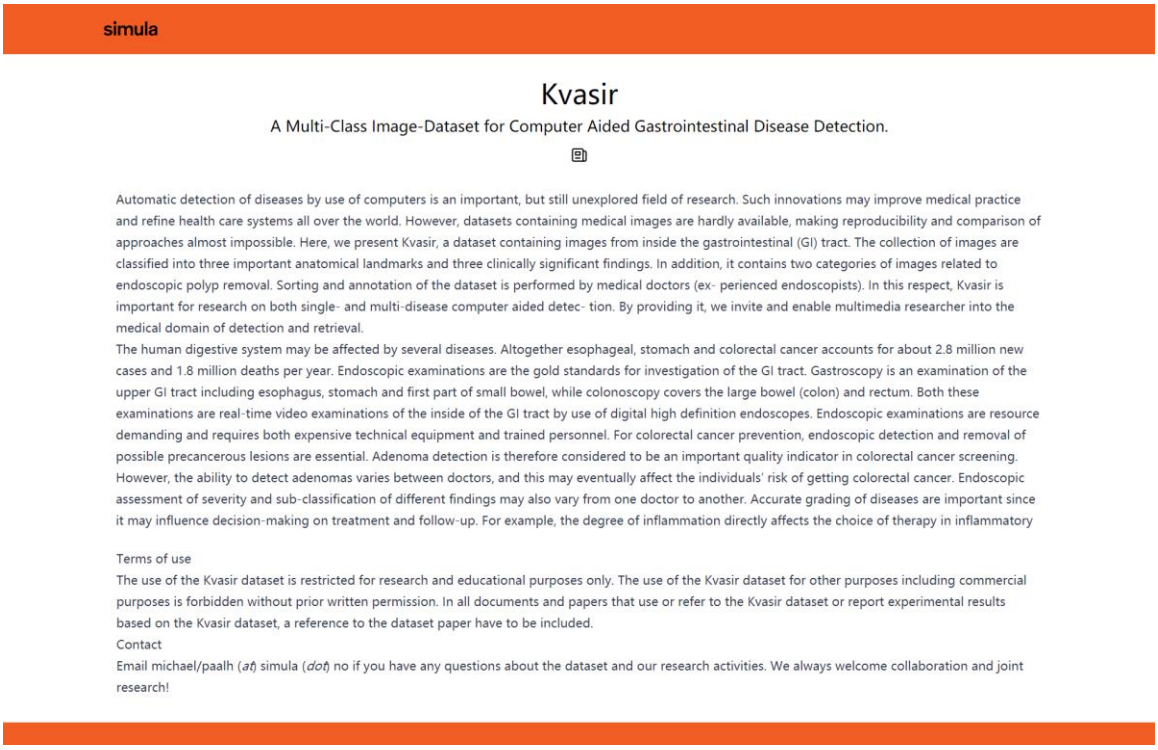
Dear editor,

The datasets from which the images represented in Figure 5 were extracted (Kvasir and HyperKvasir) were published by Simula Research Laboratory under the CC-BY-NC and CC-BY licenses, respectively, as shown:

https://datasets.simula.no//kvasir/

https://datasets.simula.no/hyper-kvasir/

Therefore, the reproduction of images in the submitted article is authorized by the licenses, as long as the relevant citations are made (which were made). Adaptations in the figure legend were made, as requested, and the new submission file is attached. Best regards.



simula

Kvasir

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

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. Here, 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 (ex- perienced endoscopists). In this respect, Kvasir is important for research on both single- and multi-disease computer aided detec- tion. By providing it, we invite and enable multimedia researcher into the medical domain of detection and retrieval.

The human digestive system may be affected by several diseases. Altogether esophageal, stomach and colorectal cancer accounts for about 2.8 million new cases and 1.8 million deaths per year. Endoscopic examinations 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. Endoscopic examinations are resource demanding and requires both expensive technical equipment and trained personnel. 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. 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. For example, the degree of inflammation directly affects the choice of therapy in inflammatory

Terms of use

The use of the Kvasir dataset is restricted for research and educational purposes only. The use of the Kvasir dataset for other purposes including commercial purposes is forbidden without prior written permission. In all documents and papers that use or refer to the Kvasir dataset or report experimental results based on the Kvasir dataset, a reference to the dataset paper have to be included.

Contact

Email michael/paalh (*at*) simula (*dot*) no if you have any questions about the dataset and our research activities. We always welcome collaboration and joint research!

simula

HyperKvasir

The Largest Gastrointestinal Dataset.

Dataset Details

The dataset can be split into four distinct parts; Labeled image data, unlabeled image data, segmented image data, and annotated video data. Each part is further described below.

Labeled images In total, the dataset contains 10,662 labeled images stored using the JPEG format. The images can be found in the images folder. The classes, which each of the images belongto, correspond to the folder they are stored in (e.g., the 'polyp' folder contains all polyp images, the 'barretts' folder contains all images of Barrett's esophagus, etc.). 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. The labeled images represent 23 different classes of findings.

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 cluster assignments in the Hyper-Kvasir 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.

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

Terms of Use

In all documents and papers that use or refer to the dataset or report experimental results based on the Hyper-Kvasir, a reference to the related article needs to be added: https://www.nature.com/articles/s41597-020-00622-y.

Contact

Please contact steven@simula.no, michael@simula.no, or paalh@simula.no for any questions regarding the dataset.