Universidade de Aveiro
Ano 2023

BEATRIZ GRAMATA NUNES

Criação de Datasets de VCE: Soluções de Active Learning para Classsificação Binária de Imagens em Informativas vs Não-informativas

VCE Dataset Generation: Active Learning Solutions for Binary Classification in Informative vs Uninformative Frames

Universidade de Aveiro
Ano 2023

BEATRIZ GRAMATA NUNES

Criação de Datasets de VCE: Soluções de Active Learning para Classsificação Binária de Imagens em Informativas vs Não-informativas

VCE Dataset Generation: Active Learning Solutions for Binary Classification in Informative vs. Uninformative Frames

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Tecnologias da Imagem Médica, realizada sob a orientação científica do Professor Augusto Silva Associado do Departamento de Eletrónica da Universidade de Aveiro e do Professor António Cunha, Professor Auxiliar da Universidade de Trás-os-Montes e Alto Douro.

o júri

presidente	Prof. Doutor Nelson Fernando Pacheco da Rocha professor catedrático da Universidade de Aveiro
vogais	Prof. Doutor Anselmo Cardoso de Paiva professor catedrático da Universidade Federal do Maranhão
	Prof. Doutor António Manuel Trigueiros da Silva Cunha professor auxiliar da Universidade de Trás-os-Montes e Alto Douro

palavras-chave

VCE, Active Learning, Dataset creation, Informative Images, Least Confidence Sampling, Margin Sampling.

resumo

A Cápsula Endoscópica é uma técnica de imagem não invasiva que permite a observação do intestino delgado. No entanto, requer revisão e anotação de vídeos de duração entre 8 a 10 horas, que necessitam de ser revistos por um profissional de saúde, o que torna esta tarefa demorada. Métodos de Machine Learning atuais já conseguem assistir os profissionais através da classificação automática de descobertas nas imagens, no entanto, para atingir este estado grandes datasets de vídeos de Cápsula Endoscópica são necessários, o que requer uma quantidade de esforço insustentável. Métodos de Active Learning podem ser usados para otimizar a anotação através da identificação inteligente de imagens para serem anotadas, num grande dataset não anotado, que vão contribuir para a aprendizagem do modelo. Nesta dissertação, um estudo de Active Learning para a criação de datasets de VCE para resolver problemas binários relacionados com a classificação de imagens em informativas e não informativas, foi realizado. Algumas técnicas de Active Learning foram exploradas, tais como Least Confidence Sampling e Margin Sampling, para se concluir sobre o esforço de anotação e a rápida criação de datasets representativos. Foi verificado que o Least Confidence Sampling foi o método que melhor se adaptou aos nossos dados, dada a precisão obtida ao dividir imagens nunca vistas pelo modelo, em informativas e não informativas; e que o Active Learning tem o potencial para expandir os datasets utilizando menos dados e menos esforço humano.

keywords VCE, Active Learning, Dataset creation, Informative Images.

abstract Video Capsule Endoscopy is a non-invasive image technique that allows the observation of the small bowel. However, it requires review and Annotation of up to 8 to 10 hours of videos that need to be reviewed by a medical expert, which is very time-consuming. State-of-the-art Machine Learning methods now have the power to assist experts by automatically classifying findings in the video frames, but big Video Capsule Endoscopy annotated datasets are needed, which requires an unaffordable effort. Active Learning methodologies can be used to optimize dataset annotation through the intelligent identification of the samples to be annotated in big non-annotated datasets that most contribute to model learning. In this dissertation, a study of Active Learning to create VCE datasets, in order to solve a binary problem related to the classification between informative and uninformative frames, was made. We explored some Active Learning techniques, such as Least Confidence Sampling and Margin Sampling, to conclude about the annotation effort and the capability to rapidly create representative datasets. It was verified that Least Confidence Sampling was the more appropriate technique for our data, given the accuracy when dividing unseen video frames into informative and uninformative; and that Active Learning has the potential to expand the existing datasets using less data and human effort.

TABLE OF CONTENTS

LIST OF FIGURES ii
LIST OF FORMULAS iii
LIST OF TABLES iv
ABREVIATIONS v
CHAPTER 1 1
INTRODUCTION 1
1.1 Context 1
1.2 Motivation 2
1.3 Objectives 2
1.4 Scientific Production 2
1.5 Document Structure 2
CHAPTER 2 5
BACKGROUND 5
2.1 Gastrointestinal Tract 5
2.2 Imaging Small Bowel 5
2.3 Image Segmentation 6
CHAPTER 3 9
LITERATURE REVIEW 9
3.1 Annotation 9
3.2 Active Learning as a Solution 9
3.3 Query Scenarios 11
3.4 Uncertainty Sampling vs Diversity Sampling 12
3.5 Informative and Uninformative Images 14
CHAPTER 4 17
METHODS 17
4.1 Active Learning Pipeline 17
4.2 Dataset 18
4.3 Deep Learning model 18
4.4 Informative and Uninformative Images 19
4.5 Active Learning methods 19
4.6 Metrics (Performance Evaluation) 19
CHAPTER 5 24
RESULTS AND DISCUSSION 24
5.1 RESULTS 24
ROUND 1: Train the base model 25
ROUND 2: Train the model in new cleaned data 25
ROUND 3: Initial model + test in video 10 26
ROUND 4: Space to use Active Learning 27
ROUND 5: Annotation in video 2 + retrain using Least Confidence Sampling 28
ROUND 6: Video 10 – test with the model trained with v1 and v2 (LCS) 29
ROUND 7: Annotation in video v2 + retrain using Margin Sampling 30
ROUND 8: Video 10 – test with the model trained with v1 and v2 (MS) 30
ROUND 9: Model trained with images from three videos (v1, v5, v7) + test in video v10 31
ROUND 10: Model trained with images from four videos (v1, v5, v7, v6) + test in video v10 32
5.2 DISCUSSION 34
CHAPTER 6 37
CONCLUSIONS AND FUTURE WORK 37
REFERENCES 39

LIST OF FIGURES

Figure2.1 This image represents the regions of the small intestine and the layers.

Figure 2.2 Example of an Endoscopy Capsule and respective components.

Figure 2.3 Example of the process of segmentation in a VCE image.

Figure 3.1. Active Learning cycle.

Figure 3.2. Example of uninformative frames of the used dataset.

Figure 3.3. Example of informative frames of the used dataset.

Figure 4.1. Pipeline of the project.

Figure 4.2. Visual explanation of ROC curve, AUC, sensitivity and specificity.

Figure 5.1. Example of images of video v10 classified by the model. Image a) and b) were wrongly classified as informative, while are uninformative. Image c) was correctly classified.

Figure 5.2. Example of images of video v10 classified by the model. Image a) and b) were wrongly classified as informative, while are uninformative. Image c) was correctly classified.

Figure 5.3. Example of images of video v10 classified by the model. Image a) and b) were wrongly classified as informative, while are uninformative. Image c) was correctly classified.

Figure 5.4. Example of images of video v10 classified by the model. Image a) and b) were wrongly classified as informative, while are uninformative. Image c) was correctly classified.

Figure 5.5. Example of the unusual content observed in video v6.

Figure 5.6. Example of images of video v10 classified by the model. Image a) and b) were wrongly classified as informative, while are uninformative. Image c) was correctly classified.

LIST OF FORMULAS

Formula 3.1. This is the basic formula to calculate the Least Confidence Sampling. x* means the most uncertain instance according to model θ, $\hat{y} = \text{argmax}_Y P_\theta(\hat{y}|x)$ is the prediction with the highest posterior probability under the model θ.

Formula 3.2. This is the basic formula to calculate the Margin of Confidence Sampling. x* means the most uncertain instance according to model θ, $\hat{y}_1$ and $\hat{y}_2$ are the most and the second most confidence samples.

Formula.3.3. This is the basic formula to calculate the Ratio of Confidence Sampling. x* means the most uncertain instance according to model θ, $\hat{y}_1$ and $\hat{y}_2$ are the most and the second most confidence samples.

Formula 3.4. This is the basic formula to calculate Entropy-Based Sampling. x* means the most uncertain instance according to model θ, y ranges over all possible labelings of x and H is the entropy.

Formula 4.1. Formula to calculate Precision.

Formula 4.2. Formula to calculate Recall.

Formula 4.3. Formula to calculate F1-score.

Formula 4.4. Formula to calculate Accuracy.

Formula 4.5. Formulas to True and False Positive Rates.

LIST OF TABLES

Table 4.1 Confusion Matrix example.

Table 5.1. Overall view of the rounds of the experiment.

Table 5.2. Results obtained from training and validation using only video v1.

Table 5.3. Results obtained from training and validation using only video v1 (cleaned).

Table 5.4. Results obtained from model evaluation, using videos v2, v3 and v4.

Table 5.5. Results obtained from training and validation using video v1 (cleaned) and 200 images from video v2, using Least Confidence Sampling.

Table 5.6. Results obtained from training and validation using video v1 (cleaned) and 200 images from video v2, using Margin Sampling.

Table 5.7. Results obtained from training and validation using video v1 (cleaned) and 600 images from videos v2, v5 and v7, using Least Confidence Sampling.

Table 5.8. Results obtained from training and validation using video v1 (cleaned) and 800 images from videos v2, v5, v7 and v6, using Least Confidence Sampling.

ABREVIATIONS

AL	Active Learning
AUC	Area Under the Curve
DL	Deep Learning
FN	False Negatives
FP	False Positives
ROC	Receive Operating Curve
TN	True Negatives
TP	True Positive
VCE	Video Capsule Endoscopy

CHAPTER 1

INTRODUCTION

1.1 Context

The early detection of pathologies can lead to a rapid improvement in patient health conditions with fewer complications [1]. For that reason, Video Capsule Endoscopy has emerged as a revolutionary non-invasive imaging technique for visualizing the small bowel [2]. Nonetheless, the full realization of its diagnostic potential comes with a significant challenge: the meticulous review and annotation of extensive videos that have a duration of 8 to 10 hours [2]. This process demands a skilled medical professional to analyze the obtained data, which can be expensive and time-consuming [2].

In this context, optimizing the video review and annotation process becomes a priority, as it directly impacts the efficiency and effectiveness of medical diagnoses. Machine Learning techniques appear to be useful in this context.

In Computer Science, Human-Computer Interaction has become the most important for Machine Learning [3]. Creating an interface for the human user to construct the training dataset includes distinct knowledge areas, such as social sciences, psychology, user-experience design and others [3].

Supervised Learning is powering about 90% of Machine Learning applications; for example, when a human user tells his in-home device to turn up the volume, the device knows what to do because humans have previously spent many hours teaching the machine how to interpret different commands [3]. These Supervised Learning models need to receive more labelled data to become more accurate in their given tasks, and to accomplish that accuracy, more training inputs must be provided [3].

When the nature of the data changes over time, like in the medical field, just a few labelled samples are not enough [3] to support the Machine Learning methods. For that reason, a large quantity of data needs to be annotated, but the human focus and the errors that can be made due to this cognitively demanding task are limiting factors to achieving the desired accuracy [3]. The annotation becomes a problem that urgently needs a quick fix and, so, a priority [4].

Active Learning appears as a solution since it can offer an effective way for selecting the most informative samples in big non-annotated datasets to be labelled by the human user, speeding up the Annotation process [9] and consequently contributing to the model learning.

1.2 Motivation

Given the potential of Active Learning techniques, I feel motivated to study Active Learning techniques and apply them to important medical challenges, such as the problem related to VCE datasets.

If Active Learning is useful, then there is a possibility of rapidly improving the datasets used in the Machine Learning classifiers, and this can be a way to improve the classification tools related to medicine, especially when dividing the images into informative/uninformative.

1.3 Objectives

The principal objective for this dissertation is the study of some Active Learning methods focused on dataset generation for binary problems such as image classification into informative/ uninformative.

For that, some sub-objectives were identified, such as:

- Study of Active Learning Methods
- Definition of a protocol to image annotation into informative and uninformative
- Dataset preparation for binary classification into informative/ uninformative
- Implementation of Active Learning methods that better adapts to the data
- Evaluation of the impact of Active Learning when there are only one video and when there are more than one video

1.4 Scientific Production

A paper entitled “Informative classification of Capsule Endoscopy videos using Active Learning”, was submitted MobiHealth 2023 Conference, Index Scopus, and was accepted.

1.5 Document Structure

This document is structured into six chapters, each serving a specific purpose within the context of our research:

Chapter 1 The present chapter briefly explains an introduction to the problem of annotation and a possible solution.

Chapter 2 A background of the Gastrointestinal tract and its significance for the human body, followed by ways to image it, is presented. Capsule Endoscopy is given, and the importance of segmentation is explained.

Chapter 3 A literature review where the extensive body of work related to Annotation, Active Learning, and the methods associated with active Learning is provided.

Chapter 4 A description of the entire process of constructing our pipeline and selecting datasets to distinguish informative and non-informative images is made. Additionally, the Deep Learning and Active Learning methods and the fundamental metrics that support our approach are detailed.

Chapter 5 The findings are presented and followed by a discussion of the results. This section enables the drawing of insights and conclusions based on the data and outcomes of the research.

Chapter 6 In the final chapter, conclusions are summarized, offering a concise overview of the knowledge gained through our research. Additionally, we provide ideas for future work, guiding the path for continued exploration and progress in this field.

CHAPTER 2

BACKGROUND

2.1 Gastrointestinal Tract

The organ system responsible for food digestion is the Gastrointestinal Tract. During the digestive process, the food is decomposed into essential nutrients and minerals to maintain the body's energy levels and repair cells [2].

The small bowel is divided into four layers (serosa, muscle, submucosa, and mucosa, which is responsible for the absorption and regulation of the intestinal flora). It can be split into three parts: duodenum, jejunum and the ileum [2], as seen in Figure 2.1.

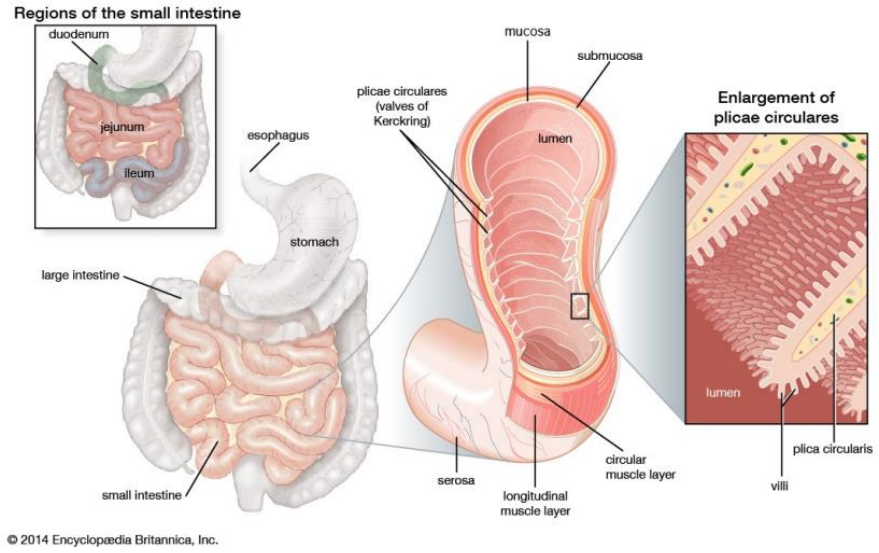

Figure 2.1 This image represents the regions of the small intestine and the layers [2].

2.2 Imaging Small Bowel

Knowing that this organ is around 3-5 meters, it becomes a challenge to diagnose pathologies since the traditional imaging methods (Ultrasound, Magnetic Resonance Imaging, Esophagogastroduodenoscopy and Colonoscopy) may not be efficient due to the intestinal gas and artefacts present in the small bowel [2]. Some of these procedures can be invasive and painful [5].

The Endoscopy Capsule is a device that was introduced in 2000 [6]. This is a pill-like, non-invasive camera that is swallowed by the patient and travels for the digestive system through the peristaltic movements while saving the information in a portable device that the patient carries [2] [6].

Figure 2.2 Example of an Endoscopy Capsule and respective components [7].

This device usually has a wide-angle camera, light source, batteries and other electronics [8], as shown in Figure 2.2.

The produced video is an eight-hour-long video that needs to be analyzed by an expert, and this task takes around two to three hours to complete. Exploring all the obtained images (about 50,000 per video) [6] is a time-consuming and laborious process since only a small number of frames are the ones that have lesions, and this increases the risk of skipping essential pathologies [9], for that reason, tools to support image analysis are convenient and desired [10].

2.3 Image Segmentation

To reduce the typical challenge of manual annotation, a variety of weakly supervised segmentation algorithms have been suggested [11], such as the one in Figure 2.3.

Image segmentation is a crucial part of biomedical image analysis. Deep Learning (DL) stands out as a potent Machine Learning tool, exhibiting promising outcomes in image analysis and recognition domains, particularly within biomedical applications [11]. Therefore, DL has been applied in the medical field, and techniques have been developed to help clinicians with the prognosis and the disease treatment regimes [11].

Deep Learning still faces a huge obstacle regarding the number of images required by the training process, considering that most of the supervised Machine Learning techniques

perform well when trained on many hundreds/thousands of labelled data, data that needs to be annotated by an expert [11].

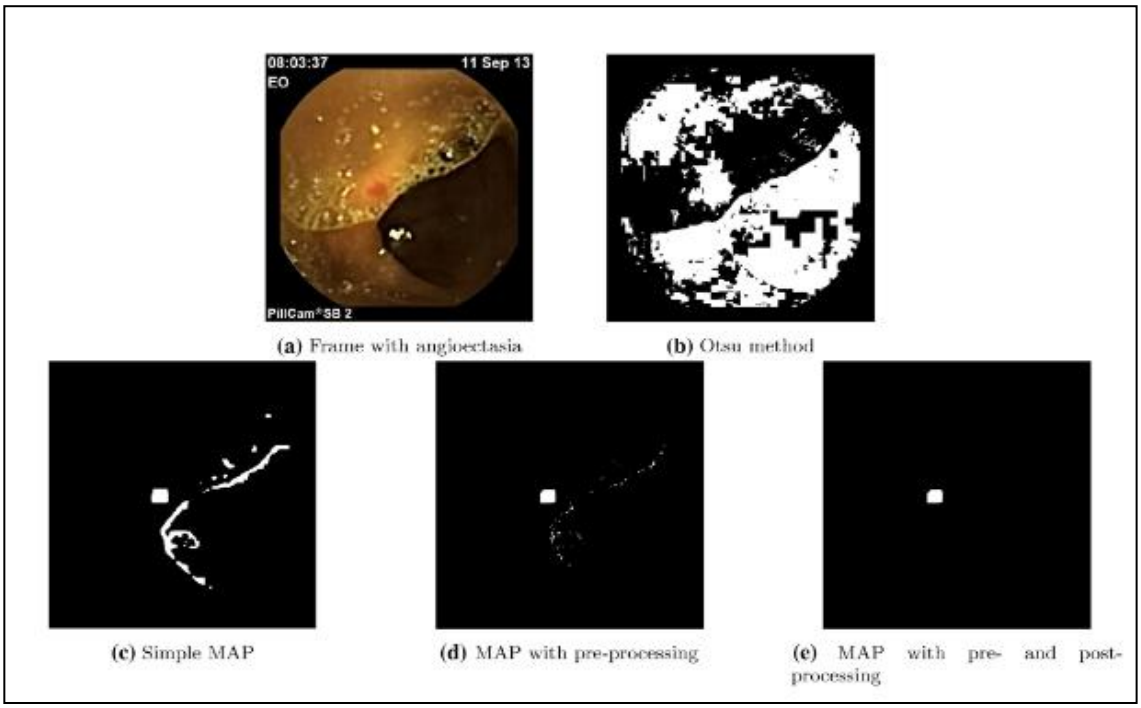

Figure 2.3 Example of the process of segmentation in a VCE image [12].

CHAPTER 3

LITERATURE REVIEW

3.1 Annotation

Annotating images is a repetitive task prone to human error due to Repetition Priming [3]. Repetition Priming happens when a sequence of functions influences the human's perception, and for that reason, it is possible to conclude that labelling is a difficult task [3]. In addition, labelling is time-consuming and medical images require professional knowledge to annotate them [4], and as a result, the annotation process became very expensive. Consequently, only a very small range of datasets are available, and most are small [11]. These are the datasets that have been used in the development of Deep Learning models [11].

To overcome this problem, Transfer Learning has been used [10], because it takes an already trained model and adapts it to another through the usage of distinct architectures and parameters [3] [6]. It is crucial to notice that some experts are questioning the performance and the results of these techniques due to the small datasets used for training. Consequently, there is a necessity to create datasets with a large number of annotated samples.

With all the pieces together, it is possible to affirm that obtaining a large dataset is relatively simple, but annotating it comes at a high cost. Therefore, maximizing the model's performance while reducing the annotation costs becomes a primary concern [4] [13].

3.2 Active Learning as a Solution

In the real world, an enormous amount of data is continuously generated from many different sources, and consequently, most real-world objects change over time [14].

In theory, if all this data is added to the training set of a model, then the model should become more accurate [3] and prepared to deal with actual data, but to add new samples to a model, these must be annotated.

Such as in the medical field, the data can require expert annotation, leading to an escalation in annotation costs [13]. As explained before, humans are susceptible to errors when dealing with repetitive tasks, which is especially critical in the medical field [13].

Active Learning came out as a potential solution to this problem because it appears to be efficient in the process of dataset creation and annotation [15].

Active Learning, also known as Query Learning [13], is a subfield of Machine Learning that is responsible for selecting a subset of valuable and unlabeled samples to be labelled by the human reviewer as part of an interactive learning process, reducing the costs of annotation without compromising the performance of the model.

For the subset selection to be labelled, the AL techniques assume, through some heuristic strategies, that different samples have different values for the model update [4]. Therefore, samples with the highest values are selected and then introduced in the annotation process to integrate the training set [15], as seen in Figure 3.1.

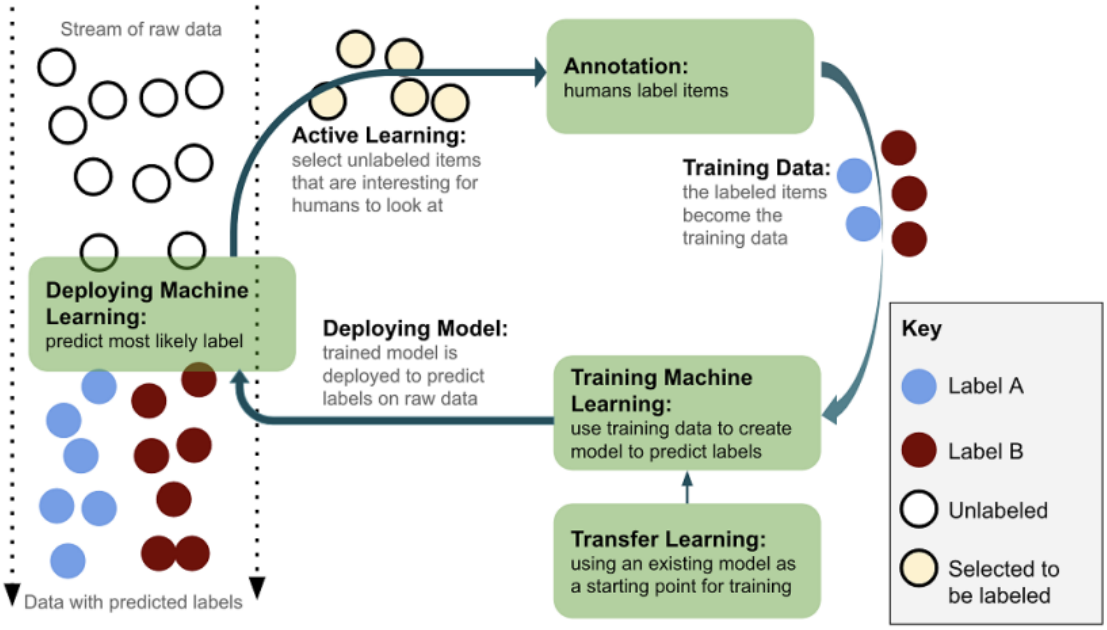

Figure 3.1 Active Learning cycle [3].

Interactive Learning processes merge humans and machines. Humans, named oracles by some authors (doctors, in the case of the medical field), are part of the system and participate actively in the learning process by annotating the set of images that are selected for annotation [16].

The system ends up being more accurate because the human inputs provide more valuable and precise information, which positively affects the final outputs [4]. The expectation is that these models would have higher accuracy with low cost and less human effort [13]. In theory, this means that Active Learning can achieve exponential acceleration in labelling efficiency [13].
It is important to notice that when only a few labelled instances are needed to train a model, it may not be appropriate to use Active Learning since it is more useful when there are a very high number of unlabeled samples that need annotation [17].

3.3 Query Scenarios

When Active Learning is useful for the problem, the learner has distinct approaches to formulate questions to select the data to be annotated, which can be through Membership Query Synthesis, Stream-Based Selective Sampling and Pool-Based Sampling [17].

Membership Query Synthesis The Active Learning model generates its own samples based on the current state of the model and the available data. This allows the model to target specific areas, accelerating the learning process. However, this could produce some examples that can be incomprehensible to the human oracle to label [2] [17].

Stream-Based Selective Sampling All the images from the database are presented to the model that will decide if the sample must be annotated. A threshold can replace this process, but with this approach, the model became more naïve [2]. Otherwise, if the input distribution is uniform, then the model does not have many advantages compared to the Membership Query Synthesis [17].

Pool-Based Sampling Assumes a big pool of unlabeled data, but only a tiny piece of that data is the informative one. Given that in the real world, there is always a big quantity of data being constantly produced and only some of that is useful, this method is the most popular [17]. Although, this strategy can be computationally demanding [2].

3.4 Uncertainty Sampling vs Diversity Sampling

There are two different Active Learning approaches: Uncertainty Sampling and Diversity Sampling. Each approach has different associated strategies.

- **Uncertainty Sampling**

This approach identifies the most uncertain samples and sends them to the oracle for annotation. Usually, this data is close to the decision boundary [17]. To calculate the uncertainty of an element, the following strategies can be used.

Least Confidence Sampling Difference between the most and the 100% confidence (maximum) predictions: this intends to calculate the uncertainty of a prediction and select the least confident sample to be annotated [3].

$$\begin{aligned} x^*_{LC} &= \underset{x}{\operatorname{argmin}}\ P_\theta(\hat{y}|x) \\ &= \underset{x}{\operatorname{argmax}}\ 1 - P_\theta(\hat{y}|x) \end{aligned} \tag{1}$$

Formula 3.1. This is the basic formula to calculate the Least Confidence Sampling. x* means the most uncertain instance according to model θ, $\hat{y} = \operatorname{argmax}_Y P\theta(\hat{y}|x)$ is the prediction with the highest posterior probability under the model θ [17].

Margin of Confidence Sampling Difference between the two most confident predictions. A small margin means that both samples are ambiguous to the model [17].

$$\begin{aligned} x^*_M &= \underset{x}{\operatorname{argmin}} \left[P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x) \right] \\ &= \underset{x}{\operatorname{argmax}} \left[P_\theta(\hat{y}_2|x) - P_\theta(\hat{y}_1|x) \right] \end{aligned} \tag{2}$$

Formula 3.2. This is the basic formula to calculate the Margin of Confidence Sampling. x* means the most uncertain instance according to model θ, $\hat{y}_1$ and $\hat{y}_2$ are the most and the second most confident samples [17].

Ratio of Confidence Sampling Ratio between the two most confident predictions [3].

$$x^*_{RC} = P_\theta(\hat{y}_1|x) \,/\, P_\theta(\hat{y}_2|x) \tag{3}$$

Formula 3.3. This is the basic formula to calculate the Ratio of Confidence Sampling. x* means the most uncertain instance according to model θ, $\hat{y}_1$ and $\hat{y}_2$ are the most and the second most confident samples [17].

Entropy-Based Sampling Is the difference between all the predictions [3].

$$\begin{aligned} x_H^* &= \underset{x}{\operatorname{argmax}}\ H_\theta(Y|x) \\ &= \underset{x}{\operatorname{argmax}}\ -\sum_y P_\theta(y|x)\log P_\theta(y|x) \end{aligned} \tag{4}$$

Formula 3.4. This is the basic formula to calculate Entropy-Based Sampling. x* means the most uncertain instance according to model θ, y ranges over all possible labellings of x, and H is the entropy [17].

- **Diversity Sampling**

This approach selects samples based on their discrepancy to the labelled data, which means that in some cases, the number of items to be annotated by the oracle is more adapted within each iteration of Active Learning [3].
Similar to Uncertainty Sampling, there are several Diversity Sampling approaches.

Model-Based Outlier Sampling Is a neural model that searches for samples that are unknown to the model through the lowest activation in a layer [3].

Cluster-Based Sampling: Divide the data into a large number of clusters and select samples to be annotated from each cluster [3].

Representative Sampling Search for samples that closely resemble our target domain, compared to the training set [3].

Sampling for Real-World Diversity Ensure that the data used to train the model represents real-world diversity as much as possible [3].

3.5 Informative and Uninformative Images

As explained before, VCE produces around 50,000 images that need to be analyzed, so strategies have been developed, but the datasets used are small. To add more data to these datasets, an expert is needed to annotate. An expert is required because of the complexity and the influence of the problem.

When the problem is the separation of images into informative and uninformative (binary classification), the oracle needs to be aware of what is considered informative/ uninformative in a VCE sample.

When an image contains food digestion, intestinal juices or bubbles and the view of the mucosa is occluded, then it is considered uninformative [18], as can be seen in Figure 3.2.

It is crucial to understand how to recognize intestinal content because, in some cases, even when the samples contain digestive material, it can be indicative of intestinal dysfunctions [18].

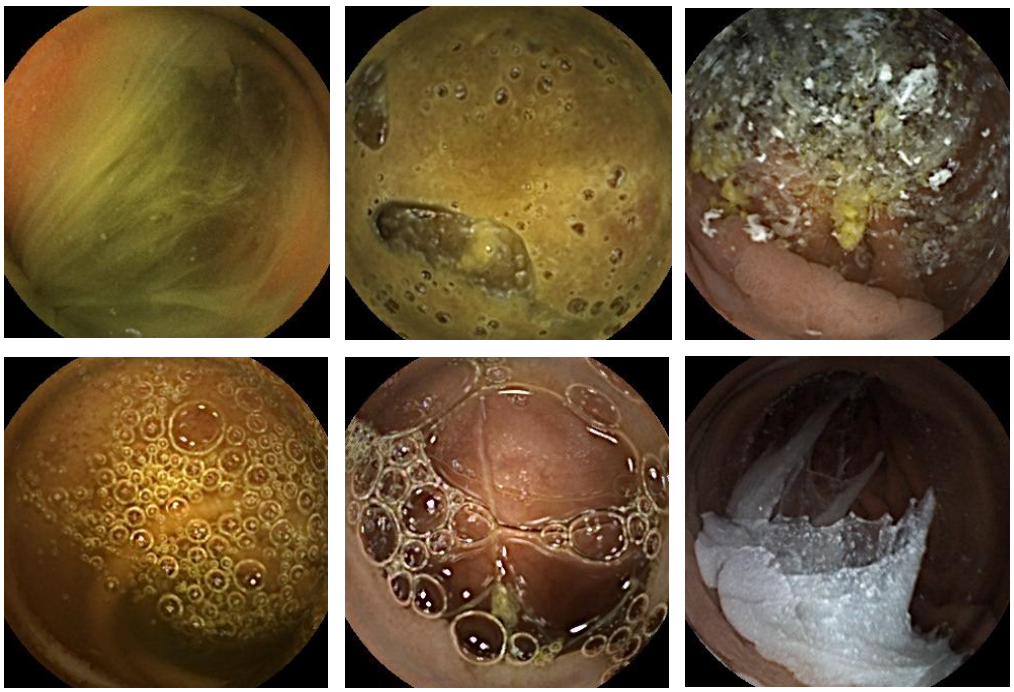

Figure 3.2. Example of uninformative frames of the used dataset.

When the image is cleaned, then is considered informative, such as the ones in Figure 3.3.

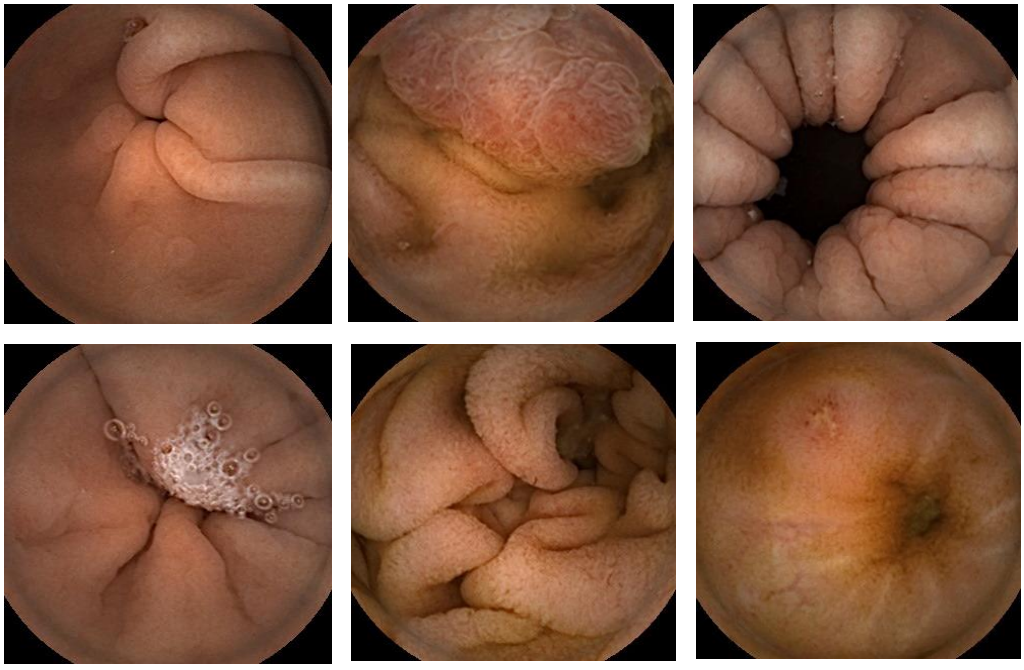

Figure 3.3. Example of informative frames of the used dataset.

CHAPTER 4

METHODS

4.1 Active Learning Pipeline

According to all the information described above, a pipeline was defined, and the main steps can be seen below in Figure 4.1.

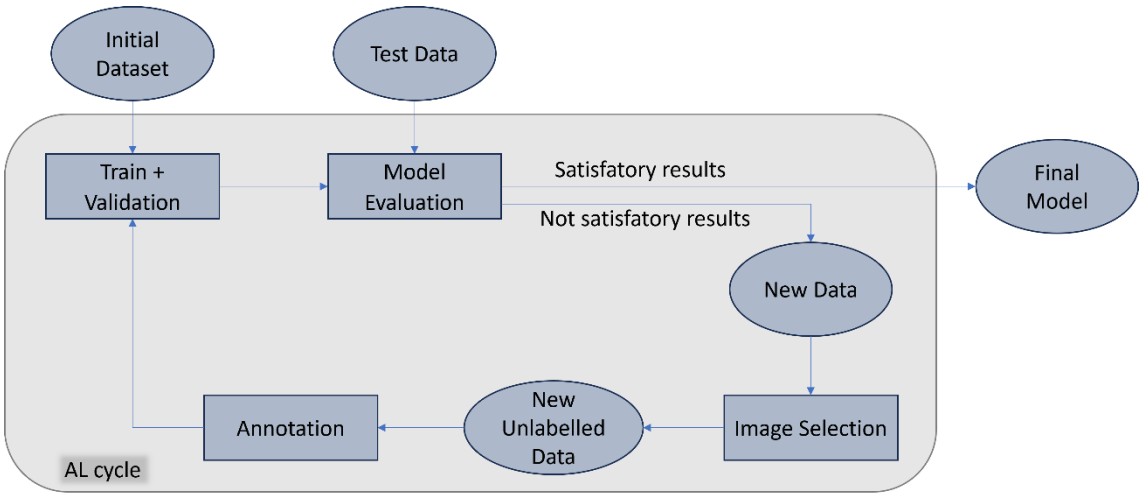

Figure 4.1. Pipeline of the project.

Firstly, an initial dataset was used to train the model, and then, the model was evaluated in the test data that includes never-seen images. If the results were satisfactory, then the model was considered the final model, but if the results were not acceptable, then the AL cycle was initiated.

In this cycle, a new amount of data was presented to the model, and according to the Active Learning method used, some samples were selected and annotated by an oracle (annotation process).

Then, these images were added to the initial training set and the model was retrained in the initial dataset plus new annotated samples, finally, the performance of the model was re-evaluated to verify if the results were better or if there were no improvements.

4.2 Dataset

A private dataset was used and consisted of 6 capsule endoscopic videos (v1, v2, v5, v7, v6 and v10) converted into images provided by a Portuguese *Public Hospital*. All these videos belong to different patients that have various pathologies.
The images were pre-selected, and each video has around 19,000 and 21,000 images. All videos were used completely unlabeled, except for video v1, which was used to train and validate the model, and for that reason, all the images from this video were annotated.
Video v10 was used as a whole to test the ability of the model to separate the images into informative and not informative. Compared with the pipeline, this video is the test data.

4.3 Deep Learning model

In previous work, a study related to the effect of Transfer Learning of three pre-trained models (ResNet50, Inception and EfficientNetB3) on the classification of VCE images was made. It was verified that ResNet50 had good results when training a model in small datasets since, through the usage of Transfer Learning, this architecture has archived values for AUC, precision, recall and F1-score, better than for the other architectures [6].

ResNet50, a deep convolutional neural network pre-trained on the extensive ImageNet dataset, has demonstrated exceptional performance across various computer vision tasks. Using ResNet50 as a base model allows us to take advantage of its learned characteristics and adapt them to our specific problem, which leads to notable savings in time and computational resources compared to training a model entirely from scratch.

To train the ResNet50 model, we employed two distinct approaches. Firstly, we updated only the final layer, effectively utilizing the pre-trained model as a feature extractor while solely modifying the classification layer weights. This approach was carried out over 50 epochs, with a batch size of 200, using the Adam optimization algorithm, which is an extension of the stochastic gradient descendent method that is based on adaptive estimation of the first-order and second-order moments. The learning rate was set at 0.00001.

Considering that our initial dataset only has video v1, this means 12,524 samples divided into 70% for train and 30% for validation, and when comparing with the total number of images produced in a Capsule Endoscopy Video (between 50 000 – 60 000 images), then this was

considered a small dataset, and for that reason, a ResNet50 model was utilized through Transfer Learning.

No data augmentation was used in this study because the objective is to analyze the direct impact of Active Learning in the improvement of a model.

4.4 Informative and Uninformative Images

Taking into account the information described above, to annotate the images, we considered that images with approximately 65% of clean mucosa were informative images. All the images were observed and this way divided into informative and uninformative.

4.5 Active Learning methods

There are different Active Learning strategies, as explained before, but only two of them were applied in this work: Least Confidence Sampling and Margin Sampling.

The objective is to evaluate two of the most used Active Learning models, understand the impact on the model's performance and evaluate which one is the most appropriate for the model update.

4.6 Metrics (Performance Evaluation)

Metrics are quantitative values that are calculated and used to conclude the performance of a model. There are many metrics, such as AUC, ROC curve, Accuracy, Mean Absolute Error, and others.

The selected metrics for this project were the most common ones: Loss, Recall, Accuracy, AUC and ROC curve, F1-Score and Precision.

To calculate these metrics, true positive (TP), true negative (TN), false positive (FP) and false negative (FN) values were needed. These values correspond to the categorization of the predictions and the actual outcomes.

True positives and true negatives are the images that are correctly predicted (the prediction and the actual outcome are the same), and false positives and false negatives are the images whose prediction and the true outcomes are different, so false positives are images that are predicted as positives but in reality are negative, and false negative are images that are predicted as false but in reality, are positive.

Confusion Matrix

The confusion matrix is useful for observing the summary of the model's performance in a matrix format and includes the FP, FN, TP and TN values, as can be seen in Table 4.1.

Table 4.1 Confusion Matrix example.

		True Class	
		Positive	Negative
Predicted Class	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Loss

Loss is a metric that quantifies the discrepancy between the predicted values and the true values.

The loss value must be as close as possible to 0.00. A small loss value implies that the predicted values are close to the actual values, and therefore, the model is considered good.

Precision

Precision quantifies the correct positive predictions in all the positive predictions made by the model and is calculated as follows. A high precision score suggests that the model's positive predictions are likely to be correct.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

Formula 4.1. Formula to calculate Precision [6].

Recall

Recall is also known as Sensitivity or True Positive Rate, and it measures the capacity of the model to correctly identify the positive class regarding the whole positive distribution. A high recall score signifies that the model is effective at identifying a significant portion of the positive instances in the dataset.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

Formula 4.2. Formula to calculate Recall [6].

F1-Score

This metric evaluates the performance of a model using recall and precision and is helpful to use this metric when there is an imbalance between classes. This value varies between zero and one, and higher scores mean better overall model performance in terms of both precision and recall.

$$F1\text{-}score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{7}$$

Formula 4.3. Formula to calculate F1-score [6].

Accuracy

The accuracy is the ratio between the number of correct predictions and the total number of predictions.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{8}$$

Formula 4.4. Formula to calculate Accuracy.

When there are imbalanced datasets, the necessity of using other metrics is high because only accuracy can cause incorrect conclusions. This is the big disadvantage of this metric.
To calculate the accuracy of the test video, a random strategy was used to select the images. Given that the video to label has 19 097 images (video v10), and we want a confidence level of 95% and a margin of error of 5%, a Python strategy was used to randomly select 400 images to be our sample. Then, these 400 images were manually labelled and became the representative sample that we used to evaluate the accuracy of the model.

ROC (Receiver Operating Characteristic) curve

This curve plots the relation between the true positive rate and the false positive rate at all possible classification thresholds [6].

The true positive rate and the false positive rate are the fraction of the real positive instances that the model accurately identifies as positive and the fraction of the actual negative instances that the model wrongly identifies as positive [2].

$$\mathrm{TPR} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{9}$$

$$\mathrm{FPR} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}} \tag{10}$$

Formula 4.5. Formulas to True and False Positive Rates [2].

AUC (Area Under the Curve)

This is the area beneath the ROC curve and is responsible for quantifying the overall performance of the model independently of the threshold [2].

Higher values of AUC mean that there is satisfactory discrimination between classes, as can be seen in Figure 4.2.

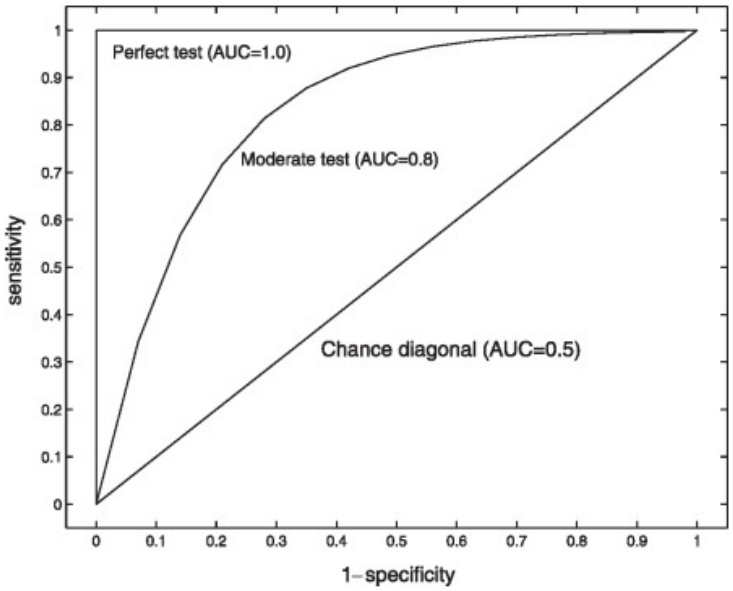

Figure 4.2. Visual explanation of ROC curve, AUC, sensitivity and specificity [19].

CHAPTER 5

RESULTS AND DISCUSSION

5.1 RESULTS

During the execution of this project, a series of rounds were executed with and without the usage of Active Learning and using different techniques.
The following table (Table 5.1) is the overall view of the model evolution over time, and it is possible to observe the impact of Active Learning on the knowledge acquired by the model.

Table 5.1. Overall view of the rounds of the experiment.

	AL Methods	Data in model	Test data	Accuracy
Round 1	-	V1	-	-
Round 2	-	V1	-	-
Round 3	-	V1	V10	0,42
Round 4	-	V1	-	-
Round 5	LCS	V1, V2	-	-
Round 6	-	V1, V2	V10	0,52
Round 7	MS	V1, V2	-	-
Round 8	-	V1, V2	V10	0,40
Round 9	LCS	V1, V2, V5, V7	V10	0,54
Round 10	LCS	V1, V2, V5, V7, V6	V10	0,41

The conclusions related to the test video (video v10) were based on the observation of the images that were classified by the model. The accuracy was calculated to prove quantitively the results observed by the team. Taking into consideration that the accuracy has the disadvantage presented above, metrics for train and validation were presented to confirm that the model did not suffer from overfitting and that it had good results and, for that reason, was ready to classify the test data.

ROUND 1: Train the base model

Initially, video 1 used to have a total of 19 815 images, which were divided into 70 % for training and 30% for validation, this means 13 871 images for train and 5 944 images for validation.

The images were separated into informative and uninformative, which makes 4 719 informative images for training, 9 152 no-informative images for train, 2 022 informative images for validation and 3 922 no-informative images for validation.

After training the model, during validation, it was verified that the model had a loss that mainly varied between 0.0 and 0.3. The AUC of the train was good, but the AUC of the validation had a lower value (0.927), as expected.

Table 5.2. Results obtained from training and validation using only video v1.

	TP	FN	TN	FP	AUC	Loss
Train	4719	0	7807	1345	1,00	[0,00 ; 0,1]
Validation	1835	187	2778	1144	0,93	[0,00 ; 0,8]

The confusion matrix for train and validation was calculated, as can be seen in Table 5.2, and it was verified that the number of samples considered as false positive or false negatives was high, which means that the initial dataset has some space for improvement.

For that reason, this initial dataset was cleaned to eliminate the images that could cause some confusion to the model.

ROUND 2: Train the model in new cleaned data

After video 1 is cleaned and all the images that could cause some confusion to the model are removed, video 1 remains with a total of 17 137 annotated images.

These images were divided into alike first round, so 12 524 images for train and 4 613 images for validation. Then, they were separated into informative and no-informative, which makes 4 719 informative images for training, 7 805 no-informative images for train, 1 835 informative images for validation and 2 778 no-informative images for validation.

The model was retrained in the initial cleaned dataset.

Table 5.3. Results obtained from training and validation using only video v1 (cleaned).

	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

As can be seen in Table 5.3, the AUC of the model is around 1.000 in training and 0.988 in validation. The loss values vary mostly between 0.0 and 0.7. After calculating the confusion matrix, we can verify that the model learned in more precise images that do not cause ambiguities.

This is considered the initial trained model.

Through the analysis of the distinct metrics, we can observe that the model already has satisfactory results, which means that there was no need to make changes to the initial ResnNet50 model.

At this point, we are ready to present the test video to the initial model and observe how efficient the model is in image classification into informative and uninformative.

ROUND 3: Initial model + test in video 10

Video v10 was introduced into the model to verify the performance in separating images into informative and uninformative, and it was verified that the model was capable of separating 848 images as informative, but when observing these images, it was noticed that a large number of uninformative images was still be considered informative.

Considering the method to calculate the accuracy of this video, 166 images were correctly predicted in the universe of 400 images, this means an accuracy of 0.415. With this accuracy, we can generalize the results and conclude that the model can correctly classify about 42% of the images.

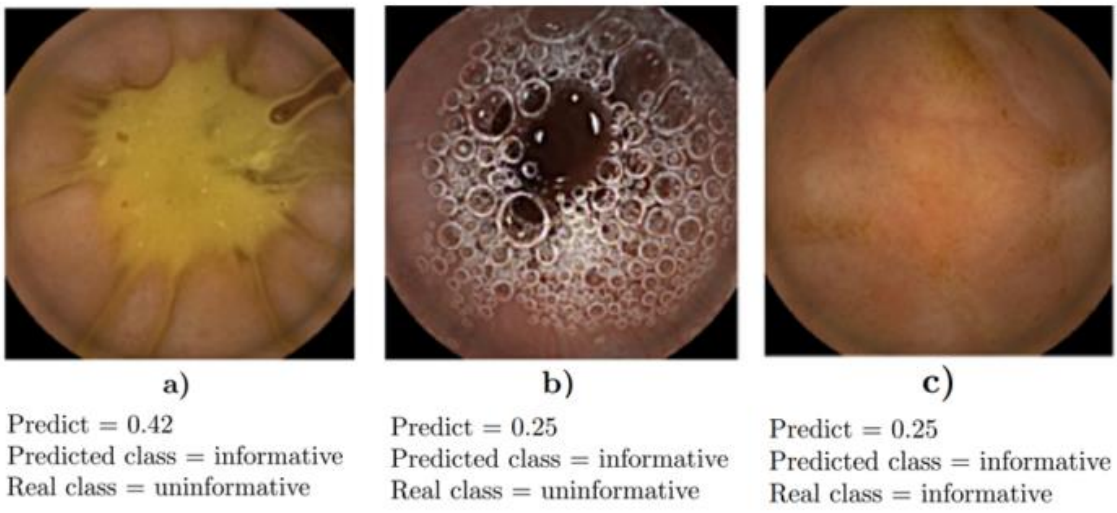

Figure 5.1. Example of images of video v10 classified by the model. Image a) and b) were wrongly classified as informative, while are uninformative. Image c) was correctly classified.

As can be observed in Figure 5.1. Images b) and c) were both considered informative and have similar predictions, even when we can clearly observe that image b) is an uninformative one. This implies that at this point, the model did not have correctly learned that images similar to b) do not have informative information.
At the end of this round, we can conclude independently of the satisfactory results in terms of metrics the model is still not very precise and still have space for improvements.

ROUND 4: Space to use Active Learning

This is a confirmation round, where a little test was executed to verify if there is some space to improve the model, considering that the model already presented satisfactory metrics. Video v2, v3, and v4 were labelled in order to obtain TP, TN, FP and FN.
In terms of the number of images per video:

- Video v2 has a total of 9 012 images: 1 417 informative and 7 595 uninformative.
- Video v3 has a total of 16 174 images: 1 023 informative and 15 151 uninformative.
- Video v4 has a total of 9 115 images: 1 417 informative and 7 595 uninformative.

After model evaluation, it was possible to observe the following results.

Table 5.4. Results obtained from model evaluation, using videos v2, v3 and v4.

	TP	FN	TN	FP	AUC	Loss	Accuracy
Video 2	107	1310	7562	33	0,89	0,15	0,85
Video 3	626	397	14955	196	0,97	0,02	0,96
Video 4	247	261	8318	307	0,93	0,03	0,94

As can be seen in Table 5.4, for videos v3 and v4, the accuracy is closer to 1, and the number of false predictions (FN and FP) is small, this means that the model can already be successful in recognizing a considerable volume of data that belong to the same type of images of video 1.

However, for video 2, considering the same factors, we can notice that the model does not learn enough to correctly predict the class of the images, so the number of false positives and false negatives is a little high.

After round 3, we can conclude that the images of videos v3 and v4 are identical to the ones used for training, and for that reason, the model is already good. This signifies that the model has good accuracy when predicting images that are very much like video 1 but not so good when predicting images like video 2.

With this round, we confirm that the initial model (only trained with cleaned video v1) still has space for improvement.

ROUND 5: Annotation in video 2 + retrain using Least Confidence Sampling

With the objective of increasing the knowledge of the model, an Active Learning strategy, in this case, Least Confidence Sampling, was selected and introduced in the model.

Video 2 was inserted, and using Active Learning, 200 images were selected, annotated, and then incorporated into the initial dataset. After retraining the model with images from video 1 and video 2, the obtained results were the following ones.

Table 5.5. Results obtained from training and validation using video v1 (cleaned) and 200 images from video v2, using Least Confidence Sampling.

	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

As can be seen in Table 5.5, compared with the results of round 2, the number of false negatives in train and validation decreased, but the number of false positives increased.

ROUND 6: Video 10 – test with the model trained with v1 and v2 (LCS)

Video v10 was employed to evaluate the model trained on videos v1 and v2. Approximately 3828 images were initially categorized as informative. However, after observing the images, it became evident that a significant number of these images had been misclassified.
Considering the method to calculate the accuracy, the accuracy was 0.52. With this accuracy, we can generalize the results and conclude that the model can classify correctly more or less 52% of the images.

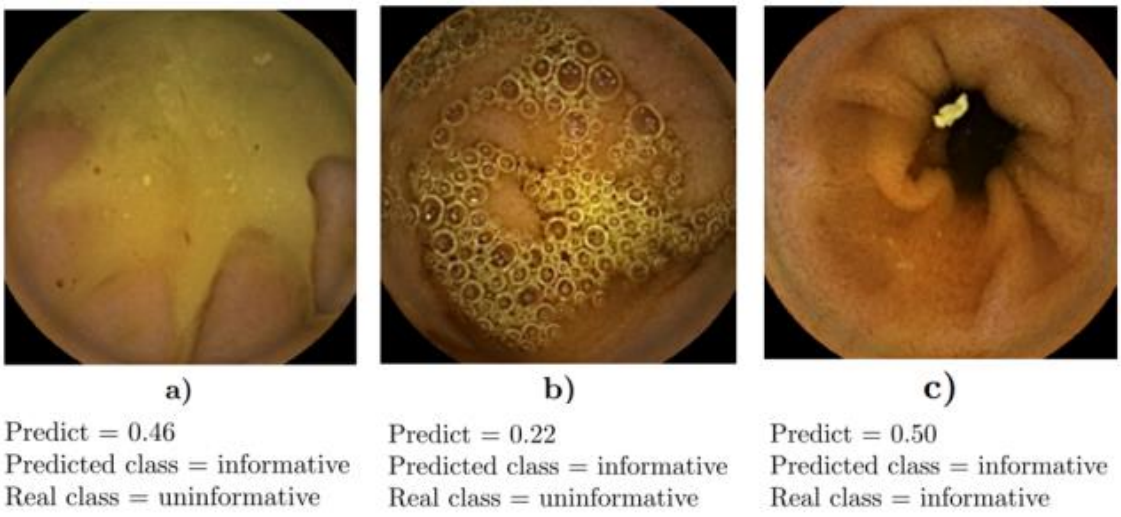

Figure 5.2. Example of images of video v10 classified by the model. Image a) and b) were wrongly classified as informative, while are uninformative. Image c) was correctly classified.

Through observation of Figure 5.2. it is possible to observe that images similar to a) and b) are still wrongly classified as informative. It is notable that image a) presents a prediction very identical to image c), which means that the model does not have enough knowledge to correctly classify images similar to a) and b).

ROUND 7: Annotation in video v2 + retrain using Margin Sampling

To analyze which method of Active Learning is more effective for our data, Margin Sampling was introduced in the model, replacing Least Confidence Sampling.
Therefore, the model obtained from round 2 was used and using Margin Sampling methods, 200 images were selected for annotation. After retaining the model with pictures from video v1 and 200 images from video v2, the obtained results were the following ones.

Table 5.6. Results obtained from training and validation using video v1 (cleaned) and 200 images from video v2, using Margin Sampling.

	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

Comparing the results of Table 5.6 with the ones obtained in round 5, it is possible to conclude that the results with Least Confidence Sampling were better than those obtained using the Margin Sampling regarding AUC, Loss and the total number of false predictions.

ROUND 8: Video 10 – test with the model trained with v1 and v2 (MS)

Even with the worst results when retraining the model, video v10 was used to test the model, and only 319 images were considered informative. The accuracy was calculated (using the method described above), and, in a total of 400 images, only 159 were correctly predicted.

a)
Predict = 0.48
Predicted class = informative
Real class = uninformative

b)
Predict = 0.01
Predicted class = informative
Real class = uninformative

c)
Predict = 0.47
Predicted class = informative
Real class = informative

Figure 5.3. Example of images of video v10 classified by the model. Image a) and b) were wrongly classified as informative, while are uninformative. Image c) was correctly classified.

Similar to round 6, images a) and c) of Figure 5.3 present close predictions and are both predicted as informative, even when only image c) is the only positive one.
Comparing the results of train, validation, and test, it is possible to assume that the model did not increase its knowledge through the usage of Margins Sampling instead of Least Confidence Sampling, and there are no important improvements when the results are compared with the ones obtained in round 5 and 6.
For those reasons, in the following rounds, only Least Confidence Sampling was used.

ROUND 9: Model trained with images from three videos (v2, v5, v7) + test in video v10

To evaluate the model's capacity to classify images into informative and uninformative, some Active Learning cycles were executed, and this way concluding about the improvement of the model's knowledge.
At this point, three Active Learning cycles were executed for videos v2, v5 and v7. The initial model (from round 2) was used, and then video v2 was introduced in the Active Learning cycle, 200 images were selected, and the model was retrained with images from v1 and v2. And the same reasoning was used for video v5 and v7. In the end, the model was trained with all images from video v1 and 600 images from video v2, v5 and v7 (200 images from each video).

Table 5.7. Results obtained from training and validation using video v1 (cleaned) and 600 images from videos v2, v5 and v7, using Least Confidence Sampling.

	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

As can be seen in Table 5.7, in terms of training and validation, the model presents more false predictions (FN and FP) when compared with the results obtained in round 2 (model only trained with video v1 – cleaned).

Even so, video v10 was introduced to test the model, and 11157 were considered informative.

Applying the method described above to calculate the accuracy, we obtained an accuracy of 0.54, this means that the model was able to correctly classify 54% of all the images.

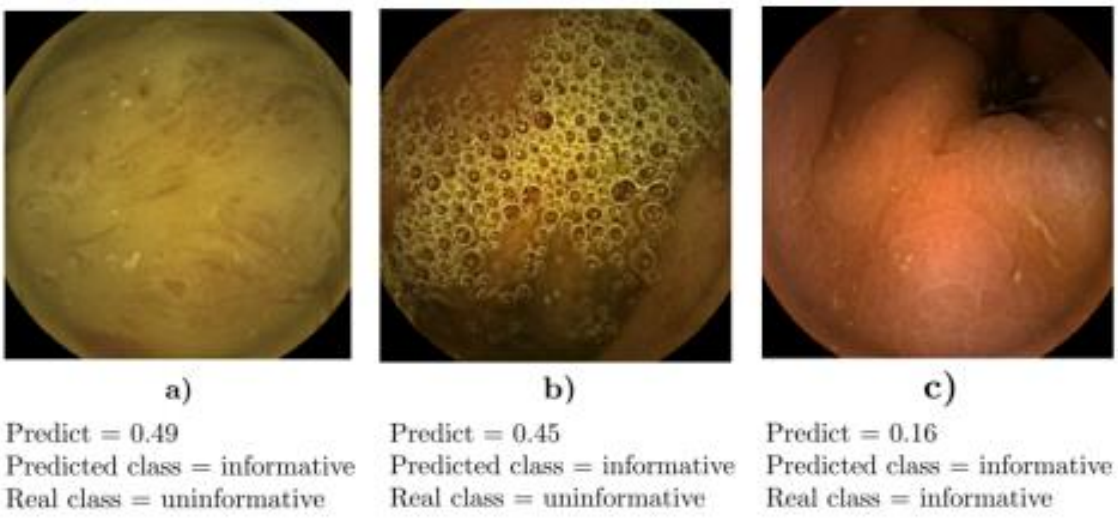

Figure 5.4. Example of images of video v10 classified by the model. Image a) and b) were wrongly classified as informative, while are uninformative. Image c) was correctly classified.

After checking all the images, it was possible to observe that samples identical to a) and b) of Figure 5.4 are still being misclassified. This indicates that the model trained with images from 4 different videos still does not have enough knowledge to correctly classify this type of image.

ROUND 10: Model trained with images from four videos (v2, v5, v7, v6) + test in video v10

It was verified that video 6 had unique samples, and the goal of this round was to evaluate the model's performance when inserting a video that contains unusual content, as can be seen in Figure 5.5.

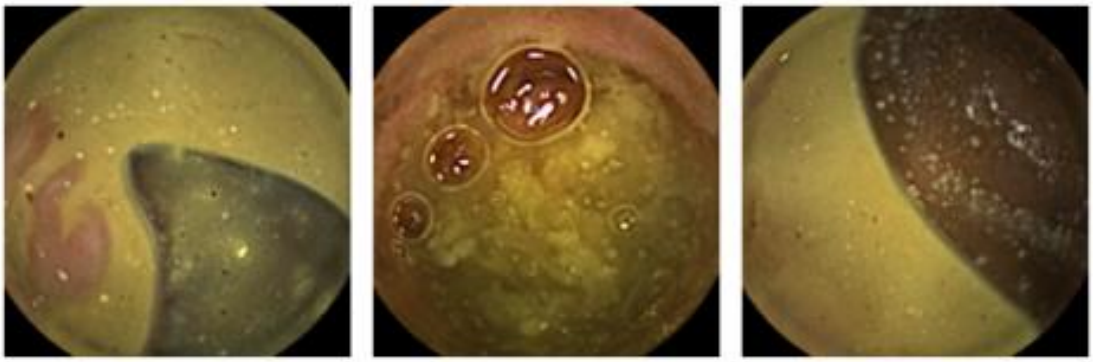

Figure 5.5. Example of the unusual content observed in video v6.

Another Active Learning cycle was executed for video v6, and the model used was the one obtained in round 9, so, in the end, the model contained samples from videos v1, v2, v5, v7 and v6.

Table 5.8. Results obtained from training and validation using video v1 (cleaned) and 800 images from videos v2, v5, v7 and v6, using Least Confidence Sampling.

	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

As can be noticed in Table 5.8, the results of training and validation of this round were not excellent, despite that, they were satisfactory, and then, tests with video v10 were executed. After checking the classification made by the model, it was verified that only 1344 images were classified as informative, and a significant percentage of them were, in reality, uninformative. The accuracy indicates that approximately 41% of the images are correctly classified.

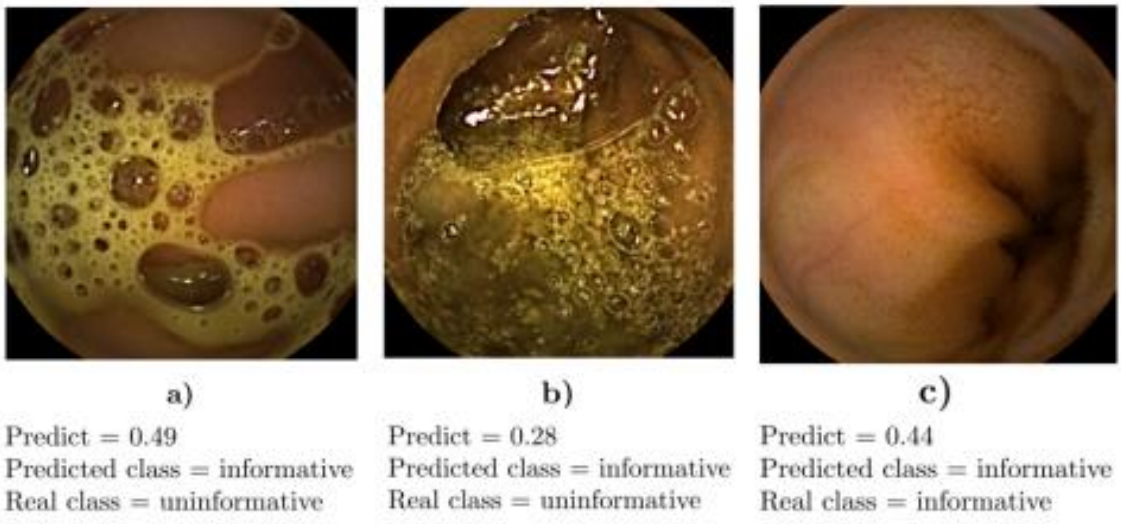

Figure 5.6. Example of images of video v10 classified by the model. Image a) and b) were wrongly classified as informative, while are uninformative. Image c) was correctly classified.

After observing some examples of images classified by the model as informative, in Figure 5.6, we could notice that samples of type b) of round 9 were not abundant, and this can mean that the model has finally learned about that type of samples. However, images from type b) are still common, which means that the model still does not have enough knowledge to learn about these examples.

5.2 DISCUSSION

Through the comparison of different rounds, some conclusions can be produced:

- **Comparing round 3 and round 6**

Introducing Active Learning to the model to classify images into informative and uninformative has significant gains in terms of reducing the time of annotation and the amount of data to label.

- **Comparing round 6 and round 8**

Active Learning can improve the knowledge of a model. However, not all Active Learning techniques are adequate for the study case. In our case, Least Confidence Sampling was more effective in improving the model's knowledge than Margin Sampling, given that the model with Least Confidence Sampling correctly predicted more images.

- **Comparing round 6 and round 9**

After training the model with more images annotated by the oracle, there is a little increase in the number of correct predictions. This indicates that the model's knowledge is rising.

- **Comparing round 9 and round 10**

Even after four Active Learning cycles, the model did not learn about samples with bobbles. This can be because the oracle is annotating incorrectly this type of image and is confusing the model instead of clarifying it. This shows the importance of the oracle in the annotation process.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

In this dissertation, a study of Active Learning to create VCE datasets, in order to solve a binary problem related to the classification between informative and uninformative frames, was made.

A protocol was defined to classify the images into informative and uninformative, with this in mind, Active Learning methods, such as Least Confidence Sampling and Margin Sampling, were used to create an extensive dataset.

To construct the dataset, a set of rounds were executed to implement both Active Learning methods and conclude which one is better for the problem in hands and the impact of Active Learning when only one video and when more than one video is annotated.

Taking this into account, we can infer that Active Learning has great potential in dataset creation, but there are some factors that should be considered when using AL techniques, such as what method is the most appropriate for the problem. The oracle is another determinant element to guarantee the excellent performance of the model.

Given the capacity of Active Learning to efficiently create datasets with more representative information without requiring large amounts of resources, there are some tasks that can be considered for future work.

- Expanding this type of experiment for image classification into a binary problem of pathology/ non-pathology, and consequently, for types of pathologies, can be a way to increase the number of datasets that are available to train the new models.
- At certain point, explore the potential for real-time annotation during the VCE procedure, reducing the post-procedure work.

As a complete set, these kinds of studies are a step forward to increase the confidence of medical experts in this type of classifier, and in the near future, they may be able to have classifiers supporting image classification.

REFERENCES

[1] L. Gueye, S. Yildirim-Yayilgan, F. A. Cheikh, e I. Balasingham, «Automatic detection of colonoscopic anomalies using capsule endoscopy», em *2015 IEEE International Conference on Image Processing (ICIP)*, Quebec City, QC, Canada: IEEE, set. 2015, pp. 1061–1064. doi: 10.1109/ICIP.2015.7350962.

[2] Maria Inês Fernandes Xavier, «Active Learning for Abnormalities Detection on Videos of Endoscopic Capsule», FEUP.

[3] Robert Munro, *Human-in-the-Loop Machine Learning Version 6*. MEAP Edition.

[4] M. Wu, C. Li, e Z. Yao, «Deep Active Learning for Computer Vision Tasks: Methodologies, Applications, and Challenges», *Appl. Sci.*, vol. 12, n.º 16, Art. n.º 16, jan. 2022, doi: 10.3390/app12168103.

[5] H. Alaskar, A. Hussain, N. Al-Aseem, P. Liatsis, e D. Al-Jumeily, «Application of convolutional neural networks for automated ulcer detection in wireless capsule endoscopy images», *Sens. Switz.*, vol. 19, n.º 6, 2019, doi: 10.3390/s19061265.

[6] F. Fonseca, B. Nunes, M. Salgado, e A. Cunha, «Abnormality classification in small datasets of capsule endoscopy images», em *Procedia Computer Science*, 2021, pp. 469–476. doi: 10.1016/j.procs.2021.12.038.

[7] S. Seshamani, R. Kumar, G. Mullin, T. Dassopoulos, e G. Hager, «A Meta Method for Image Matching», *IEEE Trans. Med. Imaging*, vol. 30, pp. 1468–79, fev. 2011, doi: 10.1109/TMI.2011.2119326.

[8] P. H. Smedsrud *et al.*, «Kvasir-Capsule, a video capsule endoscopy dataset», *Sci. Data*, vol. 8, n.º 1, 2021, doi: 10.1038/s41597-021-00920-z.

[9] M. K. Bashar, K. Mori, Y. Suenaga, T. Kitasaka, e Y. Mekada, «Detecting informative frames from wireless capsule endoscopic video using color and texture features», *Lect. Notes Comput. Sci. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma.*, vol. 5242 LNCS, n.º PART 2, pp. 603–610, 2008, doi: 10.1007/978-3-540-85990-1_72.

[10] S. Jain *et al.*, «A deep CNN model for anomaly detection and localization in wireless capsule endoscopy images», *Comput. Biol. Med.*, vol. 137, p. 104789, out. 2021, doi: 10.1016/j.compbiomed.2021.104789.

[11] L. Yang, Y. Zhang, J. Chen, S. Zhang, e D. Z. Chen, «Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation». arXiv, 15 de junho de 2017. Acedido: 2 de setembro de 2022. [Em linha]. Disponível em: http://arxiv.org/abs/1706.04737

[12] P. M. Vieira, C. P. Silva, D. Costa, I. F. Vaz, C. Rolanda, e C. S. Lima, «Automatic Segmentation and Detection of Small Bowel Angioectasias in WCE Images», *Ann. Biomed. Eng.*, vol. 47, n.º 6, pp. 1446–1462, jun. 2019, doi: 10.1007/s10439-019-02248-7.

[13] P. Ren *et al.*, «A Survey of Deep Active Learning», *ACM Comput. Surv.*, vol. 54, n.º 9, 2022, doi: 10.1145/3472291.

[14] M. Hasan e A. K. Roy-Chowdhury, «Context Aware Active Learning of Activity Recognition Models», em *2015 IEEE International Conference on Computer Vision (ICCV)*, dez. 2015, pp. 4543–4551. doi: 10.1109/ICCV.2015.516.

[15] M. Aljabri, M. AlAmir, M. AlGhamdi, M. Abdel-Mottaleb, e F. Collado-Mesa, «Towards a better understanding of annotation tools for medical imaging: a survey», *Multimed. Tools Appl.*, vol. 81, n.º 18, pp. 25877–25911, jul. 2022, doi: 10.1007/s11042-022-12100-1.

[16] X. Li e Y. Guo, «Multi-level Adaptive Active Learning for Scene Classification», em *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, e T. Tuytelaars, Eds., em Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014, pp. 234–249. doi: 10.1007/978-3-319-10584-0_16.

[17] B. Settles, *Active Learning*. Cham: Springer International Publishing, 2012. doi: 10.1007/978-3-031-01560-1.

[18] P. Radeva *et al.*, «Active labeling: Application to wireless endoscopy analysis», em *2012 International Conference on High Performance Computing & Simulation (HPCS)*, jul. 2012, pp. 174–181. doi: 10.1109/HPCSim.2012.6266908.

[19] Andy. K. Devos, S. van Huffel, A. W. Simonetti, M. van der Graaf, A. Heerschap, e L. M. C. Buydens, «Chapter 11 - Classification of Brain Tumours by Pattern Recognition of Magnetic Resonance Imaging and Spectroscopic Data», em *Outcome Prediction in Cancer*, A. F. G. Taktak e A. C. Fisher, Eds., Amsterdam: Elsevier, 2007, pp. 285–318. doi: 10.1016/B978-044452855-1/50013-1.