CHALLENGES IN THE AUTOMATIC ANALYSIS OF IMAGES FROM CAPSULE ENDOSCOPY

Edyta Rogula, Amelia Okulewicz, Julia Pietraszek

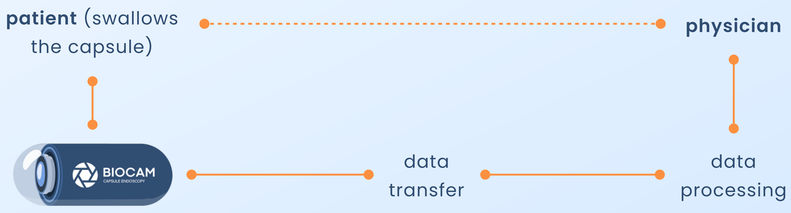
BIOCAM

CAPSULE ENDOSCOPY
PROBLEM DEFINITION

Capsule endoscopy is a noninvasive diagnostic procedure that uses a tiny wireless capsule to take pictures of the gastrointestinal tract.

At Biocam, our main task is to detect any pathologies that can be found in the digestive tract in order to speed up the diagnosis of the physician.

We use machine learning techniques to find and mark images with abnormalities, so the physician can focus on analysing them without fear of missing anything that's important.



ANALYSIS ONLY BY A PHYSICIAN

- thousands of images from one examination
- tedious
- easy to miss minor abnormalities

AI-POWERED ANALYSIS

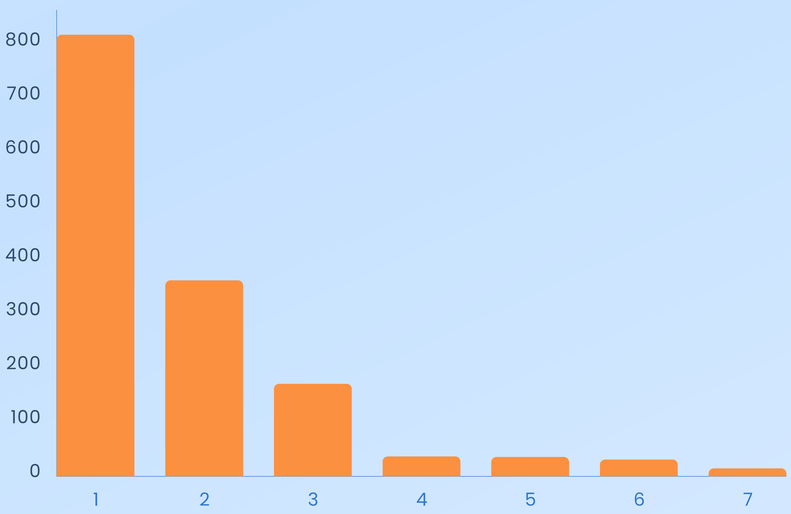
- 3 times decreased time of examination
- harder to miss abnormalities
- faster results

CHALLENGES
LACK OF LABELLED DATA

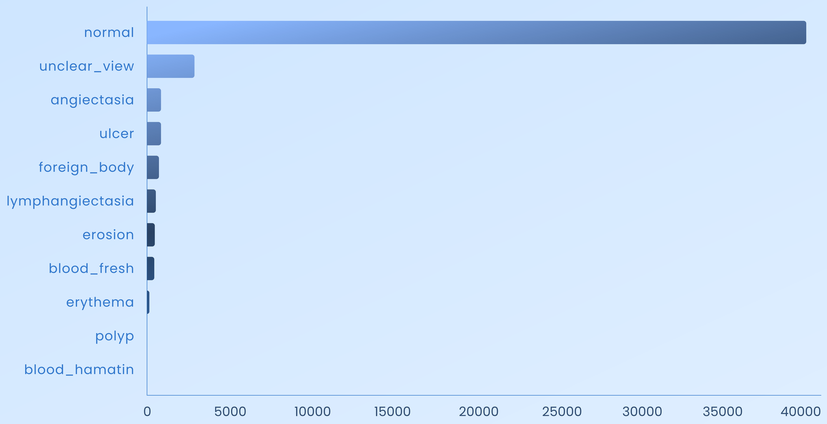
The biggest challenge in solving this problem is lack of labelled data. Although there may be tens of thousands of images from one examination, most of this data is normal (without any pathology) and images close to each other in one sequence look more or less alike. Thus, even in the rather optimistic situation of several thousand pathological images from one examination, they look similar to each other and the model cannot generalise well to images from other examinations. There is also a huge data imbalance in the pathological images - some pathologies are much more numerous than others.

For example, for the pathology named angiectasia, there are only 6 examinations that include this pathology in the Kvasir Capsule dataset, and most of the data comes from only one of them. This means that when we feed this data to the model, it mainly learns one version of this pathology (and its background) and might have a problem with generalisation.

Number of images of angiectasia in each examination



Number of images of each class in the Kvasir Capsule dataset



DATA SPLIT

Due to the similarity of images from one examination, the train-test split is not an easy task. The images next to each other in the sequence are usually almost identical and having one of them in the train set and the other in the test set can artificially increase our results. Ideally, some examinations would be included in the train set and others in the test set, but some pathologies occur in a small number of examinations (e.g. 1 or 2), so it wouldn't test them reliably. An in-between solution is to divide the sequence into packs of several dozen images and place some of these packs in the test set, and the rest in the train set.

DATA LABELLING
SUBJECTIVITY OF DIAGNOSIS VS ML MODEL

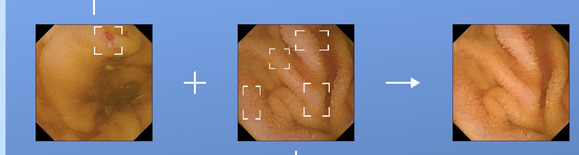
The major problem to resolve for the labelling process is addressing the gap in the image perception: one held by healthcare practitioners and the other represented by a machine learning perspective. As a representative example, doctors understand lymphangiectasia as a very broad term but in case of labelling for model training purposes it should be considered only as a single white spot on an image, without considering the nature of the pathology.

Furthermore, it is critical to remember that a diagnosis may differ from one doctor to the other and is subjective, especially in rare cases. The machine learning view stays in total opposition: the model is intended to generalise in predicted manner for similar data. This highlights the importance of ensuring that the labelling team accurately represents the viewpoint of a typical specialist in a field.

SOLUTIONS
GENERATE THE DATA

The challenges described above lead to lack of accurately labelled, well-balanced sets of diverse images required to train models in order to identify all types of pathologies. Deep learning models used in medicine cannot correctly operate on insufficient datasets, so numerous images of specific pathologies have to be acquired. However, obtaining real data can be problematic and time-consuming, especially in case of rarely occurring pathologies.

Segmentation of images with clearly visible pathologies is performed in order to distinguish relevant regions of image

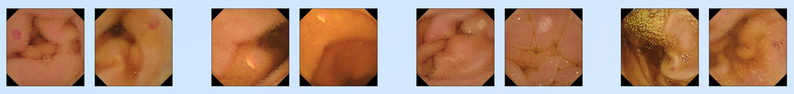


Intelligent algorithms of image blending allows us to generate necessary amounts of synthetic data of high quality for pathologies that we require

Appropriate regions must be selected from pictures without pathologies

This is where the use of synthetic data comes in handy - endoscopy is one of those branches of medicine, where the procedure of data augmentation proves to be advantageous.

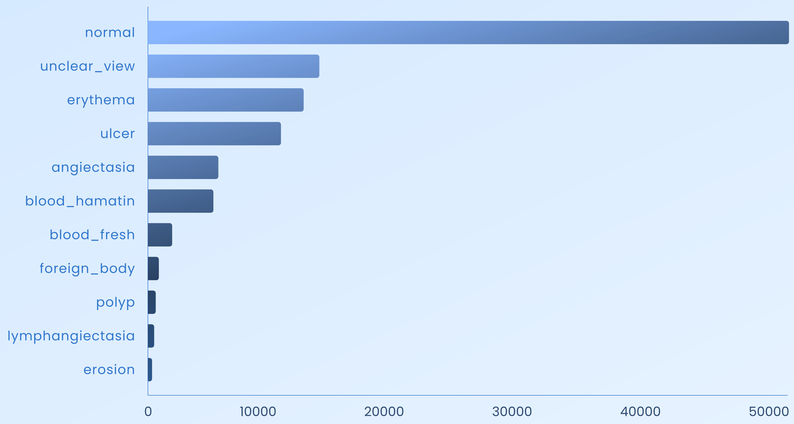
Sample images with synthetic pathologies



Angiectasia | Ulcer | Lymphangiectasia | Erosion

DATA LABELLING PIPELINE
WITH GOLDEN STANDARD

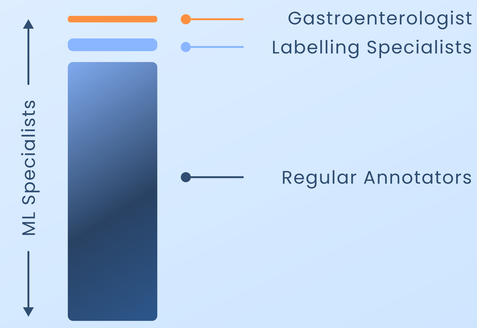
Number of images of each class in our dataset combined with the Kvasir Capsule dataset



TEAM

A structure is divided into three levels: regular annotators, labelling specialists - most experienced annotators and an expert from the gastroenterology field. Regular annotators label a dataset (addressing uneven distributions of classes) according to the manual. Annotators work in pairs to ensure cross checking ability for the set. A labelling specialist marks only part of each set to add another layer of validation. The experienced specialist takes part in meetings to resolve disputes and give a broad perspective of a field. The Labelling Team Lead - a person with a background in machine learning - takes care of datasets delivery, seamless flow between team members and includes final decisions in the manual.

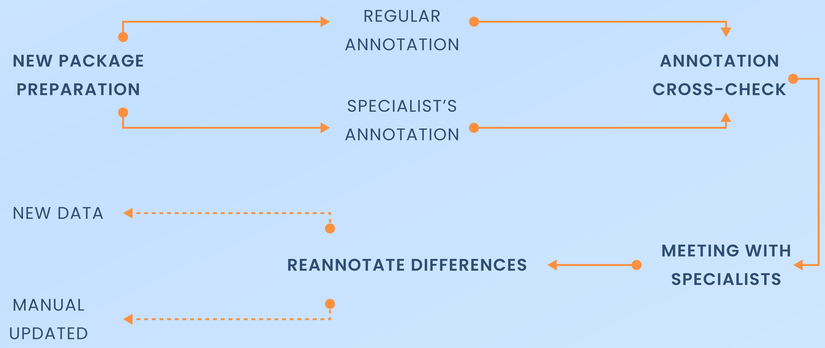
Visual representation of labelling actions' percentage by team members



It is worth mentioning that as a part of a pipeline, a multi-level training was designed: it is necessary to explain the difference between diagnosis and labelling as to minimise data incoherency and ensure smooth onboarding.

MANUAL

As a part of the golden standard, the annotation manual was created. It is a document with class descriptions - distinctive visual features, graphic scales, comparison to pathology measurements, rare pathologies and additional disambiguation for most troublesome labels e.g. types of bleeding.
Firstly, classes were adapted from the Kvasir Capsule dataset. However, provided labels were not sufficient: new classes were added based on Minimal Standard Terminology case range and annotators' feedback. The document is updated in case of important differences in annotations with an approval from an experienced gastroenterologist and a Machine Learning specialist.



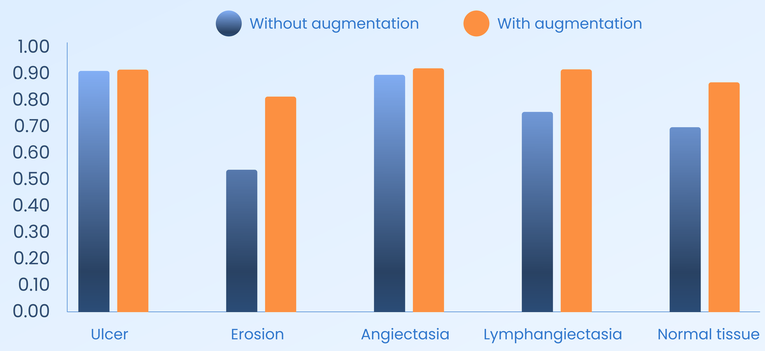
MEETINGS

Meetings with an experienced gastroenterologist and labelling team are held regularly. During a session, disputable images, labelled differently are displayed and commented on by a gastroenterologist and ML specialist. It joins the medical point of view and possible impact on a model's performance.

RESULTS

	Endoscopy class (capsule endoscopy images)	Classifier Perfomance (F1-score)	Accuracy
Without our data augmentation	Ulcer	91,2%	76,9%
	Erosion	54,0%	
	Angiectasia	89,9%	
	Lympangiectasia	75,7%	
	Normal tissue	70,0%	
Applying our data augmentation	Ulcer	91,4%	88,8%
	Erosion	81,7%	
	Angiectasia	92,0%	
	Lympangiectasia	91,7%	
	Normal tissue	86,8%	

Classifier Perfomance (F1-score) on datasets with and without data augmentation



The results prove our solutions useful - the performance of the classifier is better, when we apply our data augmentation method. As seen on the chart above, for every type of pathology the F-1 score is higher when the model is trained on datasets with data augmentation. The difference is especially apparent for erosion, lymphagiectasia and normal tissue - labels for which the results without data augmentation were the lowest. It should also be noted that overall accuracy for all classes increased from 76.9% without data augmentation to 88.8% after using synthetic data.

Pia H. Smedsrud, Vajira Thambawita, Steven A. Hicks, Henrik Gjestang, Oda Olsen Nedrejord, Espen Næss, Hanna Borgli, Debesh Jha, Tor Jan Derek Berstad, Sigrun L. Eskeland, Mathias Lux, Håvard Espeland, Andreas Petlund, Duc Tien Dang Nguyen, Enrique Garcia-Ceja, Dag Johansen, Peter T. Schmidt, Ervin Toth, Hugo L. Hammer, Thomas de Lange, Michael A. Riegler & Pål Halvorsen, "Kvasir-Capsule, a video capsule endoscopy dataset", Scientific Data 8, 142 (2021). https://doi.org/10.1038/s41597-021-00920-z

Aabakken L, Rembacken B, LeMoine O, Kuznetsov K, Rey JF, Rösch T, Eisen G, Cotton P, Fujino M. Minimal standard terminology for gastrointestinal endoscopy - MST 3.0. Endoscopy. 2009 Aug;41(8):727-8. doi: 10.1055/s-0029-1214949. Epub 2009 Aug 10. PMID: 19670144.