Galar - a large multi-label video capsule endoscopy dataset

Maxime Le Floch[1,2,†], Fabian Wolf[1,3,†], Lucian McIntyre[1,3], Christoph Weinert[4], Albrecht Palm[1], Konrad Volk[1], Paul Herzog[5], Sophie Helene Kirk[1,2], Jonas L. Steinhäuser[1,2], Catrein Stopp[1,2], Mark Enrik Geissler[2], Moritz Herzog[1,2], Stefan Sulk[2], Jakob Nikolas Kather[1,2,6], Alexander Meining[7], Alexander Hann[7], Jochen Hampe[1,2], Nora Herzog[1,†], and Franz Brinkmann[1,2,†]

[1]Else Kröner Fresenius Center for Digital Health, Technische Universität Dresden (TU Dresden), Dresden, Germany
[2]Department of Medicine I, University Hospital Dresden, Technische Universität Dresden (TU Dresden), Dresden, Germany
[3]Institute of Computer Science, Technische Universität Dresden (TU Dresden), Dresden, Germany
[4]Diakonissen Krankenhaus Dresden, Gastroenterology, Dresden, Germany
[5]Institute of Human Genetics, Ulm University and Ulm University Medical Center, Ulm, Germany
[6]Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany
[7]Interventional and Experimental Endoscopy (InExEn), Department of Internal Medicine II, University Hospital Würzburg, Würzburg, Germany
[†]Maxime Le Floch und Fabian Wolf contributed equally to this work
[†]Nora Herzog and Franz Brinkmann contributed equally to this work

ABSTRACT

Video capsule endoscopy (VCE) is an important technology with many advantages (non-invasive, representation of small bowel), but faces many limitations as well (time-consuming analysis, short battery lifetime, and poor image quality). Artificial intelligence (AI) holds potential to address every one of these challenges, however the progression of machine learning methods is limited by the avaibility of extensive data. We propose *Galar*, the most comprehensive dataset of VCE to date. *Galar* consists of 80 videos, culminating in 3,513,539 annotated frames covering functional, anatomical, and pathological aspects and introducing a selection of 29 distinct labels. The multisystem and multicenter VCE data from two centers in Saxony (Germany), was annotated framewise and cross-validated by five annotators. The vast scope of annotation and size of *Galar* make the dataset a valuable resource for the use of AI models in VCE, thereby facilitating research in diagnostic methods, patient care workflow, and the development of predictive analytics in the field.

Background & Summary

Video capsule endoscopy (VCE) is a minimally invasive gastroenterological imaging procedure used to capture video footage of a patient's digestive tract. This is especially relevant for the small intestine, which is not readily accessible through conventional endoscopic procedures like colonoscopy and esophagogastroduodenoscopy. However, this comes with limitations such as a time-consuming manual analysis[1], technical restrictions (e.g., battery runtime[2] or a lack of active locomotion), and heterogeneous image quality. In 16.5% of cases, the capsule does not pass through the ileocecal valve, resulting in incomplete small intestine examinations[3]. VCE is currently limited to a narrow range of use cases, primarily to detect internal bleeding[4,5]. However, the potential use cases for VCE are far broader; for each of the challenges outlined above, a solution utilizing Artificial intelligence (AI) methods is conceivable[6]. The technology may be used for improvement in screening and detection of pathologies such as polyps, inflammatory bowel disease, and celiac disease, and therefore enhances accuracy and efficiency of diagnosis[7].

The major drawback of VCE is the large amount of video footage generated, as medical staff are required to watch hours of recorded video. In these recordings, the section of interest is a tiny subset of the total video, and fluctuating image quality renders large parts unusable for diagnostic purposes. VCE, in combination with AI, has the potential to massively reduce the diagnostic evaluation time currently needed to interpret a rising amount of VCE footage. This could lead to VCE becoming a cost-efficient and widely used diagnostic method, as observed in other modalities, such as AI-assisted X-Ray evaluation[8].

The successful development of AI models requires substantial quantities of high-quality data, as well as precise and rigorous

annotations[9]. However, the availability of large datasets is scarce; the VCE-datasets thus far publicly available are either relatively small[10] or are limited to specific questions (e.g., quality, ulcers, bleeding, polyps, anatomy)[6,10,11]. To further drive the progress of AI in VCE, the creation of large, preprocessed, and annotated datasets is necessary[6,12]. Most academic research projects process their own data, which is tailored to their specific tasks[13,14] and do not make their datasets publicly available. The drawback of such an individualistic approach is that it necessitates a disproportionate amount of resources, limiting the progress of research. The existence of large, high-quality datasets could reduce the cost and effort involved in developing research for VCE and other medical technologies[15].

In this publication, we introduce a dataset that marks a significant advancement in the field of capsule endoscopic research. In the domain of VCE, two extensive public datasets already exist: *Kvasir*[10] (featuring 47,238 frames with 14 various anatomical and findings labels) and *Rhode Island*[11] (including a large number of annotated frames [5,247,588] on four anatomical organs). The *Galar* dataset offers a valuable contribution to the field, as it includes 29 distinct labels, incorporating a broad range of functional, anatomical, and pathological annotations while still providing 3,513,539 annotated frames.

Multidisciplinary and multicenter VCE research is needed for the clinical use of AI in patient diagnosis[6,16]. The *Galar* dataset consists of VCE data from two endoscopy centers in Germany, with two different capsule systems (Olympus™ Endocapsule 10 System, PillCam™ SB2, SB3, and Colon2 Capsule Endoscopy Systems[17,18]).

In summary, we provide a multicentric, multisystem dataset with high frame count and the most diverse and detailed annotations to date. These characteristics establish *Galar* as a robust resource for training machine learning models in video capsule endoscopy.

Methods

Videos were collected from the University Hospital Carl Gustav Carus (Dresden, Germany) and from an outpatient practice for gastroenterology (Dippoldiswalde, Germany). VCE recordings were obtained from August 2011 to March 2023 using the Olympus™ Endocapsule 10 System (Hamburg, Germany) as well as the PillCam™ SB2, SB3, and Colon Capsule Endoscopy Systems (Meerbusch, Germany)[17,18]. The videos were initially generated in proprietary data formats and were converted to the Moving Picture Experts Group (MPEG) format. The video resolution ranged from 336 x 336 pixels (Olympus™) to 576 x 576 pixels (PillCam™). Out of the 449 recordings, 80 videos were pre-selected for annotation based on the related findings by selecting only pathological videos for annotation. To de-identify VCE recordings, randomly generated study IDs were assigned, and the videos were cut. Afterwards, videos were transferred to university servers. There each video in the dataset was labeled framewise, resulting in 3,513,539 labeled frames.

For this study, no explicit patient consent was required for data sharing. The data collected were anonymized and retrospectively gathered, falling under the privilege of self-research as per Section 34, Paragraph 1 of the Saxon Hospital Act (SächsKHG). This provision allows data processing without explicit patient consent, as long as patients are informed during their treatment about the potential use of their anonymized data for research purposes, in compliance with Articles 13 and 14 of the EU General Data Protection Regulation (GDPR). The ethics review and approval for the entirety of this study were conducted by the Ethics Committee of the University Hospital Carl Gustav Carus at the Technical University of Dresden (TU Dresden), Germany. The ethics committee, formally known as the Ethik-Kommission des Universitätsklinikums Carl Gustav Carus an der Technischen Universität Dresden, granted approval on December 16, 2022 (Ethics ID: BO-EK-534122022), confirming that the study adheres to the ethical principles of the Declaration of Helsinki. This approval covers all ethical aspects of the study, including the waiver of consent for data sharing. Additionally, a separate privacy review by the Data Protection Officer of the University Hospital Carl Gustav Carus, Dresden, Germany confirmed that the anonymized data would be processed in accordance with EU GDPR regulations and national privacy laws.

Data preparation

Computer vision annotation tool (CVAT)[19] is a web-based, open-source image- and video annotation tool. Using CVAT, five annotators (a team of experienced gastroenterologists and trained medical students) labelled the data. The labels were categorized into three main groups: The **technical** group consists of labels concerning the image quality, where *good view* indicates a reduction of the view by less than 50%, *reduced view* indicates a reduction of the view by over 50%, and *no view* indicates a reduction of the view by over 95%. Furthermore, a distinction is made between *bubbles* and *dirt* as factors contributing to the degradation of image quality. The **anatomical** group consists of typical landmarks: *z-line, pylorus, papilla of Vater, ileocecal valve* and the different sections of the gastrointestinal tract: *mouth, esophagus, stomach, small intestine, colon*. The final group is the **pathological** group, which consists of the most frequent pathologies found in VCE and some less frequent findings: *ulcer, polyp, active bleeding, blood, erythema, erosion, angiectasia, inflammatory bowel disease (IBD), foreign body, esophagitis, varices, hematin, celiac, cancer, lymphangiectasis*. The pathologies *esophagitis, varices and celiac* did not occur in any of the videos. Figure 3 displays an overview of the number of annotated frames per label.

Annotation Process

An early decision was made to label each frame in the dataset individually, with the annotation process occurring in multiple stages. From each of the videos, every unique frame was extracted using Python (v3.9.8)[20] and FFMPEG (v4.0.6)[21]. Frames originating from the PillCam™ capsule system were cropped to remove black borders. A timestamp, visible in the top right corner, was also removed. No further pre-processing was done for the videos from the Olympus™ capsule system. Subsequently, the frames were uploaded to CVAT, where frames were annotated by our team. Frames containing unrecognizable features were given the label *unknown*. Then, all frames labeled with a pathology were cross-validated with the confirmation of a secondary annotator. Any frames still possessing the *unknown* label were reviewed by a gastroenterologist with 10 years of experience in endoscopy and were relabeled accordingly.

Data Records

The *Galar* VCE dataset can be found in the open access repository *figshare*[22]. It consists of 3,513,539 frames, each labeled with 29 labels, and has a total size of $\sim$ 580GB. A detailed overview of the structure of the dataset is shown in Figure 2. Each video in the dataset was labeled framewise. The dataset contains the folders *Frames* and *Labels*. The *Labels* folder contains CSV files, where each file has a header starting with the *index* column, followed by the columns of the 29 possible labels described in Data preparation and ending with the *frame* column, which refers to the corresponding frame the labels belong to. The *Frames* folder consists of 80 sub-folders, each containing the frames associated with a study. Table 1 shows the number of videos, resolution, and distribution of frames per capsule system. Additionally, a metadata file is provided, containing patient age, gender, and capsule system used.

Six videos contain technical annotations: *5, 8, 9, 13, 14, 22*. A total of 35,733 frames were annotated with this label category. The creation of technical annotations was found to be more resource intensive compared to the other categories, as the visibility is more volatile and prone to sudden change. As the category is highly relevant for machine learning (ML) applications in VCE, the labels were included for completeness.

Technical Validation

The dataset was used to train multiple ResNet-50 models[23]. The data was split into a set for training and validation consisting of 60 videos, and a test set comprised of the remaining 20 videos. K-fold cross-validation was performed on the data in the training and validation set. The videos from the test set all originate from the Dippoldiswalde practice and there is no overlap of these videos with those from the train set.

The labels *dirt* and *bubbles* form a multi-label classification problem, while the labels *good view*, *reduced view*, and *no view* as well as the section labels *mouth, esophagus, stomach, small intestine*, and *colon* require multi-class classification. Some of the other more frequently occurring pathological (e.g., *blood* or *polyp*) labels are trained on separately.

For the classification of the multi-label and the multi-class models, 5-fold cross-validation was employed. The binary classification of the pathologic labels was done using 2-fold cross-validation, as some labels were not contained in a sufficient number of videos. To ensure that the frames of one patient are not spread over the training and test set and to get the best possible distribution of the labels over the folds, sklearn's StratifiedGroupKFold[24] method was applied.

The ResNet-50 model pre-trained on ImageNet[25] was fine-tuned for 10 epochs using PyTorch (v2.0.1)[26]. Following this, fine-tuning was done for each of the target tasks. These models were trained over 100 epochs (with early stopping), with a 128 Batch size and a 0.001 learning rate. For each image, a Resize transform was applied, to scale the image down to 224x224. Additionally, the transforms ShiftScaleRotate, RGBShift, GaussNoise and RandomBrightnessContrast were each applied with a 30% likelihood to each image. The small subset of images which contain the technical annotation were fine-tuned similarly, excepting the epochs, which were capped at 50 with early stopping.

As measurements for the classification performance, the F-1 score, the Area under the Receiver Operating Characteristic Curve (AUROC) as well as the accuracy were calculated using the TorchMetrics (v1.0.3)[27] Python library.

Tables 2, 3, 4, and 5 show results for the classification models. The model fine-tuned for *dirt* and *bubbles*, along with the two multi-class models, performed decently with accuracy value up to 93% for the labels *stomach* and 92% for *small intestine*.

The binary models for pathological labels encountered challenges to accurately identify positive samples. To improve performance, weighted sampling as well as weighted loss was explored. For weighted sampling, the probability for an image to be sampled was based on the occurrence of its class, as a fraction of the total dataset. For the more complex multi-label problem, each unique combination of labels was assigned a weight, again as a portion of the total dataset. This made the weights dynamic, based on the target the model is trained on.

Although these strategies helped to improve performance on some labels, other required heavy parameter optimization. This underscores the difficulty and necessity of improving and developing AI methods to address the challenges of imbalanced

label distrubution and multi-source data. Consequently, it highlights the importance of a multicentric, multisystem dataset with extensive annotations of pathologies.

Usage Notes

With *Galar* we provide the largest public VCE dataset, both in terms of the number of features labeled per image and the total number of annotated images. The large number of ground truth labeled images allows for supervised training of machine learning models and is a significant contribution to the landscape of publicly available VCE datasets.

If the dataset is to be employed for machine learning applications, it is essential to carefully partition the data into training and validation sets. The comparative rarity of select labels, especially over others in the same class, must be respected. Additionally, the data originates from two different VCE systems and was collected at two different study sites. Patients of varying age and gender are also present in the dataset. This information must be considered when generating splits. The metadata file, found in the figshare repository, provides information regarding the capsule system and patient age and gender, per individual study.

The dataset is provided compressed, in the 7-Zip (.7z) format. The data must be uncompressed before it may be viewed and modified; common operating systems (Windows, Linux, MacOs) by default provide archive utility which enables this.

By licensing the dataset under a Creative Commons Attribution 4.0 International (CC BY 4.0) License which allows sharing, copying, and redistribution, as well as adaptation and transformation, we hope to advance research in the field. For more details about Creative Commons licensing, please refer to https://creativecommons.org.

Code availability

The code employed for the technical validation can be accessed via our public GitHub Repository: https://github.com/EKFZ-AI-Endoscopy/GalarCapsuleML. The repository contains a full guide on running the code, tuning hyperparameters, and generating statistics.

References

1. Ahmed, M. Video Capsule Endoscopy in Gastroenterology. *Gastroenterol. Res.* **15**, 47–55, 10.14740/gr1487 (2022).
2. Kwack, W. & Lim, Y. J. Current Status and Research into Overcoming Limitations of Capsule Endoscopy. *Clin. endoscopy* **49**, 8–15, 10.5946/ce.2016.49.1.8 (2016).
3. Liao, Z., Gao, R., Xu, C. & Li, Z.-S. Indications and detection, completion, and retention rates of small-bowel capsule endoscopy: a systematic review. *Gastrointest. Endosc.* **71**, 280–286, 10.1016/j.gie.2009.09.031 (2010).
4. Goenka, M. K., Majumder, S. & Goenka, U. Capsule endoscopy: Present status and future expectation. *World J. Gastroenterol.* **20**, 10024–10037, 10.3748/wjg.v20.i29.10024 (2014). Publisher: Baishideng Publishing Group Inc.
5. Iddan, G., Meron, G., Glukhovsky, A. & Swain, P. Wireless capsule endoscopy. *Nature* **405**, 417–417, 10.1038/35013140 (2000). Number: 6785 Publisher: Nature Publishing Group.
6. Iakovidis, D. K. & Koulaouzidis, A. Software for enhanced video capsule endoscopy: challenges for essential progress. *Nat. Rev. Gastroenterol. & Hepatol.* **12**, 172–186, 10.1038/nrgastro.2015.13 (2015). Number: 3 Publisher: Nature Publishing Group.
7. Mustafa, B. F., Samaan, M., Langmead, L. & Khasraw, M. Small bowel video capsule endoscopy: an overview. *Expert. Rev. Gastroenterol. & Hepatol.* **7**, 323–329, 10.1586/egh.13.20 (2013).
8. Mun, S. K., Wong, K. H., Lo, S.-C. B., Li, Y. & Bayarsaikhan, S. Artificial Intelligence for the Future Radiology Diagnostic Service. *Front. Mol. Biosci.* **7** (2021).
9. Wang, S. *et al.* Annotation-efficient deep learning for automatic medical image segmentation. *Nat. Commun.* **12**, 5915, 10.1038/s41467-021-26216-9 (2021).
10. Smedsrud, P. H. *et al.* Kvasir-Capsule, a video capsule endoscopy dataset. *Sci. Data* **8**, 142, 10.1038/s41597-021-00920-z (2021). Number: 1 Publisher: Nature Publishing Group.
11. Charoen, A. *et al.* Rhode Island gastroenterology video capsule endoscopy data set. *Sci. Data* **9**, 602, 10.1038/s41597-022-01726-3 (2022). Number: 1 Publisher: Nature Publishing Group.
12. Park, J. *et al.* Recent Development of Computer Vision Technology to Improve Capsule Endoscopy. *Clin. Endosc.* **52**, 328–333, 10.5946/ce.2018.172 (2019). Publisher: Korean Society of Gastrointestinal Endoscopy.

13. Hwang, Y. *et al.* Improved classification and localization approach to small bowel capsule endoscopy using convolutional neural network. *Dig. Endosc.* **33**, 598–607, 10.1111/den.13787 (2021). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/den.13787.

14. Mascarenhas Saraiva, M. *et al.* Artificial Intelligence and Capsule Endoscopy: Automatic Detection of Small Bowel Blood Content Using a Convolutional Neural Network. *GE - Portuguese J. Gastroenterol.* **29**, 331–338, 10.1159/000518901 (2021).

15. Zhang, A., Xing, L., Zou, J. & Wu, J. C. Shifting machine learning for healthcare from development to deployment and from models to data. *Nat. Biomed. Eng.* **6**, 1330–1345, 10.1038/s41551-022-00898-y (2022). Number: 12 Publisher: Nature Publishing Group.

16. Yang, Y. J. The Future of Capsule Endoscopy: The Role of Artificial Intelligence and Other Technical Advancements. *Clin. Endosc.* **53**, 387–394, 10.5946/ce.2020.133 (2020).

17. PillCam™ SB 3 Capsule | Medtronic (UK), https://www.medtronic.com/covidien/en-gb/products/capsule-endoscopy/pillcam-capsules/pillcam-sb-3-capsule.html.

18. Kapselendoskopie - Gastroenterologie - Olympus Medizintechnik, https://www.olympus.de/medical/de/Produkte-und-L%C3%B6sungen/Produkte/Gastroenterology/Kapselendoskopie.html.

19. CVAT, https://www.cvat.ai/.

20. Python, https://www.python.org/.

21. Ffmpeg, https://ffmpeg.org/.

22. Galar - a large multi-label video capsule endoscopy dataset. *figshare* https://doi.org/10.25452/figshare.plus.25304616 (2024).

23. Xu, W., Fu, Y.-L. & Zhu, D. ResNet and its application to medical image processing: Research progress and challenges. *Comput. Methods Programs Biomed.* **240**, 107660, 10.1016/j.cmpb.2023.107660 (2023).

24. sklearn's stratifiedgroupkfold, https://scikit-learn/stable/modules/generated/sklearn.model_selection.StratifiedGroupKFold.html.

25. ImageNet, https://www.image-net.org/.

26. PyTorch, https://pytorch.org/.

27. Detlefsen, N. S. *et al.* TorchMetrics - Measuring Reproducibility in PyTorch. *J. Open Source Softw.* **7**, 4101, 10.21105/joss.04101 (2022).

28. Handa, P., Gunjan, D. D., Goel, P. N. & Indu, P. S. AI-KODA Dataset: An AI-Image Dataset for Automatic Assessment of Cleanliness in Video Capsule Endoscopy as per Korea-Canada Scores. 10.6084/m9.figshare.25807915.v1 (2024).

29. Thakur, A., Handa, P., Goel, N. & Gunjan, D. Vce-anomalynet: A new dataset fueling ai precision in anomaly detection for video capsule endoscopy , 10.22541/au.171387106.63353485/v1 (2024).

30. Cychnerski, J., Dziubich, T. & Brzeski, A. Ers: a novel comprehensive endoscopy image dataset for machine learning, compliant with the mst 3.0 specification (2022). 2201.08746.

31. Akihito, Y. *et al.* The see-ai project dataset, 10.34740/KAGGLE/DS/1516536 (2022).

Author contributions statement

M.L.F.: methodology, investigation, data curation, writing original draft, visualization. F.W.: methodology, software, technical validation, writing original draft, visualization. L.M.: methodology, software, data curation, writing original draft, visualization. C.W.: methodology, data annotation. A.P.: methodology, data annotation. K.V.: methodology, data annotation. P.H.: data curation, investigation. S.H.K.: methodology, investigation, clinical supervision. J.L.S.: methodology, investigation, review and editing. C.S: data curation. M.E.G: methodology, data curation, data annotation. M.H.: review and editing. S.S: data curation, technical and clinical supervision. A.M.: methodology, review and editing. A.H.: methodology, review and editing. J.N.K: review and editing, supervision. J.H: conceptualization, review and editing, supervision, N.H: conceptualization, investigation, data curation, writing original draft, review, and editing, visualization, supervision, project administration, funding acquisition. F.B.: conceptualization, investigation, data curation, review, and editing, visualization, supervision, project administration, funding acquisition,

This research was funded by the German Federal Ministry of Education and Research (BMBF) within the project SEMECO-A4 (03ZU1210HB).

Competing interests

The authors declare no competing interests.

Figures & Tables

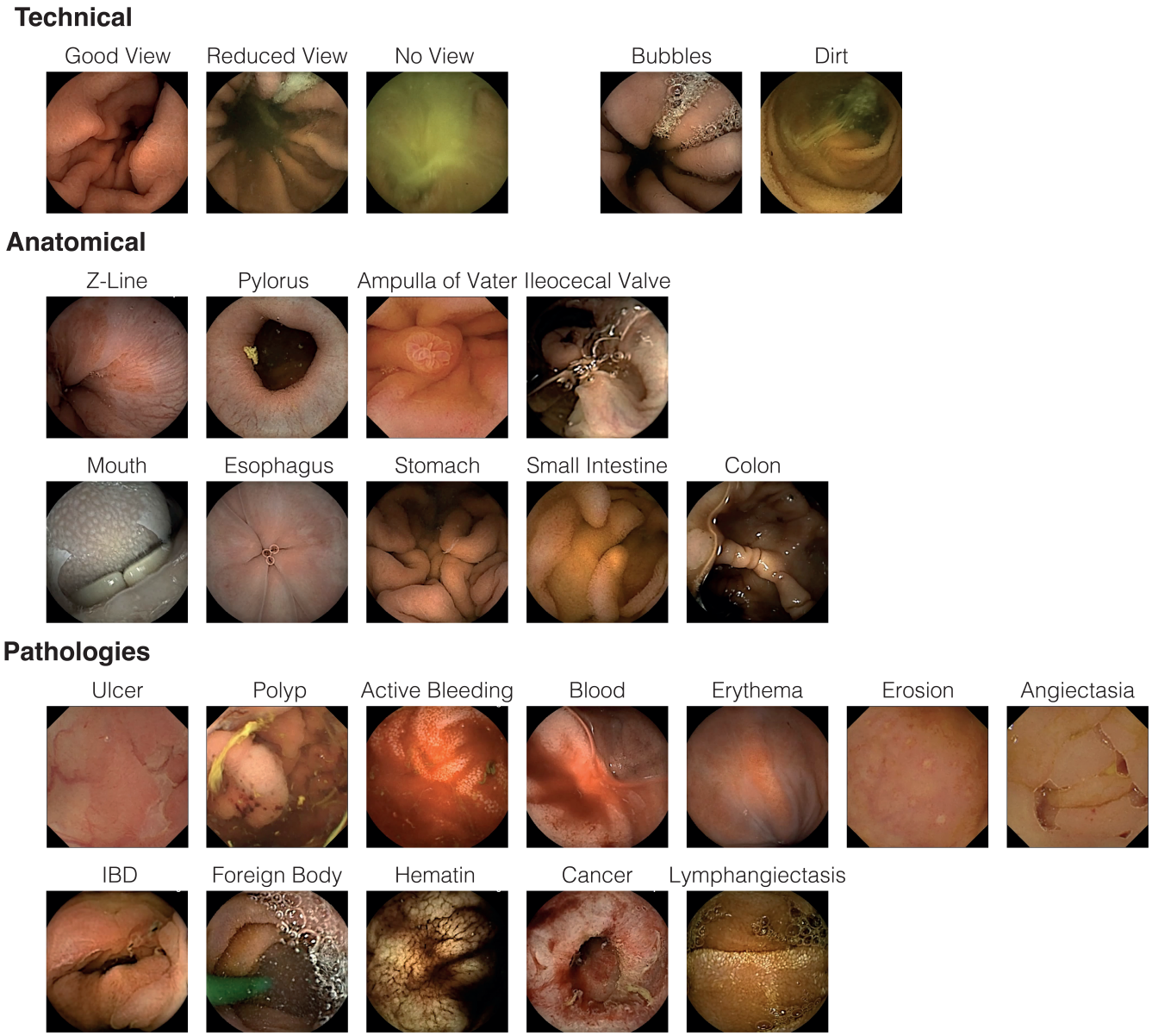

Figure 1. Example images of the 26 labels in the dataset. The figure does not contain images of the labels *esophagitis*, *varices* and *celiac*, as there were no instances of these pathologies present in the set of VCE studies.

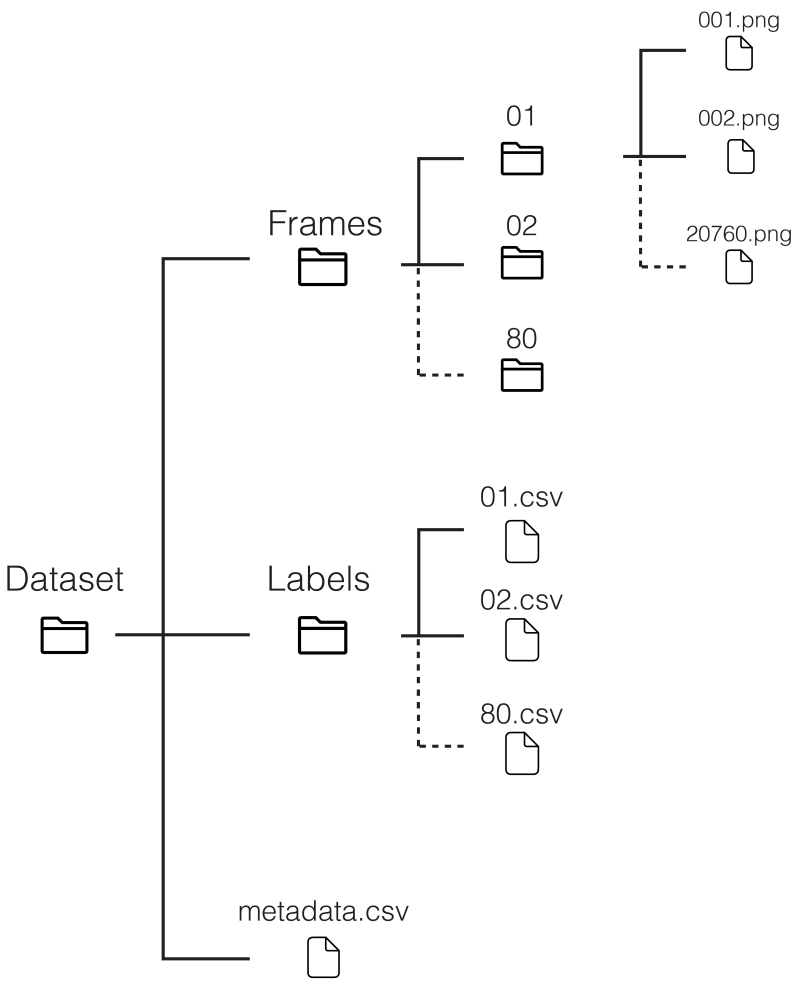

Figure 2. The file structure of the *Galar* dataset. Frames are stored chronologically in subfolders of the *Frames* folder. Labels are stored in a single CSV file, per study. The metadata file further contains data on a per study basis.

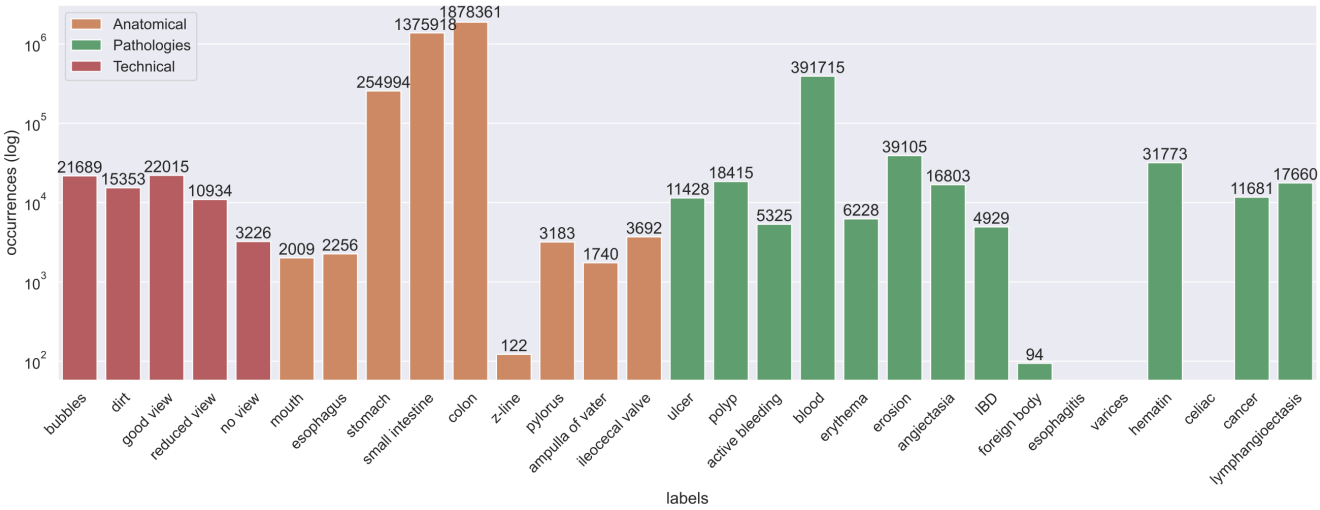

Figure 3. Overall frames per label count of the *Galar* dataset. Image occurrences per labels are displayed across the three main groups (technical, sections and anatomical). The y-axis is scaled logarithmically.

Capsule System	Resolution (Pixels)	No. of frames	No. of videos
PillCam™ University Hospital Dresden	512x512	528,470	38
Olympus™ University Hospital Dresden	336x336	2,750,514	22
PillCam™ Dippoldiswalde	512x512	234,555	20

Table 1. Overview of the data records in the *Galar* dataset. Description of the resolution, number of frames, and number of videos per capsule system.

Label	F1	AUROC	Accuracy
bubbles	0.89	0.87	0.82
dirt	0.88	0.94	0.88
macro average	0.88	0.90	0.85
micro average	0.89	-	0.85

Table 2. Classification results for a ResNet-50 fine-tuned on *bubbles* and *dirt*. The metrics were computed individually for each label, and both macro- and micro-averaged scores are calculated across all labels. The outcomes are averaged across the 5 cross-validation folds.

Label	F1	AUROC	Accuracy
good view	0.88	0.88	0.87
reduced view	0.41	0.85	0.47
no view	0.33	0.91	0.29
macro average	0.54	0.88	0.54
micro average	0.79	-	0.79

Table 3. Classification results for a ResNet-50 fine-tuned on *good view*, *reduced view*, and *no view*. The metrics are computed individually for each label, and both macro- and micro-averaged scores are calculated across all labels. The outcomes are averaged across the 5 cross-validation folds.

Label	F1	AUROC	Accuracy
mouth	0.42	1.00	0.75
esophagus	0.65	1.00	0.73
stomach	0.78	0.93	0.93
small intestine	0.93	0.95	0.92
colon	0.75	0.96	0.72
macro average	0.71	0.96	0.81
micro average	0.89	-	0.89

Table 4. Classification results for a ResNet-50 fine-tuned on *mouth*, *esophagus*, *stomach*, *small intestine*, and *colon*. The metrics are computed individually for each label, and both macro- and micro-averaged scores are calculated across all labels. The outcomes are averaged across the 5 cross-validation folds.

Label	F1	AUROC	Accuracy
blood	0.14	0.87	0.98
pylorus	0.01	0.74	0.99
z-line	0.02	0.87	1.00
ulcer	0.00	0.44	0.99
polyp	0.05	0.73	1.00
erythema	0.02	0.70	1.00

Table 5. Classification results for multiple ResNet-50 models, each fine-tuned, on points of interest (e.g. *blood*). The metrics are computed individually for each label. The outcomes are averaged across the 2 cross-validation folds.

Dataset	Size	No. of Labels
Kvasir-Capsule[10]	47,238 labelled images	3 Landmarks, 2 Technical, 9 Pathological
Rhode Island[11]	5,247,588 labelled images	4 Sections
AI-KODA[28]	2,173 labelled images from 28 patients	8 Technical
VCE-AnomalyNet[29]	108,832 labelled images	17 Pathological
ERS Dataset[30]	5,970 labelled images (982,041 imprecisely labelled *)	123 labels
SEE-AI[31]	18,481 labelled images	12 Total

Table 6. Overview of openly accessible VCE datasets. *Imprecisely labelled images inherit labels from those that are labelled by an expert, and where the image appears chronologically close.