Informative Classification of Capsule Endoscopy Videos Using Active Learning

Filipe Fonseca[1(✉)], Beatriz Nunes[2], Marta Salgado[4], Augusto Silva[2], and António Cunha[1,3]

[1] Universidade de Trás-os-Montes e Alto Douro, Vila Real, Portugal
fmiguelof@gmail.com
[2] Universidade de Aveiro, Aveiro, Portugal
[3] INESC TEC, Porto, Portugal
[4] Centro Hospitalar Universitário de Santo António, Porto, Portugal

Abstract. The wireless capsule endoscopy is a non-invasive imaging method that allows observation of the inner lumen of the small intestine, but with the cost of a longer duration to process its resulting videos. Therefore, the scientific community has developed several machine learning strategies to help reduce that duration. Such strategies are typically trained and evaluated on small sets of images, ultimately not proving to be efficient when applied to full videos. Labelling full Capsule Endoscopy videos requires significant effort, leading to a lack of data on this medical area. Active learning strategies allow intelligent selection of datasets from a vast set of unlabelled data, maximizing learning and reducing annotation costs. In this experiment, we have explored active learning methods to reduce capsule endoscopy videos' annotation effort by compiling smaller datasets capable of representing their content.

Keywords: Active learning · Deep learning · Capsule endoscopy

1 Introduction

1.1 Context

The gastrointestinal tract is a carefully studied organ system since it can harbour severe conditions [1]. Several studies in the gastrointestinal tract showed that early detection of abnormalities is correlated with a significant decrease in dangerous developments and improvement in survival rates [2]. Upper and lower gastrointestinal endoscopy allow the visualisation of a substantial portion of the gastrointestinal tract, however, the small bowel was seen as the "black box" due to its hard reach for screening [3]. With the small bowel being home to multiple pathologies [3,4], there is a clear need for taking advantage of novel screening tests for the detection of small bowel diseases and its abnormalities.

A. Cunha et al. (Eds.): MobiHealth 2023, LNICST 578, pp. 307–323, 2024.
https://doi.org/10.1007/978-3-031-60665-6_23

The Wireless Capsule Endoscopy (WCE), is a pill-like, non-invasive camera that can be swallowed by the patient to be propelled through the digestive system taking advantage of its peristaltic movements, to provide an inner screening of the tract. In the course of its route, the capsule transmits images through radio frequency to a portable device attached to the patient's body [5]. The device can record for about 8 h, producing around 60 000 images. It is considered the preferred method for the diagnosis of diseases of the small bowel, given its ability to cover the whole gastrointestinal tract and its distinguished inner imaging results compared to other methods for abnormalities visualisation [6]. Despite this, the manual analysis of the resulting eight-hour-long recorded video by an expert is very time-consuming and susceptible to human error, since identifying lesions in this type of image is a very challenging task because the quality of the acquired images is not always the best, which can imply the loss of important information. Besides this, different types of lesions can be very similar, and the same type of lesion can present different colouration and shape, which makes the process of discriminating the lesions very demanding [5].

With this in mind, computer-aided diagnosis tools have been developed for automatic lesion detection in WCE imaging. However, these attempts do not perform well enough to be used in a clinical environment and have focused much more on vascular lesions than others. Therefore, there is a clear need to develop more robust approaches for the automatic analysis of Capsule Endoscopy Videos (CEV). The automatic detection of lesions on frames can drastically reduce the number of images that a physician has to analyse, allowing him to focus his attention on the relevant data only.

1.2 Motivation

The early detection of abnormalities can lead to significant improvement in the patient's health condition and avoid complications [7]. Diagnosing intestinal motility disorders involves assessing instructive frames, which can range from a few to a large number of images per CEV. Uninformative frames, which are obstructed by gastrointestinal contents like partially digested meals, intestinal secretions, or gas bubbles, are crucial for reducing video analysis time and indicating intestinal dysfunctions. Identifying uninformative frames helps reduce analysis time and helps identify intestinal material in a significant proportion.

Machine learning methodologies are used to construct binary classifiers for uninformative frames search in CEV processing. These classifiers are aided by a training set that accurately represents the underlying data population. However, challenges arise due to significant variations in colour among uninformative frames across different videos [8]. The labelling process may require human annotation of up to 50 000 frames for each video, which may be considered a naive technique.

Active learning offers effective methods for interactive labelling, reducing human involvement. The algorithm can enhance accuracy by using a smaller number of labelled training samples, provided it can choose the data to learn from [9]. It can be effective in expanding a training set without human interaction.

1.3 Objectives

Due to the limited availability of datasets which cover the CEV reality and the meticulous effort required to review and annotate whole videos on a frame-by-frame basis, the primary objective of this study is to contribute to evaluating the efficacy of active learning strategies on fully unlabelled CEV. The strategies under investigation include selecting certain samples that accurately reflect the overall distribution of data in the videos. This approach aims to enhance the efficiency of classifiers by reducing the number of labelled samples required [10].

The study utilises a deep learning model that was evaluated and trained in a prior experiment [11] to establish an initial understanding of the content of CEV. After the investigation of active learning selection strategies on a set of unlabelled videos, the objective is to augment the proficiency of the trained model to classify informative and uninformative frames. This will facilitate the accomplishment of the initial phase out of three, in an ongoing work aimed at improving strategies that could be used in a clinical environment.

2 Methodology

The methodology for this experiment's Active Learning (AL) cycle is described in Fig. 1.

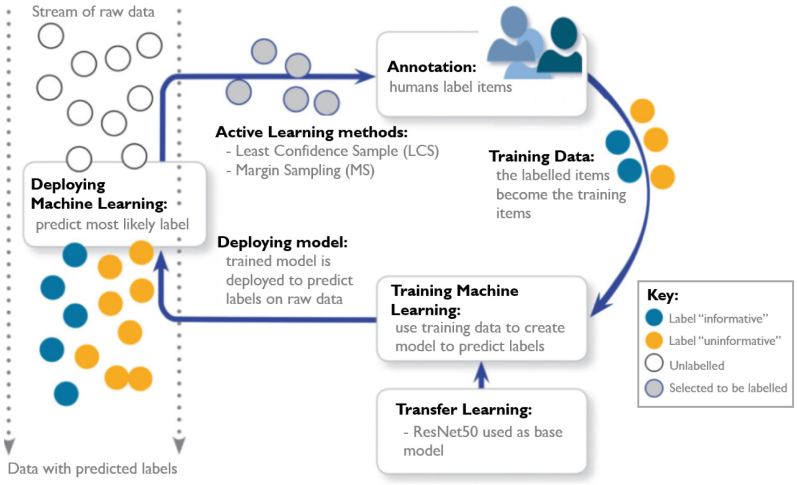

Fig. 1. The methodology used for this experiment.

Considering a base model, ResNet50, the AL cycle was initiated with the pre-trained model being introduced to new and raw data for predicting the most likely label, in our case either informative or uninformative. Then, using one AL method (either Least Confidence Sample or Margin Sample, in this experiment),

a sample of predicted labels was pooled to be annotated by an oracle. Finally, the labelled items become part of the training data for the model, and the model is retrained. Then, is tested in unlabelled CEV, for performance evaluation and given the results, a new AL cycle could be initiated.

2.1 Dataset

A private dataset was used in this experiment, a CEV dataset collected from examinations at *Centro Hospitalar Universitário de Santo António*. In total, the dataset consisted of 281 videos, belonging to 238 patients, for each video was extracted the contents of the small bowel lumen, averaging 20 000 frames per video. Only 50 videos had some labelled frames. The frames were stored using the PNG format. For the experiment depicted in this paper, were randomly selected 6 unlabelled videos, hereby named V1, V2, V5, V6, V7 and V10. Figure 2 and Fig. 3 depict example frames from the private dataset for each class to classify in this experiment.

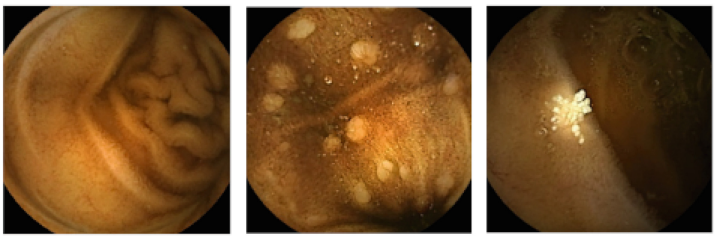

Fig. 2. Informative: Examples of CEV informative frames from the private dataset used in this experiment. From left to right: Normal Clean Mucosa; Nodular Lymphoid Hyperplasia and Xanthelasma.

Fig. 3. Uninformative: Examples of CEV uninformative frames from the private dataset used in this experiment.

2.2 Data Preparation and Feature Engineering

To evaluate the impact of active learning strategies on classifying informative and uninformative frames of CEV from unlabelled videos, we used a private dataset. Given the dataset and the random selection of videos, we have chosen V1 to be manually annotated (following the considerations described in Sect. 2.4) in order to train the ResNet50 model. For V1, 17 137 frames were manually divided into two folders, *informative* with 6 554 frames and *uninformative* with 10 583 frames. All frames were in a dimension of 320×320 pixels.

The data of V1 was then randomly divided into 2 unbalanced sets: 70% for training and 30% for validation. The dimension of all frames was adjusted to 224×224 pixels, to meet the input conditions of the model ResNet50.

2.3 Deep and Active Learning Strategies

Deep Learning: When it comes to machine learning issues, convolutional neural network models perform exceptionally well, especially when the issues include the categorization of dataset pictures [12]. Models must be sufficiently detailed to capture similarities between image samples, such as texture, colour, and shape features. While the model's initial layers gather these high-level features, later layers collect information that links those features to the outputs, allowing the model to learn to distinguish between them. Unfortunately, there are numerous instances where there is not enough data available to use this strategy. Transfer learning addresses this issue by leveraging models that have already been trained on a sizable dataset (often trained for a different task, using the same input but providing a different output). This is accomplished by previously capturing the relationships between the data's characteristics, which may later be applied to new challenges.

Given the efficacy of transfer learning in image classification, particularly in the medical field, a methodology utilising this method was used to classify informative and uninformative frames from CEV [13]. Given that the model has already learned about data patterns in the data that was used to pre-train it, transfer learning approaches use reusable features returned by some layers as input features to enable the training of a new model that only needs to learn the relations of those features for the new problem. Transfer learning provides the benefit of improved generalisation of the models, as they underlie the phenomena more than they model the data since the model has access to many types of data, in addition to having considerably fewer parameters to train.

For this experiment, following the previous work [11] and within state-of-the-art convolutional neural networks models, we have chosen the ResNet50 architecture as the base model for transfer learning. ResNet50 is a deep convolutional neural network that has been pre-trained on a large dataset (ImageNet) and has shown excellent performance in various computer vision tasks. By using ResNet50 as the base model, we can leverage its learned features for our specific problem, which saves time and computational resources compared to training a model from scratch. The model ResNet50 was trained using two different

approaches: only the last layer was updated, i.e. the pre-trained model worked as a feature extractor and only the weights of the classification layers were changed. The model was trained for 50 epochs, with a batch size of 200, and using the Adam optimisation algorithm which is an extension to the stochastic gradient descendent method that is based on adaptive estimation of the first-order and second-order moments [14]. The learning rate was set to 0.00001.

Active Learning: AL is a methodology that aims to maximize performance while minimizing the number of labelled examples required. It involves identifying the most valuable samples from unlabelled datasets and allocating them to an oracle, often a human annotator, for labelling. The goal is to minimise costs associated with labelling while maintaining performance. AL approaches can be divided into membership query synthesis, stream-based selective sampling, and pool-based AL:

- Membership query synthesis allows the learner to query the label of any unlabelled sample in the input space, even the one created by the learner [15];
- Stream-based selective sampling involves individual assessments regarding the necessity of querying the labels of unlabelled samples within the data stream, while pool-based sampling selects the optimal query sample by evaluating and ranking the entire dataset [16];
- The pool-based AL cycle involves the random selection of samples from the pool of unlabelled data, sending them to the oracle for labelling, and creating a labelled dataset. The model is trained using supervised learning techniques on the labelled dataset, and the acquired information determines the next sample to be queried. The training process is iterated until either the allocated label budget is depleted or the predetermined termination criteria are met [17].

AL differs from other approaches by using human or automated techniques to construct models with exceptional feature extraction capabilities. AL begins with datasets and designs complex query algorithms to carefully choose the most optimal samples from unlabelled datasets and request their labels. The optimisation of query criteria plays a pivotal role in determining the efficacy of AL techniques.

In this experiment, we have used two AL methods: Least Confidence Sampling and Margin Sampling, both belonging to the AL's Uncertainty Sampling technique:

- Least Confidence Sampling (LCS): This strategy selects the instances with the lowest predicted probability of belonging to the current class. It is effective when the model is uncertain about its predictions and wants to gain more confidence by labelling instances that it is least confident about.
- Margin Sampling (MS): This method is specially used when the model has some difficulties in distinguishing between multiple classes. MS is an AL

method that selects the two most confident images and calculates the uncertainty between them. Similar to LCS, these images are then presented to the oracle for annotation, added to the training set, and used to re-train the model. This approach is particularly useful when the model struggles with distinguishing between multiple classes, helping it improve its performance in such scenarios.

2.4 Annotation

In the context of CEV, the clear visualisation of internal tissue, organ lumen, or wall is sometimes hindered by the presence of intestinal juice. This fluid is often seen as a semi-opaque murky liquid, accompanied by bubbles and other artefacts related to the flux of different fluids inside the gastrointestinal tract. The presence of faecal matter, partially digested or residual foods, further complicates the problem of visibility when mixed with these secretions. Consequently, the accurate depiction of the gastrointestinal tract is impeded [18].

In this experiment, as we couldn't have an expert available to help with the annotation, it was decided that we, the technical team, would annotate the AL selected sets. This process, as well as the annotation of V1, followed the defined criteria that any frame of CEV used in this experiment would be an informative frame if it had at least 65% visible/unobstructed mucosa. A frame would be an uninformative frame if it had more than 35% of the frame obstructed.

Figure 2 and Fig. 4 depict frames from the private dataset that would be annotated as informative frames in this experiment. On the other hand, Fig. 3 and Fig. 5 depict frames that would be annotated as uninformative frames in this experiment.

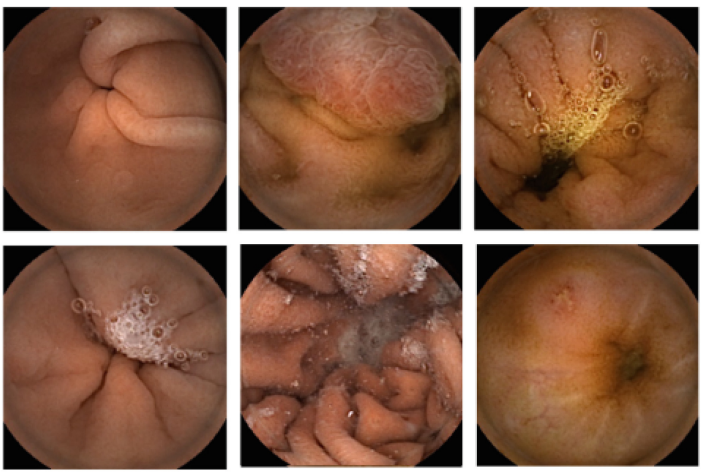

Fig. 4. Informative: Examples of CEV frames that would be annotated as informative frames from the private dataset used in this experiment.

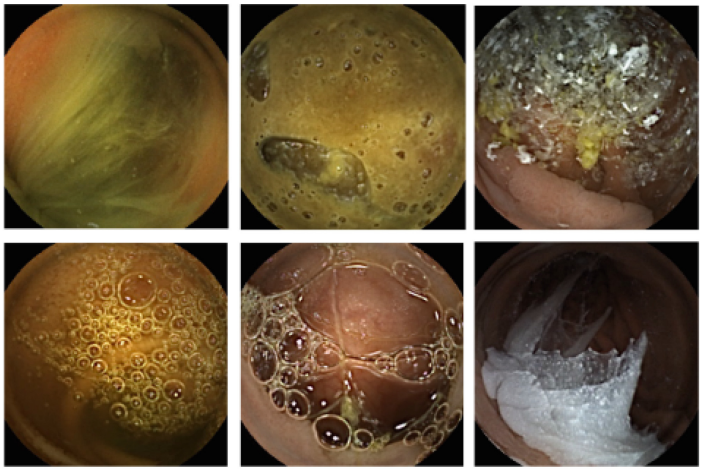

Fig. 5. Uninformative: Examples of CEV frames that would be annotated as uninformative frames from the private dataset used in this experiment.

2.5 Performance Evaluation

Considering that the efficiency was reflected in the classification distribution of true positives (TP), the correct identification of informative and uninformative frames in the set, compared with the number of false positives (FP) and false negatives (FN) classified. These values allowed the performance estimation using standard classification metrics like the Receiver Operating Characteristic (ROC) curve and the AUC (Area Under the Curve). The ROC curve sizes define the confidence of the model in its classification by plotting the true positive rate or recall (Eq. 1), versus the false positive rate (Eq. 2) at different classification thresholds. To compute the ROC, one would have to evaluate the model at different classification thresholds and plot the resulting curve. However, the AUC can express the confidence score of the model.

$$True\ Positive\ Rate = \frac{TP}{TP + FP} \tag{1}$$

$$False\ Positive\ Rate = \frac{FP}{FP + TN} \tag{2}$$

In addition, we have used loss and accuracy metrics as evaluative measures for gauging the efficacy of the ResNet50 model. The use of loss as the objective function allows for the quantification of the disparity between anticipated results and factual goal values during the training process. The aforementioned guidance plays a crucial role in facilitating the change of parameters in our model, compelling it to minimise mistakes and provide forecasts of higher accuracy. Simultaneously, Accuracy (Eq. 3) gives us a clear and simple view of how

effectively our model is identifying cases, exhibiting the proportion of properly categorised data points. Nevertheless, it is important to acknowledge that when dealing with imbalanced datasets, where the distribution of classes is unequal, relying solely on high Accuracy may not be sufficient as a comprehensive measure of performance. In such cases, it becomes necessary to incorporate additional metrics such as precision, recall, and F1-score to ensure a more precise evaluation. The combination of these measures provides a comprehensive evaluation of our model's capabilities and enables us to make well-informed judgements on its suitability for the designated task.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1\text{-}score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{6}$$

Finally, it is important to state that V10 was used to test the model throughout the steps of this experiment and served as the basis for the aforementioned metrics to be calculated. Given that the V10 had 19 097 images to be labelled, a coding strategy was implemented to randomly select 400 images to be our sample. Then those 400 images were manually labelled and became the representative sample that we used to evaluate the Accuracy of the model.

3 Results and Discussion

This section will describe the steps made throughout the experiment. The main focus of the steps was to evaluate the performance of selected AL techniques in the methodology described in Fig. 1 aimed to classify informative and uninformative frames of complete extracted videos of the small bowel lumen from full CEV.

Table 1 resumes the evolution of the model throughout the steps, as well as the impact of AL on the knowledge acquired by the model. In each step, except step 1, the model was trained and then tested in videos completely unlabelled.

The conclusions surrounding the V10 tests were drawn based on observation of the images that the model classified. The value of Accuracy will be used to prove the results that the team observed quantitatively. Given that relying solely on Accuracy may not be sufficient (as stated in Sect. 2.5), and to underline the test results, other training and validation metrics will be presented. The metrics also aim to verify that the model didn't go into overfitting and showed promising results.

Table 1. Overall view of the steps of this experiment.

# Steps	AL Methods	Data in model	Test Data	Accuracy
1	–	V1	–	–
2	–	V1	V10	0.42
3	LCS	V1, V2	V10	0.52
4	MS	V1, V2	V10	0.40
5	LCS	V1, V2, V5, V7	V10	0.54
6	LCS	V1, V2, V5, V7, V6	V10	0.41

Step 1: The model ResNet50 was trained with the V1 now labelled as stated in Sect. 2.2.

Table 2. The confusion matrix (TP, true positives; FN, false negatives; FP, false positives; TN, true negatives) and the metrics of the AUC and Loss are provided along with the Precision, Recall and F1-score for the model after Step 1.

	TP	FN	TN	FP	AUC	Loss	Precision	Recall	F1-score
Train	4717	2	7797	8	1.00	[0.0, 0.10]	1.00	1.00	1.00
Validation	1389	446	2774	4	0.98	[0.0, 0.70]	1.00	0.76	0.86

Through the analysis of the different metrics shown in Table 2 we can observe that the model already had very good results with the task in hand, which meant that no changes to the initial model were necessary. At this point, we were ready to test the model with V10 and see how efficient the model was at classifying frames into informative and uninformative, without applying any AL technique. Hence, in Step 1 there was no AL cycle.

Step 2: V10 was introduced to the model, after Step 1, to test its performance in classifying informative and uninformative frames.

It was observed, in approximately 848 images, that the model classified them inaccurately as informative frames.

Considering the method to calculate the Accuracy, 166 images were correctly predicted in the universe of 400 images, this meant an Accuracy of 0.415. With this Accuracy, we can generalise the results and conclude that the model can classify correctly approximately 42% of the images in CEV.

In Fig. 6, frames b) and c) were both considered informative and presented an equal prediction, even when we can see that frame b) is uninformative. At this point, the model hadn't yet learned that frames similar to b) weren't informative. We could see that the model, even with good results, still had room to learn.

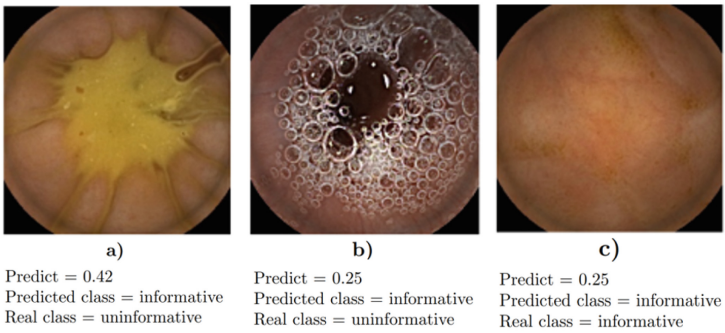

Fig. 6. Example of frames classified by the model in Step 2 as informative. The frames a) and b) were inaccurately classified as informative. The frame c) was correctly classified as informative.

Step 3: The introduction of the AL cycle to the model trained after Step 1, aimed to increase the model's knowledge with a big reduction in the effort of data annotation. The AL method chosen was Least Confidence Sampling (LCS).

V2 was the raw data for the model to predict labels and using LCS, 200 images were pooled, annotated and then inserted in the initial dataset and used to retrain the model. After re-training the model, now with the added images annotated from V2, the obtained results are stated in Table 3.

Table 3. The confusion matrix and the metrics of the AUC and Loss are provided along with the Precision, Recall and F1-score for the model in Step 3.

	TP	FN	TN	FP	AUC	Loss	Precision	Recall	F1-score
Train	4725	0	7758	241	1.00	[0.0, 0.10]	0.95	1.00	0.98
Validation	1799	36	2716	62	0.99	[0.0, 0.50]	0.97	0.98	0.97

Compared to the results of Step 1 (see Table 2), the number of FN in training and validation decreased but the number of FP increased.

V10 was again used to test the model and around 3828 images were considered as informative. After checking them, it was verified that there is still a large number of images that were inaccurately classified.

The Accuracy at this step was 0.52.

Taking into consideration Fig. 7, we could see that images similar to a) and b) were still wrongly classified as informative. It can also be noted that frame a) presented a prediction very similar to the prediction in frame c) which is an informative frame.

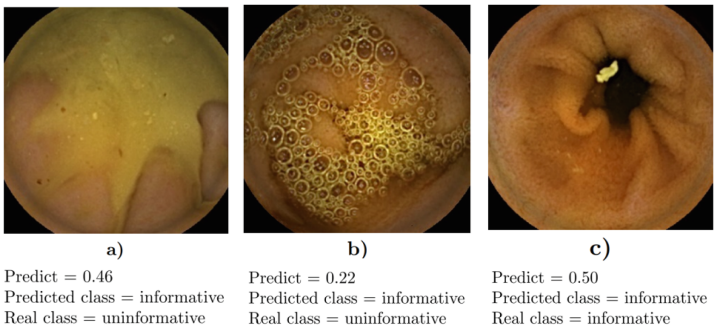

Fig. 7. Example of frames classified by the model in Step 3 as informative. The frames a) and b) were inaccurately classified as informative. The frame c) was correctly classified as informative.

Step 4: The initial conditions of Step 3 were repeated but the AL method LCS was replaced with MS, in order to assess whether another AL method was more suitable for the task that we were dealing with.

With V2 as the raw data for the model to predict labels and using MS, 200 images were pooled, annotated and then inserted in the initial dataset. After re-training the model, now with the added images annotated from V2, the obtained results are stated in Table 4.

Table 4. The confusion matrix and the metrics of the AUC and Loss are provided along with the Precision, Recall and F1-score for the model in Step 4.

	TP	FN	TN	FP	AUC	Loss	Precision	Recall	F1-score
Train	4213	513	7991	7	0.99	[0.0, 0.10]	1.00	0.89	0.94
Validation	855	980	2777	1	0.98	[0.0, 1.00]	1.00	0.47	0.64

The results obtained with MS and compared with the results from LCS, it was possible to conclude that the results with LCS presented to be better, regarding AUC, Loss and the total number of false predictions. Even with the worst results after re-training the model, we still used V10 to test the model and verified that only 319 images were classified as informative.

The model, using MS, had an Accuracy of 0.397.

Considering Fig. 8, as in the previous step with LCS, we could see that images similar to frames a) and b) were still wrongly classified as informative and images similar to frame a) presented predictions very similar to the prediction of frame c) which is an informative frame.

We concluded that with MS the model had an overall regression when compared to the results of Step 3 with LCS.

Given the aforementioned information, LCS was the AL method used for the following steps, until the end of the experiment.

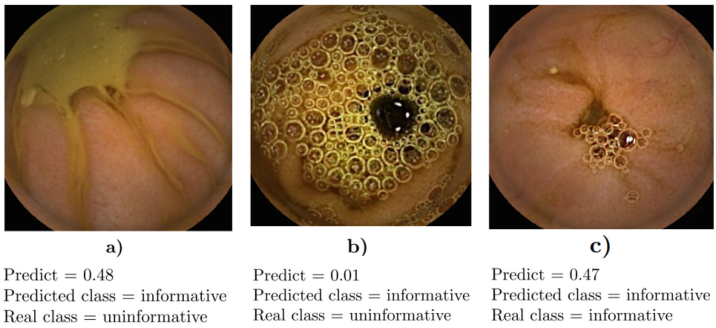

Fig. 8. Example of frames classified by the model in Step 4 as informative. The frames a) and b) were inaccurately classified as informative. The frame c) was correctly classified as informative

Step 5: The main goal of this step, was to evaluate the performance of the model's classification capability, after 3 iterations of the AL cycle.

Three iterations of the AL cycle over the trained model after Step 1, were triggered using V2, V5 and V7 as the raw data for the model to predict labels. Using LCS, in each iteration 200 images were pooled, annotated and then inserted in the initial dataset, used to train the model. After re-training the model, now with the added images annotated from V2, V5 and V7, the obtained results are stated in Table 5.

Table 5. The confusion matrix and the metrics of the AUC and Loss are provided along with the Precision, Recall and F1-score for the model in Step 5.

	TP	FN	TN	FP	AUC	Loss	Precision	Recall	F1-score
Train	4885	14	5451	2774	0.99	[0.0, 0.05]	0.64	1.00	0.78
Validation	1791	44	1822	956	0.97	[0.0, 0.20]	0.65	0.98	0.78

The results of training and validation in this step were slightly worse when compared with the results of Step 3, given that the total number of FN and FP increased.

The V10 was used to test the model and around 11 157 images were considered informative. After checking them, it was verified that there were still a large number of images that were inaccurately classified.

The Accuracy at this step was 0.54.

The Fig. 9 depicted that images similar to a) and b) continue to be difficult for the model to correctly classify, which could mean that 3 AL cycles weren't enough to testify major differences in the capability of the model to classify images of that type.

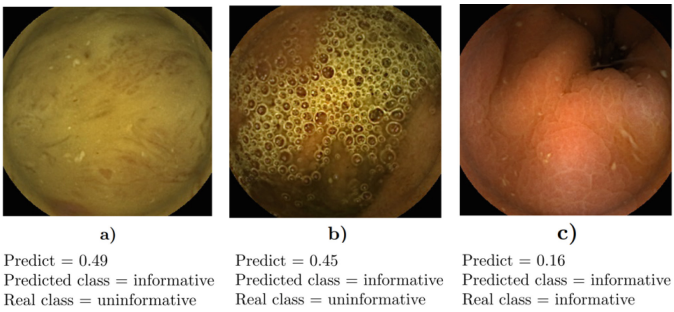

Fig. 9. Example of frames classified by the model in Step 5 as informative. The frames a) and b) were inaccurately classified as informative. he frame c) was correctly classified as informative.

Step 6: We came across V6 and verified that this video had unusual content and for that reason, we took an extra step to evaluate how the model would react to this type of content (Fig. 10).

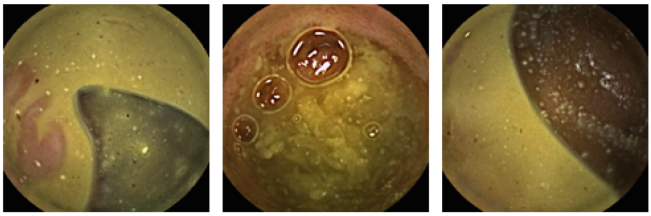

Fig. 10. V6 had an unusual amount of frames similar to images shown in this figure.

Taking the resulting model from Step 5, V6 was the raw data for the model to predict labels and using LCS, 200 images were pooled, annotated and then inserted in the initial dataset. After re-training the model, now with images from V1, V2, V5, V6 and V7, the obtained results are stated in Table 6.

Table 6. The confusion matrix and the metrics of the AUC and Loss are provided along with the Precision, Recall and F1-score for the model in Step 6.

	TP	FN	TN	FP	AUC	Loss	Precision	Recall	F1-score
Train	4770	140	8384	30	0.99	[0.0, 0.05]	0.99	0.97	0.98
Validation	1093	742	2741	37	0.96	[0.0, 0.40]	0.97	0.60	0.74

When V10 was used to test the model, it was verified that only 1314 images were classified as informative and a large percentage of those images were in

reality uninformative. The Accuracy indicated that approximately 40% of the images were correctly classified.

After checking some examples of images classified as informative by the model and considering Fig. 11, we noticed that images similar to frame a) in Fig. 9 were no longer abundant, which could mean that the model had an increased capacity to deal with those images. However, it can be seen that there is still a large number of images of type b) that are wrongly classified. This could mean that the model hasn't yet learned enough to be able to correctly classify images of this type.

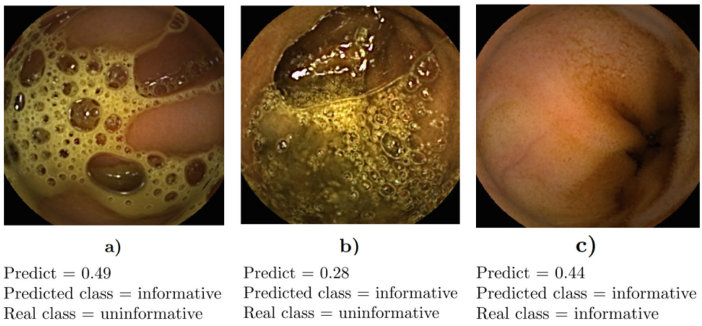

Fig. 11. Example of frames classified by the model in Step 6 as informative. The frames a) and b) were inaccurately classified as informative. The frame c) was correctly classified as informative.

4 Conclusions and Future Work

With this experiment and each of its steps, we arrived at the following conclusions:

- Introducing an AL cycle to a trained model, for classifying informative and uninformative frames of full CEV, had significant gains in reducing the time and the amount of data to be labelled, without a significant impact on the quality of the results (by comparing the results of Step 2 with Step 3);
- It was possible to observe that between LCS and MS (AL methods) the first method produced better results, given that the number of images correctly predicted is higher in the case of MS (by comparing the results of Step 3 with Step 4);
- The more iterations a model had with the AL cycle it was possible to observe that there was an increase (little one, but still an increase) in the number of correct predictions of classification (by comparing the results of Step 5 with Step 3);

- In the early stages, introducing the AL cycle to a CEV with unusual content could impact the performance of the model (by comparing the results of Step 6 with Step 5);
- Even after several iterations the model had difficulties distinguishing images with bubbles, mainly, but also with other artefacts related to the flux of different fluids inside the gastrointestinal tract from images with clear mucosa. This fact could be an indication of some ambiguity in the oracle annotation that impacted the model performance and/or the model needs more data and iterations with AL cycle to be more capable of distinguishing those types of images.

The outcome of this experiment verified the approach that using AL in classification tasks for CEV had room to grow and gave valuable cues for the following tasks as future work:

- Improving the quality of a reference dataset for tasks like abnormality classification using full CEV;
- Introducing deep active learning in classification tasks for CEV by evaluating its performance with the results of this experiment.

Acknowledgements. National Funds finance this work through the Portuguese funding agency, *FCT - Fundação para a Ciência e a Tecnologia*, within project LA/P/0063/2020.

References

1. Rawla, P., Sunkara, T., Barsouk, A.: Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. Gastroenterol. Rev. **14**, 89–103 (2019)
2. Simadibrata, M., Adiwinata, R.: Precancerous lesions in gastrointestinal tract. Indon. J. Gastroenterol. Hepatol. Digest. Endosc. **18**, 112–117 (2017)
3. Flemming, J., Cameron, S.: Small bowel capsule endoscopy: indications, results, and clinical benefit in a university environment. Medicine **97**, e0148 (2018)
4. Spada, C., et al.: Performance measures for small-bowel endoscopy: a European society of gastrointestinal endoscopy (ESGE) quality improvement initiative. United Eur. Gastroenterol. J. **7**(5), 614–641 (2019). https://onlinelibrary.wiley.com/doi/abs/10.1177/2050640619850365
5. Lee, N.M., Eisen, G.M.: 10 years of capsule endoscopy: an update. Expert Rev. Gastroenterol. Hepatol. **4**(4), 503–512 (2010)
6. Muñoz-Navas, M.: Capsule endoscopy. World J. Gastroenterol. WJG **15**(13), 1584 (2009)
7. Gueye, L., Yildirim-Yayilgan, S., Cheikh, F.A., Balasingham, I.: Automatic detection of colonoscopic anomalies using capsule endoscopy. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 1061–1064. IEEE (2015)
8. Dray, X.: Artificial intelligence in small bowel capsule endoscopy-current status, challenges and future promise. J. Gastroenterol. Hepatol. **36**(1), 12–19 (2021)
9. Radeva, P., et al.: Active labeling: application to wireless endoscopy analysis, pp. 174–181 (2012)

10. Folmsbee, J., Liu, X., Brandwein-Weber, M., Doyle, S.: Active deep learning: improved training efficiency of convolutional neural networks for tissue classification in oral cavity cancer. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 770–773. IEEE (2018)
11. Fonseca, F., Nunes, B., Salgado, M., Cunha, A.: Abnormality classification in small datasets of capsule endoscopy images. Procedia Comput. Sci. **196**, 469–476 (2022)
12. Shin, H.C., et al.: Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging **35**(5), 1285–1298 (2016)
13. Kim, H.E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M.E., Ganslandt, T.: Transfer learning for medical image classification: a literature review. BMC Med. Imaging **22**(1), 69 (2022)
14. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Angluin, D.: Queries and concept learning. Mach. Learn. **2**, 319–342 (1988)
16. Dagan, I., Engelson, S.P.: Committee-based sampling for training probabilistic classifiers. In: Machine Learning Proceedings 1995, pp. 150–157. Elsevier (1995)
17. Settles, B.: Active learning literature survey (2009)
18. Malagelada, C., et al.: New insight into intestinal motor function via noninvasive endoluminal image analysis. Gastroenterology **135**(4), 1155–1162 (2008)