri, Bobby Prochnow, Quoc V Le, and Andrew Y Ng. 2011. On optimization methods for deep learning. In *Proc. of ICML*. 265–272.

[26] PMLiVE (Dominic Tyer). 2014. European Commission forms EUR2.5bn big data partnership. (2014). https://goo.gl/NeKb7H Accessed: 2017-04-19.

[27] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Peter T Schmidt, Carsten Griwodz, Dag Johansen, Sigrun L Eskeland, and Thomas de Lange. 2016. GPU-accelerated Real-time Gastrointestinal Diseases Detection. In *Proc. of CBMS*.

[28] Ariadna Quattoni, Michael Collins, and Trevor Darrell. 2008. Transfer learning for image classification with sparse prototype representations. In *Proc. of CVPR*. 1–8.

[29] Douglas K Rex, Philip S Schoenfeld, Jonathan Cohen, Irving M Pike, Douglas G Adler, M Brian Fennerty, John G Lieb, Walter G Park, Maged K Rizk, Mandeep S Sawhney, Nicholas J Shaheen, Sachin Wani, and David S Weinberg. 2015. Quality indicators for colonoscopy. *American J. of Gastroenterology* 110, 1 (2015), 72–90.

[30] J.-F. Rey, R. Lambert, and the ESGE Quality Assurance Committee. 2001. ESGE recommendations for quality control in gastrointestinal endoscopy: guidelines for image documentation in upper and lower GI endoscopy. *Endoscopy* 33, 10 (2001), 901–903.

[31] Michael Riegler, Mathias Lux, Carsten Griwodz, 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 *Proc. of ACM MM*. 968–977.

[32] Michael Riegler, Konstantin Pogorelov, Pål Halvorsen, Thomas de Lange, Carsten Griwodz, Peter Thelin Schmidt, Sigrun L. Eskeland, and Dag Johansen. 2016. EIR - Efficient Computer Aided Diagnosis Framework for Gastrointestinal Endoscopies. In *Proc. of CBMI*.

[33] Michael Riegler, Konstantin Pogorelov, Jonas Markussen, Mathias Lux, Håkon Kvale Stensland, Thomas de Lange, Carsten Griwodz, Pål Halvorsen, Dag Johansen, Peter T Schmidt, and Sigrun L. Eskeland. 2016. Computer Aided Disease Detection System for Gastrointestinal Examinations. In *Proc. of MMSys*.

[34] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567* (2015).

[35] Nima Tajbakhsh, Suryakanth Gurudu, and Jianming Liang. 2015. Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information. *IEEE Transactions on Medical Imaging* 35, 2 (feb 2015), 630–644.

[36] R. Valori, J-F. Rey, W. S. Atkin, M. Bretthauer, C. Senore, G. Hoff, E. J. Kuipers, L. Altenhofen, R. Lambert, and G. Minoli. 2012. European guidelines for quality assurance in colorectal cancer screening and diagnosis. First Edition – Quality assurance in endoscopy in colorectal cancer screening and diagnosis. *Endoscopy* 44, S03 (2012), SE88–SE105.

[37] A.J. Walsh, A. Ghosh, A.O. Brain, O. Buchel, D. Burger, S. Thomas, L. White G.S., Collins, S. Keshav, and S.P.L. Travis. 2014. Comparing disease activity indices in ulcerative colitis. *Journal of Crohn's and Colitis* 8, 4 (2014), 318–325.

[38] Yi Wang, Wallapak Tavanapong, Johnson 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.

[39] Yi Wang, Wallapak Tavanapong, Johnny Wong, Jung Hwan Oh, and Piet C de Groen. 2015. Polyp-Alert: Near Real-time Feedback during Colonoscopy. *Computer methods and programs in biomedicine* 3 (2015), 164–179.

[40] World Health Organization - International Agency for Research on Cancer. 2012. Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012. (2012). https://goo.gl/IgZpVl Accessed: 2017-04-19.

[41] Mingda Zhou, Guanqun Bao, Yishuang Geng, B. Alkandari, and Xiaoxi Li. 2014. Polyp detection and radius measurement in small intestine using video capsule endoscopy. In *Proc. of BMEI*. 237–241.