Chapter 1

Methodology

1.1 Data collection

The datasets used in our experiments are Kvasir Pillcam, Kvasir and Hyper-Kvasir. This section will demonstrate the main differences between the three datasets and explain how they can be found and used for fact checking. All three datasets are collected using endoscopic equipment at Vestre Viken Health Trust in Norway. The VV consists of 4 hospitals and provides health care for 470.000 people. One of the hospitals is Bærum Hospital, which has a large gastroenterology department from where the data is collected.

1.1.1 Kvasir PillCam

The dataset we used in our experiments consist of endoscopic videos collected from Bærum Hospital. Unlike Kvasir and Hyper-Kvasir datasets we have made the Kvasir PillCam dataset for the purpose of this thesis. In total we have 44 videos which have gone through some re encoding to reduce the file sizes, and also because the original encoding is proprietary Sony technology. After that the videos are uploaded to Augere Medical [1] tagging tool. The data export from Bærum also contains some findings for each video (if there is any) and are extracted, converted to frame number and that helped us a great deal with tagging the videos. We also have Thomas de Lange to thank, because he helped us a lot with the medical aspect of the classification process. When all 44 videos have been precisely labeled the dataset is exported from Augeres tagging tool and split into folders for each class. The folders/classes are given in table **??**. In total we have 44 000 labeled images in 8 classes. The sample distribution across the eight classes is skewed depending on how many findings there are in the videos. Some findings occur often and some very rarely. The dataset also contain one class for 'normal' images, which there is quite a bit more of than findings.

Imbalanced dataset pose a challenge for predictive algorithms as most learning algorithms are based on the assumption of an equal number of samples for each class. This results in models that have poor predictive performance, especially for minority class or classes. This is a great problem because in many medical datasets the minority class is the most important and therefore more sensitive for classification errors.

In addition to labeling the images the dataset also contain a JSON format file which stores coordinates for where in the frame the finding is located. The Kvasir Pillcam dataset will be an open-source dataset available for others scientists, and will later be grown to include more PillCam videos, both labeled and unlabeled samples.

[1] `https://augere.md/`

Class number	Class name
0	normal
1	polyp
2	polyrus

Table 1.1: PillCam class names and corresponding class numbers.

Class number	Class name	Number of samples
0	normal	8000
1	polyp	8000
2	polyrus	8000

Table 1.2: Kvasir class names and corresponding class numbers.

1.1.2 Kvasir-V2

The Kvasir dataset **KVASIRMultiClass17** contains images from inside the gastrointestinal (GI) tract. The samples are classified into three important anatomical landmarks and three clinically significant findings. In addition it has two classes related to the removal procedure of polyps. The dataset is sorted and annotated is performed by medical doctors. The class names and findings for each class is given in table **??**. One of the most important aspects of the Kvasir dataset is that it makes it easy to reproduce and compare results in scientific computing.

1.1.3 Hyper Kvasir

The Hyper-Kvasir dataset **HyperKvasirComprehensive19** is one of the largest medical datasets containing 110.079 images and 373 videos where it captures anatomical landmarks and pathological and normal findings. Resulting in more than 1.1 million images and video frames all together. The dataset contain four parts, labeled images, unlabeled images, segmented images and lastly, videos. In total the dataset is 70 GB in size, but can be downloaded and stored in parts from `https://datasets.simula.no/hyper-kvasir/`.

Labeled images

Hyper-Kvasir contains 10.662 labeled images. The images are split into 23 different classes, and are stored in a folder with the same name as its corresponding class. All of the images are stored in JPEG format **JPEGStill92**, which means it has some image quality loss but quite insignificant compared to the reduction in file size. Like in situations most often encountered the classes has a different number of samples, this is a challenge in the medical field because some findings occur more often than others.

Unlabeled images

This part of the dataset contains 99.417 images

segmented images

In figure **??** we can see an example of the segmented Kvasir images.

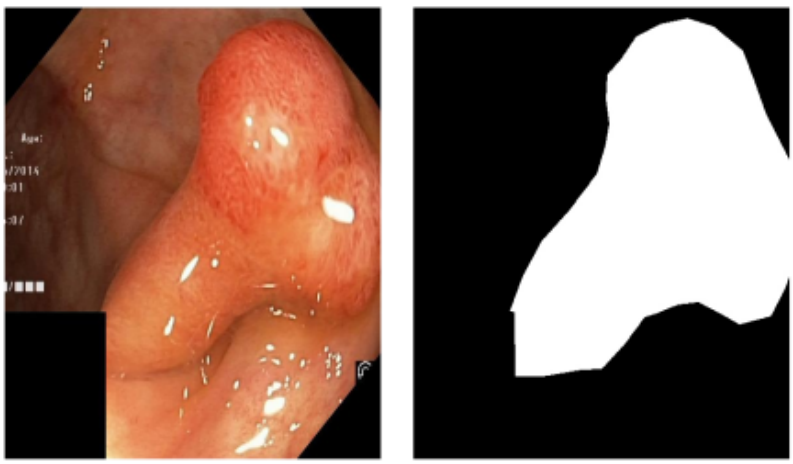

Figure 1.1: Example of a segmented image from Hyper-Kvasir dataset.

Videos

1.2 Data process

1.2.1 Data preprocessing

1.2.2 Data pipeline

1.3 System implementation

1.4 Summary