FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Active Learning for Abnormalities Detection on Videos of Endoscopic Capsule

Maria Inês Fernandes Xavier

Mestrado Integrado em Engenharia Eletrotécnica e de Computadores

Supervisor: António Manuel Trigueiros da Silva Cunha

Co-Supervisors: Tânia Maria Pereira Lopes

Hélder Filipe Pinto de Oliveira

October 30, 2020

Resumo

A introdução da cápsula endoscópica (CE) na área de gastroenterologia tornou possível a observação interna de todo o intestino delgado, de forma indolor e não invasiva. No entanto, em comparação com outros métodos de imagiologia, este exige um tempo de revisão significativamente mais longo por parte dos especialistas. De forma a facilitar este processo, a comunidade científica tem vindo a propor várias estratégias de machine learning para a detecção automática de anormalidades em vídeos CE. Particularmente, métodos mais recentes de deep learning, apresentam grande potencial para a assistência dos especialistas. No entanto, devido ao enorme esforço necessário para anotar cada imagem dos vídeos completos gerados, os algoritmos publicados são treinados e avaliados em datasets muito pequenos. Estes datasets, compostos por imagens com e sem lesões, não são capazes de proporcionar uma base de conhecimento suficientemente informativa sobre o conteúdo encontrado no intestino delgado, e os modelos acabam por ser ineficazes na análise de vídeos inteiros, impedindo a sua integração para uso clínico.

As técnicas de Active Learning (AL) permitem a seleção inteligente de pequenos conjuntos de dados retirados de bases de dados não anotadas, que maximizem a expansão de conhecimento em modelos de aprendizagem máquina, reduzindo de forma significativa a quantidade de instâncias que necessitam de ser anotadas.

Com este trabalho, foram explorados e aplicados métodos de AL para diminuir o esforço de anotação de vídeos completos de CE, compilando datasets mais abrangentes que contribuissem para o treino de modelos de deep learning mais aptos na análise de vídeos completos.

Assim, começou-se por preparar um dataset com 1812 imagens de endoscopias anotadas, para um treino inicial de dois modelos de deep learning, o VGG16 e o DenseNet121. Para além disso, foi organizada uma pool de dados com 272 videos para que os algoritmos de AL selecionassem novos dados a anotar. Estes datasets de treino foram incrementalmente adicionados ao conhecimento dos modelos. A biblioteca modAL foi utilizada para implementar as estratégias de seleção ativa em Python. Estes algoritmos de seleção incluem estratégias de Uncertainty Sampling, aplicados ao conhecimento do modelo VGG16, e Query-By-Comittee, que avaliavam a experiência dos dois modelos treinados.

Como forma de avaliar as plataformas propostas, foram implementadas 5 experiências cada uma com multiplas iterações. Os critérios de avaliação do conteúdo foram, sucessivamente, refinados para se aproximarem mais dos usados em ambientes clínicos. Os melhores resultados foram obtidos na primeira etapa, na qual os critérios eram mais intuitivos, mas não cobriam todos os requisitos dos especialistas. Na étapa 2, a redefinição dos critérios levou a uma reduçação no número de imagens que continham anormalidades de tal forma que os modelos não convergiram. Para mitigar estas adversidades, foram implementadas diferentes estratégias de atualização dos modelos nas etapas 3, 4 e 5, levando ao retreino destes usando técnicas de balanceamento dos sets de dados, com fine-tuning e transfer learning. A última étapa incluiu uma técnica de seleção dos sets que promovia a diversidade entre os elementos do mesmo, juntamente com a métrica de seleção Least Confident sampling, esta provou ser eficaz para maximizar a aprendizagem do modelo.

Abstract

Capsule endoscopy (CE) made it possible to observe the inner lumen of the small intestine in a painless and non-invasive way. This device, which covers the entire gastrointestinal tract, however, in comparison with other imaging methods, comes at the cost of a longer reviewing.

Therefore, to facilitate this process, the scientific community has developed several machine learning strategies to help in the detection of abnormalities in these videos. In more recent approaches, these include deep learning (DL) methods that show great potential to solve this issue. Nonetheless, labelling full CE videos requires a significant effort; hence, there is a lack of labelled data on this medical area. The published algorithms are typically trained and evaluated on small sets of images with and without lesions, ultimately not proving to be efficient when applied to full videos, which does not allow for their clinical integration.

Active Learning (AL) strategies allow for an intelligent selection of datasets from a vast set of unlabelled data, based on the model's knowledge, to maximise its learning. These set selections provide informative batches of data and reduce the cost of annotating full sets.

This dissertation explored AL methods to reduce the CE videos' annotation effort by compiling smaller datasets capable of representing their content, which could improve deep learning methods for its integration in clinical analysis.

For that purpose, a dataset of 1812 images of labelled CE was prepared for initial training of the DL models, VGG16 and DenseNet121. Additionally, a pool of 272 videos was organised to implement the AL selection strategies, to query frames that should be annotated. These sets were incrementally added to the models' knowledge. The modAL library was used to implement the selection strategies in Python. These included Uncertainty Sampling strategies, regarding the VGG16's knowledge, and Query-by-Committee strategies, involving the collective reality perception of both DL trained models.

The AL exploiting was evaluated with 5 experimental sets, in which the content analysis criteria were, gradually, refined to meet the guidelines clinically implemented. The best results were obtained in the first stage, in which the criteria were more intuitive, but did not cover the entirety of the specialists' requirements. In the second stage, the re-defined guidelines resulted in a reduced number of images with abnormalities, which led to the non-convergence of the models. Stages 3, 4 and 5, attempted to mitigate this impact by experimenting with other model updating strategies, which integrated dataset balancing techniques, and fine-tuning and transfer learning approaches. The last stage also included a diversity batch selection strategy to be incorporated alongside the primary AL selection strategy of Least Confident sampling, which showed improvement over the simple informativeness measure.

Acknowledgements

In this section I would like to express my gratitude to all the people who, directly or indirectly, accompanied and backed the elaboration of this dissertation.

First and foremost, I wanted to thank prof. António Cunha, for his guidance, availability and motivation, that enable me to progress so much in these months. I am also very grateful to Tânia Pereira and prof. Hélder Oliveira, for all the support throughout this year. The work developed would also not be possible without the collaboration of Dr^{a} Marta Salgado.

To all my friends that I had the honour to meet thanks FEUP, thank you so much for making these five years so spectacular. Thank you for turning FEUP into a second home, and for always making the work fun and worthwhile. I hope we will find the ultimate lucky Forint to guide us through this new engineer life, and other adventures.

I also wanted to thank the friends that I made throughout my life, that were always there and always will be. I hold a special place in my heart for you, and I would not have made it without you.

Lastly, a special thanks to my family, for always being so patient, loving, and supporting. I am extremely grateful to you, for showing me the brightest things in life and for teaching me so much without leaving out the fun. Thank you for the opportunities you provided me with, and for moulding me into the person I am today.

Inês Xavier

"Where does it all lead? What will become of us?
These were our young questions, and young answers were revealed.
It leads to each other. We become ourselves."

Patti Smith

Contents

Resumo i

Abstract iii

1 Introduction 1
- 1.1 Context . . . 1
- 1.2 Motivation . . . 1
- 1.3 Objectives . . . 2
- 1.4 Contributions . . . 2
- 1.5 Document Structure . . . 3

2 Background 5
- 2.1 Gastrointestinal Tract . . . 5
- 2.2 Imaging of the Gastrointestinal Tract . . . 6
- 2.3 Abnormalities in the Gastrointestinal Tract . . . 7

3 Literature Review 11
- 3.1 Computer-Aided Diagnosis . . . 11
- 3.2 Abnormalities Detection and Characterisation in Videos of Endoscopy Capsule . . . 12
  - 3.2.1 Preprocessing of Video Endoscopy Frames . . . 13
  - 3.2.2 Methods for VEC images Detection and Characterisation . . . 15
- 3.3 Datasets Used in the Literature . . . 20
- 3.4 Active learning . . . 23
  - 3.4.1 Query Scenarios . . . 23
  - 3.4.2 Query Strategies . . . 24
  - 3.4.3 Batch-Mode Active Learning . . . 28
  - 3.4.4 Active Learning in Medical Imaging . . . 29
- 3.5 Summary . . . 29

4 Methodology 31
- 4.1 Datasets Characterisation and Preparation . . . 32
  - 4.1.1 Datasets Characterisation . . . 32
  - 4.1.2 Datasets Preparation . . . 36
- 4.2 Base Model Selection and Initialisation . . . 38
- 4.3 Performance Evaluation . . . 39
- 4.4 Active Selection Strategies . . . 40
  - 4.4.1 Query Strategy . . . 40
  - 4.4.2 Batch Selection Strategy . . . 41
  - 4.4.3 Annotation . . . 42

4.5 Model Updating . . . 43
4.5.1 Fine-tuning . . . 43
4.5.2 Transfer Learning . . . 43
4.6 Summary . . . 44

5 Results and Discussion **45**
5.1 Stage 0 - Base Model Initialisation . . . 48
5.2 Stage 1 - Active Learning and Fine-tuning . . . 48
5.2.1 Results . . . 49
5.2.2 Discussion . . . 50
5.2.3 Criteria redefinition . . . 51
5.3 Stage 2 - Active Learning with Annotation Criteria Correction . . . 51
5.3.1 Fine-tuning . . . 52
5.3.2 Transfer Learning . . . 54
5.3.3 Criteria refinement . . . 57
5.4 Stage 3 - Active Learning and Fine-tuning with Refined Criteria . . . 58
5.4.1 Results . . . 58
5.4.2 Discussion . . . 59
5.5 Stage 4 - Active Learning and Transfer Learning on Increasingly Large Dataset . . . 60
5.5.1 Hyperparameter Selection . . . 61
5.5.2 Results . . . 62
5.5.3 Discussion . . . 63
5.6 Stage 5 - Greedy Selection and Fine-tuning . . . 63
5.6.1 Results . . . 64
5.6.2 Discussion . . . 64
5.7 Summary . . . 65

6 Conclusions and Future Work **67**
6.1 Conclusions . . . 67
6.2 Future Work . . . 68

References **71**

List of Figures

2.1 Small bowel's composition illustration from Encyclopædia Britannica 6
2.2 Vascular Lesions in VEC frames 7
2.3 Inflammatory Lesions from VEC images 8
2.4 Lynfagiectasia, Polyp and Bleeding images retrieved from a private dataset . . . 9

3.1 Vascular disease, VEC image and prepreocessed output 16
3.2 A general overview of the CNN used by Ding et al. 16
3.3 Model for CNN architecture used by Sadasivan and Seelamantula 17
3.4 Segmentation resulting images from Sharif et al.'s work 18
3.5 Representation of the MM-CNN used by Vasilakakis et al. 19
3.6 General pipeline used in active learning strategies. 23

4.1 Pipeline followed in the Methodology. 32
4.2 Abnormalities in the GIANA Dataset 33
4.3 Healthy content in the GIANA Dataset 34
4.4 Abnormalities in the Private Dataset 35

5.1 Results for the first stage 50
5.2 Results for the second stage with Fine-tuning 53
5.3 Results for the second stage with Transfer Learning 56
5.4 Results for the third stage with Fine-tuning 59
5.5 Results for the forth stage with Transfer Learning 63
5.6 Results for the fifth stage with Diversity sampling 64

List of Tables

3.1 Characterisation of the Datasets used in Publications. 21

4.1 Used Datasets Characterisation. 33

5.1 Datasets Selection and Transformations. 46
5.2 Models' Updating and Testing approaches. 47
5.3 Hyperparameters explored for Transfer Learning. 61
5.4 Training approach selected for Transfer Learning. 61

Abbreviations

AL	Active Learning
AUC	Area Under the Curve
BoW	Bag-of-visual-Words
CAD	Computer-Aided Diagnosis
CE	Capsule Endoscopy
CT	Computed Tomography
CNNs	Convolutional Neural Networks
DL	Deep Learning
fc-CNN	fully connected CNN
FN	False Negatives
FP	False Positives
GI	Gastrointestinal Tract
QBC	Query-by-Committee
LMSCBs	Large-Medium-Small Convolution Blocks
LBP	Local Binary Pattern
LC	Least Confident
ML	Machine Learning
MM-CNN	Multi-label Multi-scale CNN
MRI	Magnetic Resonance Imaging
MSE	Mean-Squared Error
SB	Small Bowel
SGD	Stochastic Gradient Descent
TN	True Negatives
TP	True Positives
US	Ultrasound
VEC	Videos of Endoscopic Capsules

Chapter 1

Introduction

1.1 Context

The gastrointestinal (GI) tract is a carefully studied organ since it can harbour severe conditions, such as colorectal cancer, the third most deadly cancer in the world [1]. Studies in GI cancer showed that early detection is correlated with a significant decrease in dangerous developments of lesions, and improvement in the survival rates [2].

Upper and lower GI endoscopy allow the visualisation of a substantial portion of the GI tract, however, the small bowel (SB) was seen as the "black box" of the GI tract, due to its hard reach for surveillance [3]. With the SB being home to multiple pathologies such as Crohn's disease or obscure gastrointestinal bleeding [3, 4], there is a clear need for taking advantage of novel screening tests for the detection of SB diseases and its abnormalities.

The wireless capsule endoscopy (CE), introduced in 2000, 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. It is considered the preferred method for the diagnosis of diseases of the SB, given its ability to cover the whole GI tract, including this area of difficult access, and its notable inner imaging results compared to other methods for abnormalities visualisation, such as in Crohn's disease in patients [5].

Despite this, considering the resulting eight-hour-long recorded video, the main disadvantage with this procedure resides in a long (40−60 minutes), monotonous and susceptible to human error reviewing of the screening by experts [6]. Thus the interest for developing techniques for the automatic detection and diagnosis of lesions to save time, and aid gastroenterologists on the examining of the videos of endoscopic capsules (VEC).

1.2 Motivation

The work on detection and characterisation of GI tract lesions is extensive, and it covers the most common SB threats, which are vascular and inflammatory lesions, lymphangiectasias, polyps and bleeding. The early detection of these abnormalities can lead to significant improvement of the

patient's health condition, and avoid complications, such as the development of ulcerative colitis, Crohn's disease or cancer [7].

The strategies developed include the usage of machine learning (ML) algorithms, with deep learning (DL) strategies being the most successful among them. Given the difficulty in the frame-wise annotation of full videos, there is a shortage of available large data sets with labelled VEC data. Most of the explored work apply the techniques to small sets of images, with and without abnormalities, but ultimately the resulting learners fail when performing over full-length videos [8, 9]. This suggests that the datasets used for training lack in representativity of the full VEC reality for the networks to generalise correctly. Thus, the existing work on detection and identification of abnormalities in VEC still lacks in performance for its clinical application.

Active learning (AL) consists of examining the model's familiarity with a set of unlabelled data and querying the elements that are expected to maximise the model's knowledge over the whole dataset. The informativeness of the samples can be evaluated using several techniques and allows for a reduction of labelling cost. These methods, to the best of our knowledge, are yet to be explored in the selection of unlabelled frames from entire CE videos and can be beneficial to employ cost-efficient querying of representative data from the VEC reality.

1.3 Objectives

Given the lack of available datasets which cover the VEC reality and the fastidious work of revising and frame-wise annotating full videos, the main goal of this thesis is to experiment with AL strategies over full unlabelled VEC. The explored approaches query samples that can represent the overall distribution of the data in the videos and improve the efficiency of the classifiers with fewer labelled samples [10].

The work incorporates a labelled publicly available dataset, included in the training of DL models, to provide an initial familiarity of the content in VEC. Followed by the exploration of AL selection strategies over an unlabelled pool of videos, to expand the knowledge of the trained models, and provide as improved analysis over full videos.

1.4 Contributions

The present dissertation attains the following contributions:

- An investigation on the viability of integrating AL strategies to select relevant data to train DL approaches for VEC abnormalities detection;
- Development of different VEC abnormalities detection approaches, including model updating strategies with Fine-tuning and Transfer Learning;
- Analysis of complete CE videos' structure for the preparation of an organised dataset for application in DL strategies;

- Development of VEC annotation criteria with three evolving levels of medical standard.

1.5 Document Structure

This dissertation is organised into 6 chapters, starting with the present Chapter, which briefly introduces the importance of CE and the difficulties posed by VEC analysis. It also presents the proposed strategy to tackle the issues regarding the clinical performance of state-of-the-art VEC abnormalities detection approaches.

A second chapter examines the clinical concepts regarding the SB, its imaging and the abnormal content which can be found. Chapter 3 looks into the existing VEC diagnosis platforms and abnormalities detection and characterisation works that were developed. The third chapter also reviews AL strategies commonly used and applied in the medical field, finishing with an overview of the datasets typically used in the CE exploitation strategies.

Chapter 4 explains the steps followed to implement the proposed solution, and Chapter 5 exposes the stages of the developed work and the corresponding results followed by their conclusions. This document finishes in the sixth chapter, which contains conclusions regarding this dissertation's development and future work considerations.

Chapter 2

Background

This chapter will introduce the issue under study. It will be divided into three sections, starting with the explanation of concepts regarding the medical field, along with a further exploration of the techniques developed for examination, and the abnormalities commonly found in the SB.

2.1 Gastrointestinal Tract

The GI tract is responsible for the decomposition of the food ingested into the nutrients and minerals needed for maintaining the body's energy levels, growth and repair of cells. The SB (or small intestine), roughly located in between the stomach and the colon, is the portion of the GI tract held responsible for most of the nourishment absorption [11].

The lumen, the inner cavity of the whole SB, is coated in 4 layers: mucosa, submucosa, muscularis propria (muscular layer), and adventitia (serosa) [12, 11], as it is represented in 2.1.

Facing the lumen surface is the mucosa, primarily responsible for the direct absorption of nutrients and regulation of the intestinal flora. The absorption cells, enterocytes, are contained in the villi, small folds that are filled with capillary and lymphatic vessels. Following the mucosa, the submucosa is a connective tissue layer constituted with a vast network of arteries and lymphatic tissue with a supportive role in the absorption [11].

Peristalsis of the SB is made possible in the muscularis propria, composed of two layers of muscle that contract and relax the tract to propagate the sustenance in a wave down the SB. Finally, the serosa is responsible for the connecting of the tissue and protection of the tract [12, 11].

The SB tract, illustrated in 2.1, with approximately 3 to 5 meters in length, can be parted into three great consecutive regions, them being the Duodenum, the Jejunum and the Ileum [12]. Starting at the pylorus with the duodenal bulb the duodenum rapidly starts descending into the major and minor duodenal papilla, responsible for receiving bile and pancreatic ducts. The ligament of Treitz then marks the entry to the jejunum, that after approximately 2.5m, transitions to the ileum without an anatomic delineation. Thicker lumen and more prominent mucosal folds, distinguish the jejunum from the ileum, for the absorption of any nutrients that got past the jejunum, ending in the ileocecal valve responsible for isolating colonic contents from the SB [12, 11].

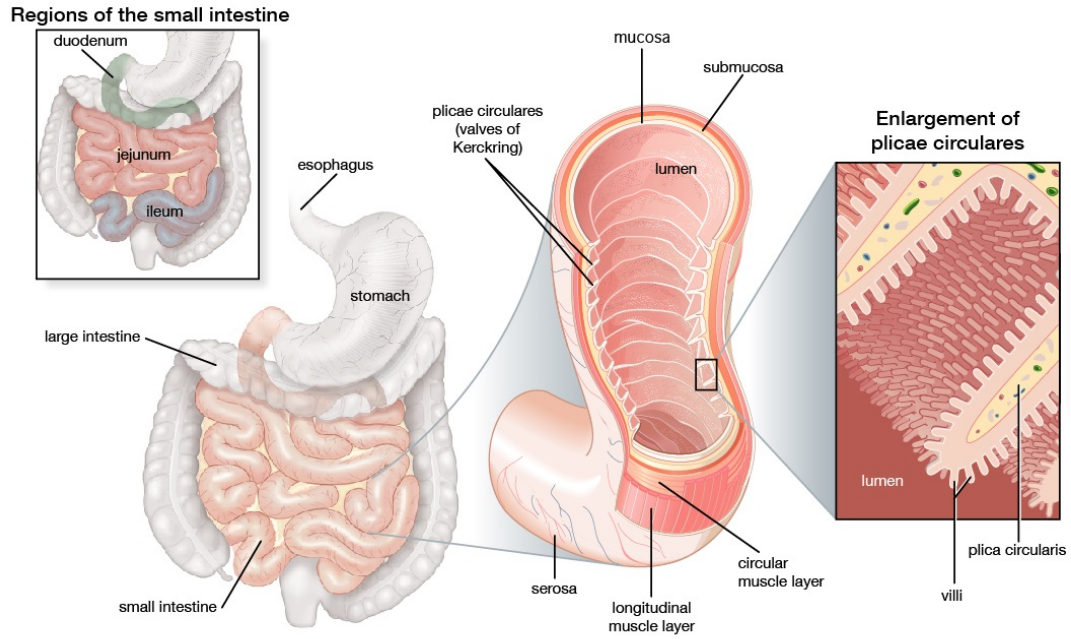


Figure 2.1: Small bowel's composition illustration from Encyclopædia Britannica©

2.2 Imaging of the Gastrointestinal Tract

The GI's examination is essential to diagnose potential pathologies in this vital organ. Along with the traditional endoscopy approaches (that includes esophagogastroduodenoscopy and colonoscopy), ultrasound (US), computed tomography (CT) and magnetic resonance imaging (MRI) also serve as conventional methodologies for its visual analysis.

US can be used for intraluminal imaging, and provides high-resolution images of soft tissue, being a good alternative for the detection of GI tract inflammations. Still, despite also giving information on the intestinal wall and its surroundings, the quality of the images can be conditioned due to intestinal gas, artefacts can be challenging to identify and complete visualization of the intestine is not always possible [13].

Multidetector CT is considered a valuable tool in the study of intestinal loops, allowing multiplanar reconstructions, and total evaluation of the intestine and its surroundings. It is particularly useful in the detection of inflammatory and neoplastic intestinal lesions, yet it still uses contrast and is highly exposed to radiation, which can lead to health complications [13].

On the other hand, MRIs also provides full coverage of the tract, without the exposure to radiation, and is superior for soft tissues imaging. However, it is still susceptible to poorly identified artefacts, due to intestinal movements, and potential long term effects with the usage of contrast [13].

With endoscopes, it is possible to visualize the GI mucosa directly, and even intervene for biopsies, polypectomy or endoscopic surgery; however, the device is invasive, and through the standard endoscopy methods it is impossible to reach the SB [13]. For accessing this organ,

enteroscopy procedures have been used. Yet, the best rate for total enteroscopy still varies from 45% to 86%, given its limited access through full oral, or combined oral and anal intervention, which makes this technique preferable for SB therapeutic approaches [14].

The endoscopic Capsule was introduced in 2001 and, due to its success, the original device already has a third-generation, PillCam^{TM} SB3, and evolved with different adaptations that vary in features such as size, resolution and battery life [15].

This method consists of a pill-shaped and sized device with an incorporated endoscope camera that is propelled through the digestive system by peristaltic movements. It communicates with a recording device worn by the patient saving, on average, up to 8hs of screening, corresponding to the battery life of the capsule [5].

Experts typically resort to this method as an assistance to the diagnosis of different situations or clarify dubious results found in MRI's, since it allows direct visualisation of the mucosa, wherein 80% of the cases the expected pathologies can be confirmed using VEC [3]. This method is typically used for the identification of lesions that can be the cause of iron deficiency anaemia or to verify cases of obscure GI bleeding detected in other endoscopies, which have an 80% probability of occurrence in the SB [15, 4]. The suspicion of Crohn's disease incidences in this section of the GI tract, or to verify the evolution of existing polyps are also reasons to turn to this method [3, 5, 4].

The main disadvantage with resorting to this method is the fact that it relies on an expert reviewer for the analysis of the findings, subjected to a reading of 40 to 60 min varying on the SB transit, which requires a considerable amount of concentration and patience [6].

2.3 Abnormalities in the Gastrointestinal Tract

With the increasing undergoing of VEC analysis to prevent and confirm complicated health conditions, it is common to find several types of lesions. These can be roughly grouped into inflammatory and vascular lesions, lymphangiectasia, polyps and bleeding. Bellow the visual characterisation and symptoms of the lesions will be presented.

(a) Angiectasia/Angiodysplasia (b) Erythematous Patch (c) Phlebectasia

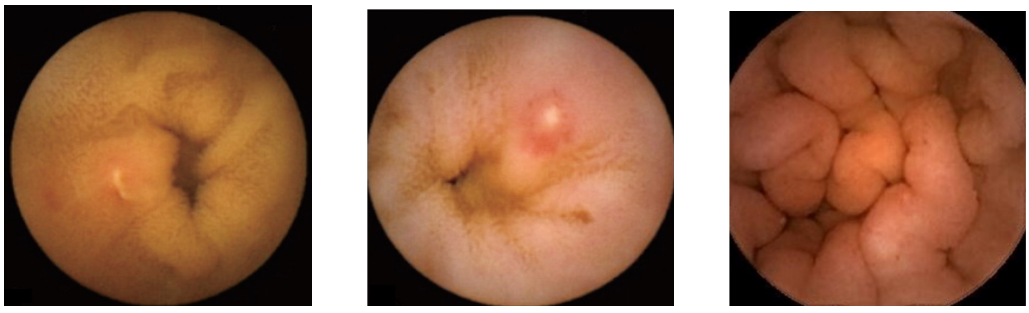
Figure 2.2: Vascular Lesions in VEC frames retrieved from [16]

Vascular lesions are distinguished for their colouration on the villi. The most commonly found lesions regarding bleeding in the SB include Angioectasias, angiodysplasias, or arteriovenous malformations [17, 18].

The terms Angiectasia or Angiodysplasia can be used to define the lesion in the Figure 2.2a, where the second term tends to be referred to its more evident form, with bright-red colouration, consisting of capillary dilatation and presenting a vessel appearance. Diminute Angiectasias and Red Spots are also similarly coloured lesions, differentiated by more delineated and punctuated shaped abnormalities with also reduced size [16].

Erythematous Patch in Figure 2.2b is a more subtle evidence of vascular abnormalities, consisting of a reddish area appearance, surrounded by intestinal villi [16].

More distinguishable, the bluish colouration on Phlebectasias present in 2.2c is the consequence venous dilatation running below the mucosa, giving it also a flat to slightly-elevated shape [16].



(a) Erosion image retrieved from [19]

(b) Ulcer image retrieved from [19]

(c) Oedema image from the GIANA[1] dataset

Figure 2.3: Inflammatory Lesions from VEC images

Inflammatory lesions in the SB can be translated into complications such as Crohn's disease or celiac disease. Common findings count with erosions, ulcers, and oedema in the mucosa [18].

Ulcers [20] and erosions are very similar openings in the mucosal layer caused by inflammations in the tract, presented in 2.3b and 2.3a, respectively.

Mucosal oedema, shown in 2.3c is thick and swollen mucosa, which is pooled with fluid, leading to thickened haustral folds. In severe oedema, the mucosal narrowing can be found [21].

[1]Gastrointestinal Image Analysis - Endoscopic Vision Challenge - "https://giana.grand-challenge.org/"

(a) Lynphagiectasia (b) Polyp (c) Bleeding

Figure 2.4: Lynfagiectasia, Polyp and Bleeding images retrieved from a private dataset

Lymphangiectasia is an intestinal reaction from the villi, where it dilates, as it can be visualized in Figure 2.4a, provoking nutrients absorption deficiency. One can identify it by its yellowish colouration in its projections [22].

Polyps, shown in Figure 2.4b, are inflamed or eroded projected lesions, that rise above the plane of the mucosal surface [20].

Bleeding can be a consequence of several lesions in the SB, including vascular lesions [17], and other diseases [23]. Observed in Figure 2.4c, it is worth locating to trace its origin and possible complications associated with it.

Chapter 3

Literature Review

The procedure for reviewing VEC frames in search of abnormalities is exhausting and susceptible to human error. Platforms for visualising the resulting CE videos do not provide much support to their analysis, so ML strategies were developed to assist experts in this field with this time-consuming task. The ML techniques, however, fail to generalise for this assistance to diagnosis to be included in clinical practice. This is due to a lack of representativity in the datasets that provide the VEC content. For that purpose, it should be interesting to analyse methods that can optimise the learning procedure by selecting more rich samples to represent the VEC reality.

Therefore, this chapter will be parted into five sections, a first will introduce the computer-aided diagnosis (CAD) applications used in the study of VEC, followed by the developed work on the aiding the practitioners in automatic abnormalities detection and characterisation. A third section will present the public datasets which are typically used for the latter's application, and a fourth section will analyse active learning (AL) strategies that can tackle these datasets limitations. The final section will summarise all the topics covered.

3.1 Computer-Aided Diagnosis

CADs are growing tools in the medical field, which assist the specialist in analysing medical content to support in diagnostics or assist in deciding on adequate treatment approaches. The interest in this subject in VEC reviewing comes from the excessive amount of content that needs to be analysed by specialists, as well as the advantage that computer imaging can provide over regular human surveillance. Thus some support should account for faster and improved diagnosis [24].

This technological implementation of CAD has been applied to other endoscopic imaging methods [24] and explored for application in images from VEC. However, the success drawn from the latter has not met the expectations given its lack of application in full videos [9]. Nonetheless, this section will introduce the functionalities provided by CE's reviewing platforms, to aid in the VEC reviewing.

The Medtronic company, the distributor of the PillCam™, provides a PillCam™ Reader application for the reviewing of the videos called the RAPID reader. The corporation has been investing in assisting the viewer by integrating assistance functionalities to its software. These include image enhancement through increased contrast or blue filtering the image, the latter having been found useful for the confirmation of vascular lesions. Previous versions came with a suspected blood indicator functionality, that highlighted bleeding and potentially bleeding frames. However, this reading mode proved to merely achieve a sensitivity of 55.3% and a specificity of 57.8% [25].

The Smart QuickView software in the RAPID application also allowed the user to view solely the images that potentially had relevant areas. Nevertheless, results proved to be discordant with conventional human reading in 28.3% of cases [25].

Currently, their most recent software, PillCam™ Reader Software v9[2], provides image contrast levelling and regulative blue filtering. It also includes a detection algorithm named "Top 100" that automatically maps the top 100 most relevant study images. However, this update is found to be equally unreliable, given its high dependability in the detection of red spots, which accounts for a high rate of irrelevant information. In addition to this, it is also limiting since it only detects the first 100 defections starting on a chosen frame.

Therefore, there is still plenty of room for improvement in the development of a properly timesaving, functional and polished CAD.

3.2 Abnormalities Detection and Characterisation in Videos of Endoscopy Capsule

For detecting and characterising lesions in VEC frames, several strategies and algorithms were adopted. However, an overall structure is maintained throughout the methods, involving two main stages, a preprocess of the VEC frames followed by abnormalities detection.

Preprocessing techniques vary concerning the abnormalities that they ought to detect, enhancing features in the frames that can result in its better detection [8]. Abnormalities detection count with different approaches, where some specify on identifying a specific abnormality [23], and others count with the detection of several abnormalities [25, 26]. These detection methods sometimes perform segmentation and classification as well, and other times characterisation is conducted in different subsequent steps.

Several studies have taken on the challenge to aid the work of a specialist in finding content as quickly and as unmistakably as possible. Among the work on VEC images' characterisation and detection, a reasonable quantity is based on supervised learning with DL, with Convolutional Neural Networks (CNNs) being the most popular used approaches [27]. Nevertheless, given the fastidious work on labelling the VEC frames, and the CNNs great on dependability on the amount of data, this section will also look into weakly-supervised learning strategies.

[2]PillCam™ Reader Software v9 Review - "https://www.medtronic.com/covidien/en-us/support/software/gastrointestinal-products/pillcam-software-v9.html"

3.2.1 Preprocessing of Video Endoscopy Frames

What distinguishes human visual analysis is the innate perception of multiple features in an image and its association, which allows the brain to recognise it. Thereby, the translation of human perception of the semantic of a VEC frame can be potentiated by the enhancement of relevant features that are inherent to its diagnosis.

The CE capsule, with its reduced size, has limitations in the quantity and quality of its components, so it still cannot live up the image quality of conventional endoscopy [27]. It is also subjected to low-quality lighting, and its peristalsis propulsion contributes to very irregularly lighted images. Therefore, this imaging exam could especially benefit from image processing to improve the exhibition of its content [27].

The abnormalities that can be found in the SB are distinguishable through colour, shape, and texture, so these can serve as a cue for topographic location in the current image [27]. Consequently, these are features that are typically emphasised in VEC frame processing. Thus an approach of each of these explorations will be presented.

Colour-based enhancement

Different colour space manipulation is a common approach for the analysis of VEC images, given the distinct colour characterisation of some of the lesions, as presented in the background section on abnormalities 2.3. The Lab or CIELab colour space is a transformation that is often used [23, 26, 28, 29]. It is characterised by three channels, L, α and β, L containing the achromatic luminance channel, and the α and β chromatic channels varying between green (-) to red (+), blue (-) to yellow (+) respectively, and it was designed to approximate human vision's behaviour [21].

In the case of Pan et al. [23], when transforming the RGB formatted frames to the CIELab colourimetric system, the large colour difference between two pixels is translated into a great distance, providing a more significant distinction between colour. After this transformation, a covariance matrix is calculated to increase the Euler distance between different coloured pixels. The bleeding areas are then distinguished using a threshold value.

Texture-based enhancement

This characteristic for feature analysis in VEC is explored in the paper developed by Li et al. [30] that uses texture to spot ulcer sections in frames. Pixel intensity spatial variation in the neighbourhood of this abnormality's region is explored by starting with the signal processing method, continuous curvelet transform. In this phase, a spacial grid is used for the translation of curvelets in each scale and angle obtaining a U_j product after applying a 2D Fourier transform. The resulting product is then wrapped around the origin and goes through the inverse 2D fast Fourier transform returning discrete curvelet coefficients.

The second texture-based enhancement in the work of Li et al. [30], involves Local Binary Pattern (LBP), in which each pixel in the frame is labelled by thresholding the difference between

the central intensity value and its eighth surrounding neighbours. This thresholding step acknowledges the surrounding pixels that have equal or higher intensity value, and these are then multiplied by binomial wights given to the corresponding pixels, the values of these are finally summed to obtain the LBP value of its neighbourhood. The final step is a 256-bin LBP computation over the region for texture description.

Shape-based enhancement

In this particular case, some preprocessing can be applied in VEC frames for the filtration of non-informative content, such as when a frame only has intestinal fluids with bubble-like regions. Wang et al. [31] seek to resolve this problem by automatically detecting the frames using an RSS filter constructed by eigenvalues of the Hessian matrix. Starting by using the Hessian matrix as a shape filter, where the Hessian matrix is the second-order partial differential of the image density function f. The density is partially differentiated twice in the x and y directions, to create a squared matrix H:

$$H = \begin{bmatrix} f_{xx} f_{xy} \\ f_{yx} f_{yy} \end{bmatrix} \tag{3.1}$$

With the f's index expressing the direction of the first and second derivative respectively, expressing the variation in these four directions. Two eigenvalues are then retrieved to represent the geometric implications of these variations so that the approximate shape of a local curve on its gradient direction is oriented on the λ_1 and λ_2 eigenvalues as length and long axis.

Then the bubble-like the RSS filtered is constructed on functions to make the correspondence between a line and a circle, by analysing the features of the points' eigenvalues relation, creating a p_{line} and p_{circle} probability functions, where the shape is defined according to the proximity to 1 in both functions.

$$p_{line} = \frac{|\lambda_1| - |\lambda_2|}{|\lambda_1|} \tag{3.2}$$

$$p_{circle} = \frac{|\lambda_2|}{|\lambda_1|} \tag{3.3}$$

This step is followed by an analogous approach for group eigenvalue functions also to describe the various degrees of similar shapes accurately. The shape-selective filters are then obtained for each form represented by the products of the group and point of each shape, originating a z_{ring} resulting filter.

$$g_{line} = |\lambda_1| \tag{3.4}$$

$$g_{circle} = |\lambda_2| \tag{3.5}$$

$$p_{ring} = \frac{|\lambda_1 - \lambda_2| \cdot |\lambda_2|^2}{|\lambda_1|} \tag{3.6}$$

In sum, each frame is processed by the RSS filtering considered light pixels, and bubble sections are taken into account according to an adaptive threshold. The images with a bubble area ratio of 30% or higher are distinguished as non-informative frames.

3.2.2 Methods for VEC images Detection and Characterisation

The present section will look into the procedures that have been developed to detect and characterise abnormalities in VEC frames automatically. It will explore several techniques applied to multiple lesion detection, its characterisation and explore strategies regarding weakly labelled data.

Multiple Lesion detection

In the field of lesion detection in VEC, most approaches are limited to only a few specific lesions, where the most successful were the ones performing on the search of one particular abnormality [27].

For multiple abnormalities detection, most of the work that has been developed binarily classifies VEC into normal and abnormal frames. Nonetheless, this section will start by exposing work produced on the investigation of the ideal CNN-based strategy for detection and characterisation of inflammatory and vascular lesion. Two recent approaches that successfully cover more than nine lesion types will additionally be explained in this section, as well as a distinct method that focuses on feature extraction for differentiating bleeding and ulcer occurrences in the SB.

To start with, the work developed by Teresa et al.[20] included the exploitation of preprocessing algorithms, looked into the most appropriate CNNs for the task of inflammatory and vascular lesion detection and also methods for lesion segmentation.

Although they experimented with Adaptive contrast diffusion, Homomorphic filtering and Automated Multi-Scale Retinex with Colour Restoration (MSRCR), the best detection results were obtained using VGG16 and without preprocessing, reaching a precision of 0.95. In segmentation, vascular lesions were better isolated without preprocessing and using U-net with a Jaccard coefficient of 0.821 and a Dice of 0.902, whereas GAN was better succeeded using Automated MSRCR in inflammatory lesion segmentation, achieving a Jaccard of 0.602 and a Dice of 0.751. The latter result proves that in segmentation of vascular lesions, given its distinctive colouration, preprocessing is not as essential as it is for inflammatory lesions.

Despite the significant positive results of the developed work, these models did not succeed when applied to 449 VEC videos from a private dataset. Therefore, it would be of interest to either train the model on some of the full VEC for an update of the model's parameters on more general information or reevaluating the methodologies for reducing the overfitting problem.

Ding et al. [25] applied a CNN-based auxiliary reading model to a dataset of 6970 patients, distinguishing normal and abnormal images, with the latter including inflammations, ulcers, polyps,

lymphangiectasias, bleeding, vascular diseases, protruding lesions, lymphatic follicular hyperplasias, diverticulums, and parasites.

Using the ImageNet, they started by training a model for preprocessing the VEC data. After 1970 VEC videos are put through this first step, as shown in Figure 3.1, they serve as input into a ResNet model represented in Figure 3.2 for training. This high number of training examples was chosen to have 1000 representative images for each abnormal category. The CNN then outputs a result according to the likelihood of containing an abnormality based on a probability threshold value.

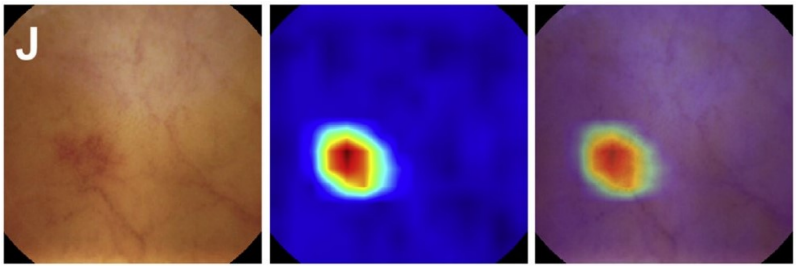


Figure 3.1: Vascular disease, VEC image and prepreocessed output from [25].

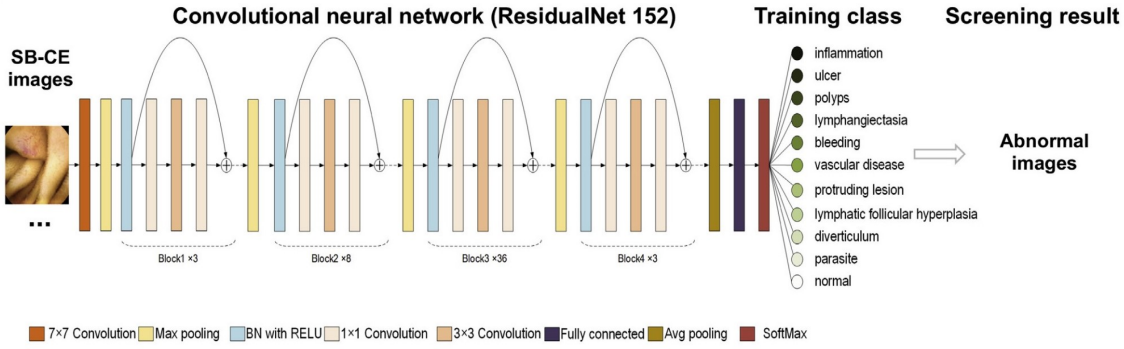


Figure 3.2: A general overview of the CNN used by Ding et al.[25].

The validation included 5000 patients, in which the cases were read by both conventional reading and the CNN-based supplemental reading by 20 gastroenterologists, that verified the output filtered abnormal images from the CNN. The comparison of both methods in sensitivity improved from 76.89% to 99.90% with the CNN-based auxiliary reading, and the time for analysing the screenings was reduced from 96.6 minutes to 5.9 minutes, which significantly tackles the issue regarding video analysis by experts.

The most differentiating aspect of Ding et al.'s [25] study is that the resulting CNN-based auxiliary CE reading group achieved a rate of 70.91% per-lesion detection compared to conventional reading which reached 54.57%, proving that CAD can effectively improve the analysis performance.

By contrast, in the paper developed by Sadasivan and Seelamantula [26], they were able to perform multiclass characterisation, as well as abnormality detection, achieving a very satisfying

result, with a minimum AUC (area under receiver-operating-characteristic curve) of 92.36% in the detection of polypoids. This work, although, used a very reduced dataset, with a total of 137 images, and to tackle this issue employed a patch-based approach of 100 patches per image, translating into a total of 13700, as well as performed data augmentation, through rotation and reflection.

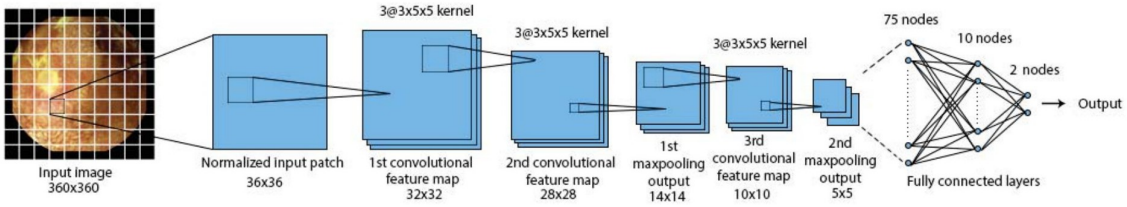


Figure 3.3: Model for CNN architecture used in [26].

A CNN presented in Figure 3.3 was trained by inputting 36 by 36 normalised patches by three colour channels per image, that suffer a zero-padding from 320x320 to 360x360 for improved learning of the background. The normalisation before the inputting to the max-pooping layer speeds up the conversion of the model and alleviates local neighbourhood accentuated responses. ReLu is employed in the activation for both convolutional and fully-connected layers, except in the last layer which evaluates the scores into class probabilities. Classification is performed by a perceptron network structure 75-10-2, then resulting in an output fully-connected hidden layer with a dropout keep probability of 0.90 for regularisation of the network. To avoid overfitting, they employed early stopping, and a loss function of the square of softmax cross-entropy with logits was used. An overall sensibility of 94.95% and specificity of 97.72% was achieved through this procedure in abnormality detection of multiclass detection.

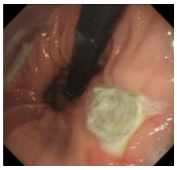
The only employed enhancement involved the conversion to the CIE-Lab colour space, that has been commonly used for better extraction of colour (as is presented in 3.2.1) and texture [26]. Still, their reflection on the work developed states that better bleeding detection failed to be achieved due to its lack of texture information, which is odd, given its higher detectability with colour enhancement through the L*a*b colour space, as is studied in [23].

In the work of Sharif et al. [28] the detection of abnormal areas in frames is used to remove healthy regions from the analysed VEC frames, and to extract the geometric features from the abnormal areas. A fusion of a CNN feature classifier and the extracted geometric features allows for a consequent classification of the frames into healthy, bleeding and ulcer frames with a K nearest neighbour algorithm.

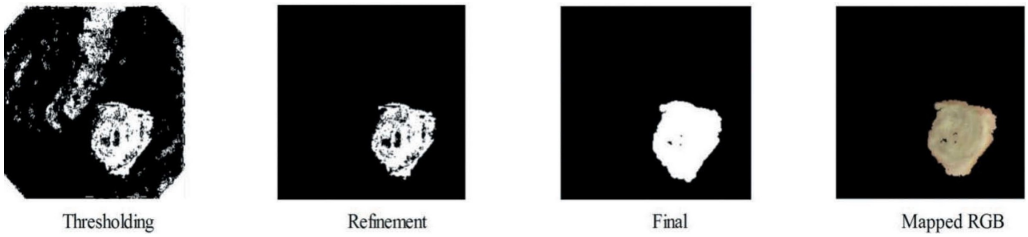
For the abnormality detection task, once again, the L*a*b colour space is explored over median filtered RGB image channels, to extract mean, variance, standard deviation, skewness and kurtosis as colour features. These are combined in a mean deviation matrix corresponding to a weight value for clustering similar valued weighted pixels. The features are extracted based on the resulting clusters, as normal and abnormal, through a probability-based obtained threshold value. Morphological operations (dilation and filling) are finally applied to obtain a segmented section.

The area, solidity, major axis length, minor axis length, orientation, eccentricity, height and width geometrical features are extracted to be combined in the matrix that is used for feature fusion.

Bellow is the original and resulting isolated deceased area mask 3.4, providing yet another strategy for multi-classification, despite including the detection of less lesion diversity as the previously presented strategies.



(a) Lesion example used for segmentation from [28].



(b) Thresholding refinement and resulting segmentation from [28].

Figure 3.4: Segmentation resulting images from Sharif et al.'s work [28].

Weakly-supervised learning in multiple lesion detection

The methods using weakly-supervised learning are indicated for datasets with frame annotations of the identified lesions, instead of pixel-wise pinpointing. Despite lacking localisation information on the occurrences of lesions, the annotation cost of these frames is inferior, given the lower demand that this labelling strategy requires.

The work of Iakovidis et al. [32] applies a weakly-supervised convolutional neural network for the binary classification of the frames, employing cluster unification for the localisation of GI anomalies. An approach that has proven that pixel-wise annotations are unnecessary for a model to learn and locate abnormalities in a VEC accurately. Besides using a VEC dataset, a colonoscopic images database was also integrated, making it more widely applicable.

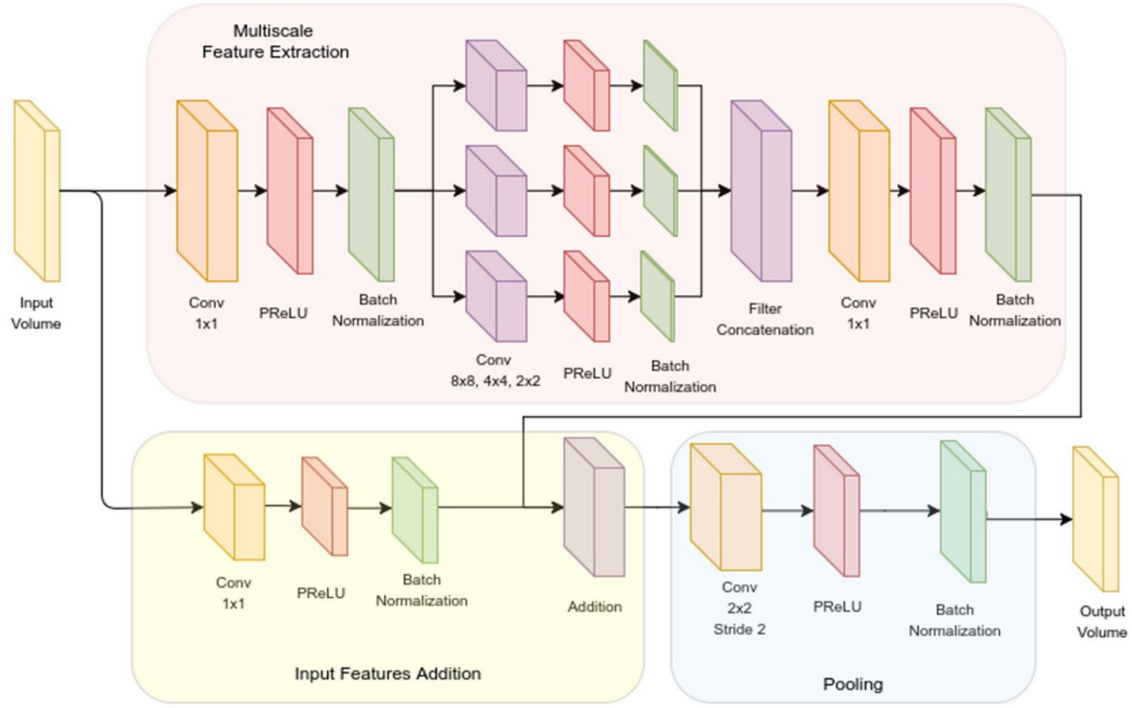
By comparison of accuracy, sensitivity and specificity in classification, with conventional learning algorithms, achieving sensitivity results of 93% and 92.4% for each database was the only parameter in which their approach did not exceed the compared methods. This is due to the patch usage of the CNNs that applied different proportions of abnormal to normal patches, so smaller lesions are less likely to be missed.

An example of weakly-supervised learning this implementation can be found in the work of Vasilakakis et al. [29], which presents a strategy based on saliency-enabled Bag-of-visual-Words (BoW) and another CNN approach using weakly-supervised learning strategies. This work focuses explicitly in expressing the diversity in the VEC image's content, by seeking to distinguish the

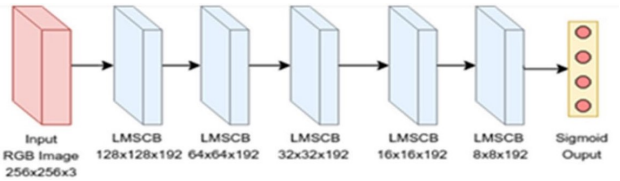
semantics in normal and abnormal frames clearly; thus their goal is to be able to label one frame with multiple findings through multi-label classification.

The content in the frame can be composed of multiple normal and abnormal content, with each being semantically explicit given that normal content includes categories such as normal mucosa, bubbles or intestinal debris, and abnormal information counts with a diversity of lesions that are typically found in VEC analysis, including inflammatory lesions and vascular lesions.

To begin with, in the BoW methodology, Vasilakakis et al. [29] focus on salient point detection, using the difference of maxima algorithm over the CIE-Lab colour space resulting image. It looks for the points in which the channel values diverge the most, compared to the mean euclidian distances between all the *a* channel values, in order to pinpoint the regions where salient points can be found. Two different sized patches are formed around the salience points.



(a) LMSCBs structure scheme



(b) MM-CNN with LMSCBs representation

Figure 3.5: Representation of the MM-CNN used by Vasilakakis et al. [29].

From the extracted patches, each of the three-channel Lab values from the point, and the max and min values of the patch allows for the creation of a 9-dimension colour feature vector associated with each patch. These patches are then clustered and translated into visual words;

hence, the BoW vocabulary associated with the frame is formed. Finally using a histogram, the frequency of each word is measured regarding the video under study.

In this research [29] a multi-label multi-scale CNN (MM-CNN) is also developed, composed of five large-medium-small convolution blocks (LMSCBs) as shown in 3.5. The input of the LMSCB 3.5a is fed to the multi-scale feature extraction component in parallel with the addition operator, which receives the multi-scale extractor output after batch normalisation to create an enhanced input volume that is finally down-sampled in the pooling module. The five LMSCBs layers in MM-CM 3.5b correspond to multi-classification that allows for four output sigmoid neutrons, one for each label.

For testing the BoW in lesion detection, they used its feature vector on an support-vector machine classifier, achieving a sensitivity of 44%, 90% specificity and an AUC of 81%. By comparison with other methods, BoW only failed to surpass the CNN approach in sensitivity.

In multi-label detection, standard healthy features were used for learning (bubbles, residues, lumen hole). The MM-CNN outperformed the BoW, now applying the vector to a multi-layer perception classifier, with 90% and 83% in the AUC metric.

3.3 Datasets Used in the Literature

The methods in Section 3.2 used distinct datasets, where private databases were primarily used. The data counts with CE extracted frames and even a case of annotated patches from frames. All of which is briefly compared in Table 3.1. In the literature, the best performing method [25] is also the one with the most extensive dataset and that accounts for the widest variety of abnormalities. However, this was achieved with a high effort from multiple specialists, and the whole set was still privately acquired and annotated entirely. Given the great dependency of CNNs on large amounts of data [33], and the cost of annotation, this is not always attainable.

Other approaches explored smaller data by employing data augmentation [20, 26] or patch-based techniques [30, 26] to counter these data size limitations. These same approaches, however, are applied to detect a small variety of abnormalities in the work of [20] and [30]. This lack of diversity in content is also verified in [23, 28, 31] and the approaches that include the KID Dataset 2[3].

Therefore, given the low representativity of the full VEC reality in most of the used datasets, the work developed still lacks in performance when applied to full videos [9]. Besides, the annotation effort necessary to acquire a significantly representative dataset, like in the case of [33], would require mobilising numerous experts and examining a tremendous amount of clinical cases, in which presumably most of the frames would have redundant information on the VEC reality. Thus, one should seek to optimise the data selection to minimise the labelling cost associated with such a task.

[3]KID Datasets - "https://mdss.uth.gr/datasets/endoscopy/kid/"

Table 3.1: Characterisation of the Datasets used in Publications.

Publication	Applied Dataset	Characteristics	Size	Content
Teresa et al.[20]	GIANA[1] Challenge Dataset	SB content abnormalities from an unknown number of patients.	1812 video frames	600 healthy frames, 607 inflammatory and 605 vascular.
Ding et al. [25]	Private Dataset	Composed of 6970 clinical cases with SB-CE, from 77 medical centres from July 2016 through July 2018.	113,426,569 Video Frames	Images with normal, inflammation, ulcer, polyps, lymphangiectasia, bleeding,vascular disease, protruding lesion, lymphatic follicular hy-perplasia, diverticulum, and parasite content.
Pan et al. [23]	Video clips available in the Endoscopy Organization website.	Images obtained from 10 full video clips.	960 video frames	546 bleeding images and 414 non-bleeding images.
Sadasivan and Seelamantula [26]	Private Dataset	-	137 Video Frames	77 images with 9 different types of abnormalities and 60 healthy images.
Sharif et al. [28]	Private Dataset	Composed of 10 videos obtained from POF Hospital, Pakistan.	4500 Video Frames	2000 bleeding, 2000 healthy and 1500 ulcer disease.

Li et al. [30]	Private Dataset	Composed of medically annotated patches from 3 patients from Prince of Wales Hospital in Hong Kong.	Set composed of 3600 patches.	Representative abnormal (1800) and normal (1800).
Wang et al. [31]	Private Dataset	Cases from different patients provided by Nanjing Jinling Hospital of Nanjing University School of Medicine.	2000 Video Frames.	Normal frames with non- or small bubble and non-useful frames with plentiful bubbles.
Vasilakakis et al. [29] Iakovidis et al. [32]	KID Dataset 2[3]	SB lesions, as well as normal video frames obtained from the oesophagus, the stomach, the SB and the colon.	2371 VEC MiroCam® frames.	303 frames are weakly annotated as vascular, 44 as polypoid, 227 inflammatory and in also includes 1778 normal video frames.

3.4 Active learning

AL is a domain in machine learning that seeks to overcome a lack of labelled data, relying on the principle that a learning algorithm should choose the data to be annotated. The selected data is, then, redirected to an oracle, or annotator, for its labelling; thus, this strategy is believed to improve the model's performance with fewer labelled data.

The central hypothesis in AL attempts to reduce the annotation cost taking a large amount of unlabelled data $\mathcal{U}$, and assuming that there are one or multiple oracles to annotate a labelled dataset $\mathcal{L}^*$. The typical approach would be to annotate all the elements, x, in $\mathcal{U}$, creating a fully labelled database $\mathcal{L}$. However, AL strategies seek to define a subset $\mathcal{L}^*<\mathcal{L}$ to train the model on, where it can achieve an equivalent performance as if $\mathcal{L}^*=\mathcal{L}$ [34].

Hence, this section will present some of the various approaches to take into consideration in this field. The general strategy can be found in the framework bellow 3.6, consisting of iteratively, selecting the data samples x according to a sample active selection strategy. Finally, they are presented to the oracle for its annotation and incorporated into new labelled data and included in the model's knowledge [35, 34].

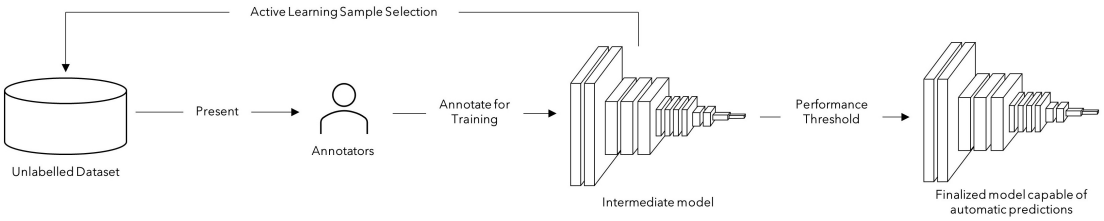


Figure 3.6: General pipeline used in active learning strategies.

3.4.1 Query Scenarios

The initial decision to outline an AL framework is the query type used to feed the unlabelled data to the model, counting with the three popular query types presented below.

Stream-based Selective Sampling

This type of query presents the learner with each sample of the database individually, for the query strategy to evaluate its informativeness and decide if it should be added to the training query. By doing this, it isolates the decision for each element, since it is not taking into account the whole distribution of the database, providing a less computationally expensive technique with a tendency for more imbalanced exploration and exploitation approach [35, 34].

The selection of the instances to query can be achieved by measuring its informativeness, likely querying the ones that present highest expected utility based solemnly on that individual evaluation. This can also be achieved by computing a *region of uncertainty*, the part of the instance space where learner's knowledge is more ambiguous, and query the instances that fall within that

region. A threshold can impose the instances selection, that delineates the area. However this is a somewhat naive way to approach the choice, another approach would seek to define this region according to the current labelled set, creating hypotheses that can comparatively select the instances that rise the highest discordance [35].

Membership Query Synthesis

Membership Query Synthesis consists of the creation of each sample instead of drawing examples from a natural distribution of data. The generated samples seek to provide the most informative examples to improve the model's knowledge; nonetheless, this technique falls into the same flaw as the previously presented method, given that it might be unaware of unseen knowledge areas from which the model lacks in performance [34]. This autonomous synthesising of samples to mimic real-world distributions is an area with great potential; however, some of the generated examples may be incomprehensible to the human understanding, making it harder for the oracle to label [35].

Pool-based Sampling

Similarly to *Stream-based Selective Sampling*, Pool-based Sampling scans through an existing pool of data; however, this type of querying takes the unlabelled dataset as a whole, and the query is extracted by ranking the informativeness from all the samples. This scheme is the most commonly used compensating the imbalances presented in the previous methods; nevertheless, it requires substantial computational power, making it is not eligible for all applications [35, 34].

3.4.2 Query Strategies

Regardless of the selected query scenario chosen, an informativeness measure will be required to evaluate the knowledge of the model at each stage of the AL framework. A wide variety of parameters can be assessed regarding the familiarity of the model with the reality that it should learn; thus, distinct strategies have been developed. In the extensive review of AL by Settles [35], he groups the techniques into the following:

- **Uncertainty Sampling**: Instances are queried with the highest uncertainty in the model's output;
- **Query-By-Committee**: Having a committee of learners, the instance with the highest disagreement between the members of the committee is queried;
- **Expected Model Change**: The queried instance should be the one that can change the model the most;
- **Expected Error Reduction**: The queried example should be the one that will reduce the model's generalisation error the most;

- **Variance Reduction**: The queried instance should be the one that minimises output variance the most;
- **Density-Weighted Methods**: Queried samples concerning their representativeness in the data distribution.

Uncertainty Sampling

This group of strategies queries samples for which the model's classification is more dubious, often relying on probabilistic models to evaluate this certainty. One of the most typically used measures of uncertainty is the ***Least Confident*** strategy that takes the class with the highest posterior probability, for each instance, under the model θ, and returns the sample for which this probability is the smallest, as shown in the equation below 3.7.

$$x_{LC} = \underset{x}{\text{argmax}}\ (1 - P_\theta(\hat{y}|x)), \quad \hat{y} = \underset{x}{\text{argmax}}\ P_\theta(y) \tag{3.7}$$

One can interpret this posterior probability as of how certain the model is of its classification; hence, a drawback associated with this strategy is that it only takes into account the highest posterior probability for each instance, instead of the posterior distribution [35, 36].

Margin, on the other hand, is a strategy that takes into account the two most probable labels for each instance and measures the model's ability to differentiate between the two most probable classes, as in the following equation 3.8.

$$x_M = \underset{x}{\text{argmin}}\ [P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x)] \tag{3.8}$$

Where $\hat{y}_1$ and $\hat{y}_2$ are respectively the first and second most probable classifications for that instance; thus, the selection of the instance to query would correspond to the smallest distance between the posterior probability of these two [35, 36].

Lastly, ***Entropy*** sampling uses entropy [37], denoted by H, to measure uncertainty through comparison of each instance y of the available labels as follows 3.9.

$$x_H = \underset{x}{\text{argmax}}\ H_\theta(Y|x), \quad H_\theta(Y|x) = -\sum_y P_\theta(y|x) log P_\theta(y|x) \tag{3.9}$$

This being the evaluation of the variable's average information content, which can be thought of as the log-loss expected for each instance [35, 36].

Query-By-Committee

These query strategies involve the integration of a committee of classifiers $C = \{\theta^{(1)}, ..., \theta^{(C)}\}$, trained on the current $\mathcal{L}$ labelled dataset, that constitutes hypotheses to evaluate the unlabelled dataset $\mathcal{U}$. The members of the committee vote on the labels for each input and the query selection

is based on disagreement sampling strategies. Hence, relying on the establishment of a *region of uncertainty* composed of the samples where these hypotheses disagree.

The first measure of disagreement strategy to be explored is ***Vote Entropy*** sampling. In this approach, each classifier labels all instances, and the probability distributions of the votes of each class y for every instance's x is computed. The queried sample is selected according to the maximum entropy of the voting distributions, presented in the equation below 3.10.

$$x_{VE} = \underset{x}{\operatorname{argmax}} - \sum_{y} \frac{vote_C(y,x)}{|C|} log \frac{vote_C(y,x)}{|C|}, \quad vote_C(y,x) = \sum_{\theta \in C} \mathbf{1}_{\{h_\theta(x)=y\}} \tag{3.10}$$

In this approach, the $vote_C(y,x)$ component, presented in 3.10, is the number of votes a label y received for the instance x by the elements of the committee C [35].

Another strategy that is typically employed in ***Consensus Entropy*** sampling, that instead of considering the votes measures the average class output distribution of each learner for all instances x. In the 3.11 equation, this operation is presented, where the previously mentioned distribution is $P_c(y|x)$, which is consequently compared through the entropy measured [35].

$$x_{SVE} = \underset{x}{\operatorname{argmax}} - \sum_{y} P_C(y|x) log P_C(y|x), \quad P_C(y|x) = \frac{1}{|C|} \sum_{\theta \in C} P_\theta(y|x) \tag{3.11}$$

The last committee strategy is ***Maximum Disagreement*** sampling, in which the disagreement in the consensus probability is measured for each learner, and the queried instance is the one with the highest disparity for one of the members of the committee.

$$\begin{aligned} x_{KL} &= \underset{x}{\operatorname{argmax}} \frac{1}{|C|} \sum_{\theta \in C} KL\left(P_\theta(Y|x) \parallel P_C(Y|x)\right), \\ KL(P_\theta(Y|x) \parallel P_C(Y|x)) &= \sum_{y} P_\theta(y|x) \; log \frac{P_\theta(y|x)}{P_C(y|x)} \end{aligned} \tag{3.12}$$

As it is demonstrated in the equation above, the *Kullback–Leibler divergence (KL)* measures the difference between the two probability distributions $P_\theta(y|x)$ and $P_C(y|x)$. The first being the average divergence of each member of the committee, and the second the above-mentioned consensus probability [35].

Expected Model Change

Measuring the impact of each instance, if the learner was to learn it, can be achieved with ***Expected Gradient Length***. This approach for discriminative probabilistic models takes advantage of their typical gradient-based optimisation and uses the length of the training gradient as a measure.

$$x_{EGL} = \underset{x}{\operatorname{argmax}} \sum_{i} P_\theta(y_i|x) || \nabla l_\theta(\mathcal{L} \cup \langle x, y_i \rangle) || \tag{3.13}$$

In equation 3.13 the largest gradient $\nabla l_\theta(\mathcal{L})$ is calculated using the objective function l and with the model parameters θ. The gradient's euclidian length is obtained through speculation over all the possible labels y creating the pairs $\langle x, y_i \rangle$ according to the expectation $P_\theta(y_i|x)$. This estimation of the impact, however, is rather expensive when the feature space or the labelling set is large, and inappropriate scaling of the features can lead to an overestimation of x [35].

Expected Error Reduction

The queried instance should be the one that will directly minimise the model's generalisation error the most. Thus the expected error from the future train model is calculated for $\mathcal{L} \cup \langle x, y \rangle$ for each pair in the $\mathcal{U}$. The loss function can be calculated through the $0/1-$loss approach or with the log-loss, both presented in 3.14 and 3.15 respectively. In both equations the $\theta^{+\langle x, y_i \rangle}$ notation stands for the resulting model when adding the $\langle x, y_i \rangle$ pair to the $\mathcal{L}$ dataset.

$$x_{0/1} = \underset{x}{\operatorname{argmin}} \sum_i P_\theta(y_i|x) \left(\sum_{u=1}^{U} 1 - P_{\theta^{+\langle x, y_i \rangle}}(\hat{y}|x^{(u)}) \right) \tag{3.14}$$

In the 3.14 equation one can interpret that contrary to the case in the Expected Model Change strategy, in this approach one expects to reduce the number of wrongly predicted labels.

$$x_{log} = \underset{x}{\operatorname{argmin}} \sum_i P_\theta(y_i|x) \left(-\sum_{u=1}^{U} \sum_j P_{\theta^{+\langle x, y_i \rangle}}(y_j|x^{(u)}) log P_{\theta^{+\langle x, y_i \rangle}}(y_j|x^{(u)}) \right) \tag{3.15}$$

The 3.15 minimisation presents us with a similar expression to the expected entropy over $\mathcal{U}$. Both implementations, allow for the optimisation of a measure of interest like precision or recall. However, this is the most computationally expensive procedure, since it requires iterative retraining over all instances of the pool [35].

Variance Reduction

This strategy targets the prospective generalisation error, attempting to reduce the output variance. In the review by Settles [35], the expected output $\sigma^2_{\hat{y}}$ is approximated in 3.16.

$$\sigma^2_{\hat{y}} \approx \left[\frac{\partial \hat{y}}{\partial \theta} \right]^T \left[\frac{\partial^2}{\partial \theta^2} S_\theta(\mathcal{L}) \right]^{-1} \left[\frac{\partial \hat{y}}{\partial \theta} \right] \approx \nabla x^T F^{-1} \nabla x \tag{3.16}$$

With ∇x being the gradient of the models predicted output and F being the covariance matrix of the second-order expansion of the objective function S the squared error of the model in the current state θ. From this, the estimated output variance $\langle \tilde{\sigma}^2_{\hat{y}} \rangle^{+x}$ when the model is retrained on

the x instance, assuming the prediction for x is quite close to the real one and that ∇x is locally linear, querying with variance reduction can be achieved as presented in 3.17.

$$x^*_{VR} = \underset{x}{\operatorname{argmin}} \langle \tilde{\sigma}^2_{\hat{y}} \rangle^{+x} \tag{3.17}$$

This method, as the previously presented, is relatively slow given the considerable computational resources required to go over all the input space.

Density-Weighted Methods

Density-Weighed methods intend to measure the resemblance of each of the instances x, with the average distribution of the input space. Hence, a similarity measure $sim(x, x^{(u)})$ is employed to ensure that the querying strategy accounts for the representativity of the selected samples, to avoid the selection of outliers.

$$x_{ID} = \underset{x}{\operatorname{argmax}} \Phi_A(x) \times \left(\frac{1}{U} \sum_{u=1}^{U} sim(x, x^{(u)}) \right)^{\beta} \tag{3.18}$$

In the 3.18 equation, the similarity measure, $sim(x, x^{(u)})$, supports the main querying strategy $\Phi_A(x)$, creating an information density framework. With the last component of the equation, being the average similarity meterage of an instance to all the other data points of the input space. This element then is weighted with a β relatively to the main informativeness measure $\Phi_A(x)$.

This similarity measure is typically a distance between each instance and the entire distribution, such as euclidean or Chebyshev distance [38]; thus it can be precomputed for all samples to be integrated into the information density function for a more representative selection of data [35].

3.4.3 Batch-Mode Active Learning

In the traditional approach of AL, only one sample is queried at each iteration, followed by the retraining of the model. For the implementation of NN, however, retraining at each iteration is not only a costly process, but it is also not significant in the evolution of the network's knowledge [39]. Therefore, the selection process should be adapted to a batch setting by selecting a set of data samples that are expected to maximise the models' knowledge of the database reality.

An intuitive approach would be to query a batch of the "top k" most informative instances. Nonetheless, the selected highly ranked examples risk being identical, given the models' difficulty in recognising a specific region of the instance space [35, 39]. For that purpose, other strategies have been developed to ensure diversity within the selected batches. Similarly to the previously introduced density-weighted method, in the work of Xu et al. [40], the set is incremented upon with the samples that maximise the sum of $\alpha(informativeness) + \beta(diversity)$, where $\alpha + \beta = 1$, employing a distance of all unlabelled samples from the current batch as a diversity measure. Another strategy applies k-means clustering to the highly informative selected set, filtering redundant instances by querying the centre of the cluster [41, 36].

3.4.4 Active Learning in Medical Imaging

In medical imaging, the study of AL is especially relevant, concerning the increasing amount of data that is available, and the high cost associated with the required specialist's attention to annotate it [34]. For VEC exploitation, however, these active selection techniques are yet to be explored, even so, other image-based medical multi-classification challenges have employed it.

Gu et al.[33] experimented on confocal endomicroscopy images and gastrointestinal endoscopic images, starting by pretraining ResNet-18 networks on ImageNet data, following with a fine-tuning using labelled data. Then they search for the 64 samples that show the most considerable uncertainty levels, to update the CNN. These samples are then subjected to iterative supervised learning, in which they are sent to the oracle for classification. This way, the network's update is always based on where it would be least confident in classifying, reinforcing the clarification of the model.

For the selection of the most dubious data, Gu et al.[33] aim at distinguishing the cases with a small inter-class variation where the differences lie in detailed features. They start by using a dropout strategy, in which, at each iteration, it predicts the labels, and measures the mean and variance of the prediction. This adds up to an evaluation of its susceptibility to visual variation from imaging setting, by randomly transforming the original images, and calculating its variance. Evenly summing up these uncertainty values, the unlabelled dataset is then organised in decreasing order of uncertainty value result, and clustered into subsets. Finally, for each iteration, a selected subset is sent to be labelled by experts, and the model is updated.

The work of Folmsbee et al. [10], for example, explicitly compares a random learning CNN and AL training for the classification of tissue in oral cavity cancer. This strategy presents a patch-based CNN classifier, that is converted into a fully connected CNN (fc-CNN) and returns pixel maps for each image.

The AL approach iteratively starts by randomly selecting a small number of regions of interests for manual labelling, that generates the training set and define the classifier. The latter's performance is evaluated, and a resulting pixel map is obtained on the classified examples. Quality assurance is performed for each of the resulting pixel maps, and the n worse classified instances are used for training in the following iterations by repeating the previous steps.

Comparing these two methods after three iterations, the accuracy of AL returned a value of 96.44% contrasting the lower accuracy of the CNN with 93.50%, which demonstrates better efficiency by the AL method.

3.5 Summary

As is stated in the research presented above, there is still a need to improve the novel technologies for VEC abnormalities detection and characterisation.

The majority of the most successful approaches rely on CNNs, with detection including binary distinctions between normal and abnormal frames [27], and multi-classification of different

abnormalities [26, 20]. These networks, however, are generally limited to the universe represented in the databases used for training, e.g. Teresa [20].

The comparison between different VEC abnormalities detection software is also challenging given the inexistence of a public representative standard database for evaluation of its performance [8, 9]. Thus, the implementation of learning strategies to optimise the knowledge acquired on the existing datasets should be employed.

AL strategies have proven to improve the performance of machine learning strategies with fewer labelled instances, which is especially of interest in the field of medical imaging where the annotation process comes at a high cost [34]. Thus the exploitation of AL techniques to optimise the learning process can be very beneficial. Lacking a large public dataset, VEC analysis could profit from such AL approaches, which are yet to be explored in this field of study.

Chapter 4

Methodology

This chapter will present the methodology that makes up the work developed by this dissertation, exploring AL strategies to improve detection of abnormalities in VEC using CNNs. The selected procedures considered the following issues pointed out in Chapter 3:

- The revised literature highlights the implementation of DL based methods, namely CNNs, as the best performing methods for the task of abnormalities detection in VEC frames. Therefore, the employed techniques resort to these models;

- The need to use data augmentation techniques over the small available CE images datasets for its implementation in DL models, to develop a model more robust to transformations and avoid overfitting;

- Public datasets are not sufficiently representative to reflect the clinical reality contained in full-length VEC, which is where the published methods are not applicable. Thus, this suggest that new VEC labelled data is required for the correct generalisation of ML models;

- AL strategies provide efficient dataset selection techniques to expand the learner's knowledge towards a more representative sample set, enhancing the model's performance. The implementation of these strategies also allow for a cost-efficient integration of unlabelled datasets.

On that basis, the defined methodology main steps are described as presented in Figure 4.1. In a first stage, VEC data is provided and was characterised and prepared, resulting in organised datasets for the following tasks. Consequently, starting models were initiated, a first implementation to determine the networks capability to represent the VEC reality. From then on, if the base models are unable to satisfy the requirements, they enter an AL cycle. Here, there is a sequential active selection of datasets that are included in the model update until satisfactory final results are achieved. Each of these tasks will be further explored in the present chapter.

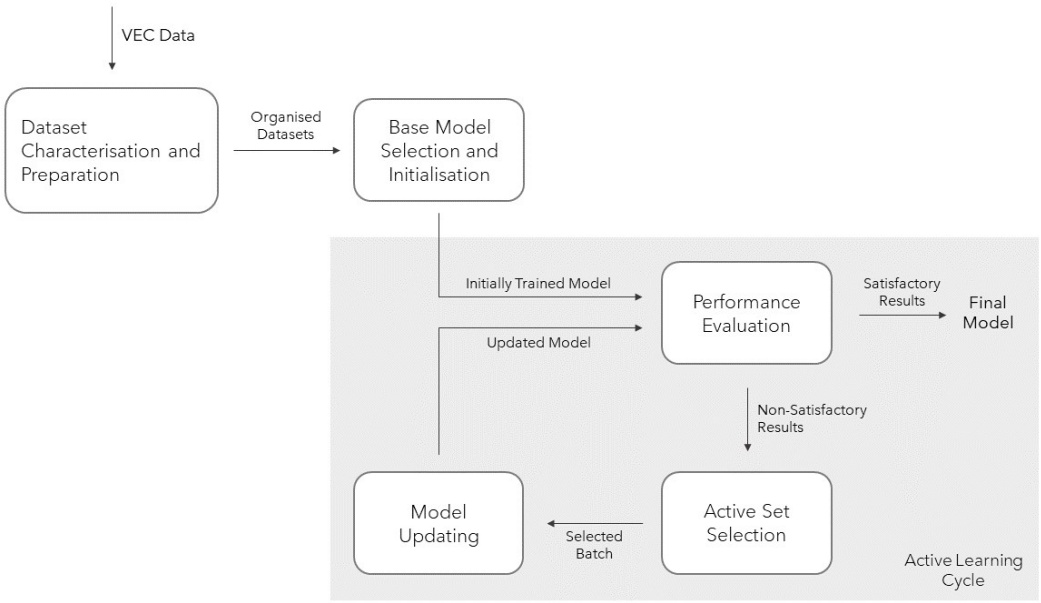


Figure 4.1: Pipeline followed in the Methodology.

4.1 Datasets Characterisation and Preparation

Two datasets were available for the implementation of the proposed solutions; the public annotated dataset GIANA[1] presented in Section 3.3, composed of inflammatory, vascular and normal frames, and additional private unlabelled dataset, further exploited using AL strategies. Both of these datasets are generally presented in Table 4.1 and underwent a more profound characterisation and preparation which will be presented in this section.

4.1.1 Datasets Characterisation

The step of characterising the data has extreme importance given that the issue which this dissertation wants to tackle is presenting the learner with data that can effectively represent the universe in VEC. Thus, having a solid knowledge of the content to be exploited serves as the foundation for the development of the proposed methodology presented in this chapter.

GIANA Dataset

The publicly available, and widely used GIANA[1], comes with a testing and training sets in which the later contains 1812 CE annotated frames. From the images in the training set, 600 were healthy, 607 inflammatory and 605 vascular. The abnormalities depicted in this dataset feature clean static images with explicit content, as can be observed in the examples presented in Figure 4.2.

Table 4.1: Used Datasets Characterisation.

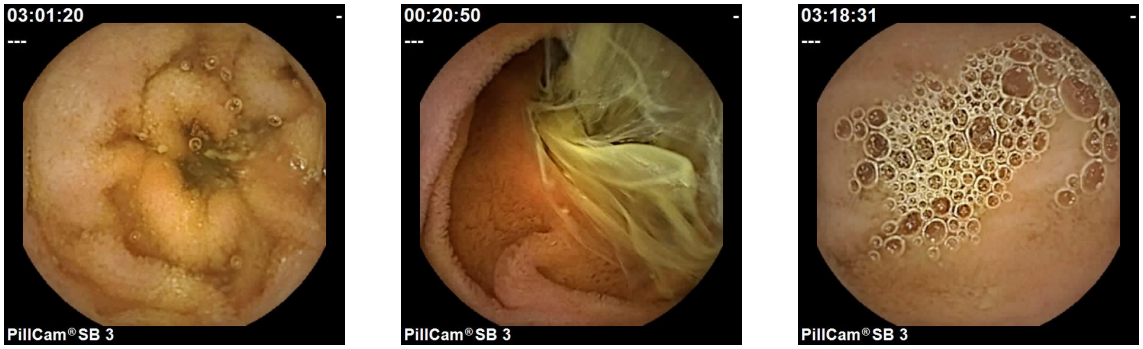
	GIANA[1] Dataset	**Private Dataset**
Data Format	VEC Images	VEC Full Videos
Used Videos	Unknown Ammount	272 Clinical Cases
Data Size	3619 Video Frames	5.44M Video Frames
Number of Labelled Frames	1812 Video Frames	2351 Video Frames
Number of Registered Abnormalities	1212 Video Frames	1832 Video Frames
Content Variety	Healthy, Vascular and Inflammatory lesions	Healthy, Vascular and Inflammatory lesions, Polyps, Lynfagiectasias, Bleeding, Xanthelasmas, Submucosal Masses, Nodular Lymphoid Hyperplasia, Parasites, etc.
Annotation Type	Frame-wise Semantic Annotation	Frame-wise Semantic Annotation

(a) Erosions (b) Edema (c) Ulcer

(d) Angiectasia/Angiodysplasia (e) Erythematous Patch

Figure 4.2: Abnormalities found in the GIANA[1] Dataset.

The healthy frames that are presented in this dataset also consist of clear instances, including conditions with mucus and bubbles, which are also essential to characterise the normal content found in VEC, examples of these instances are shown 4.3.



(a) Normal Content (b) Normal Content with Mucus (c) Normal Content with Bubbles

Figure 4.3: Healthy content in the GIANA[1] Dataset.

This variety of normal and abnormal content makes this dataset an eligible starting point for the learners to get familiarised with instances that can be found in VEC. Nonetheless, as it was concluded in the literature review (Chapter 3), this selected set is insufficient for the learner to be able to generalise and detect abnormalities in full videos. Thus, these contain more ambiguous and diverse content that is not typically included in the available datasets.

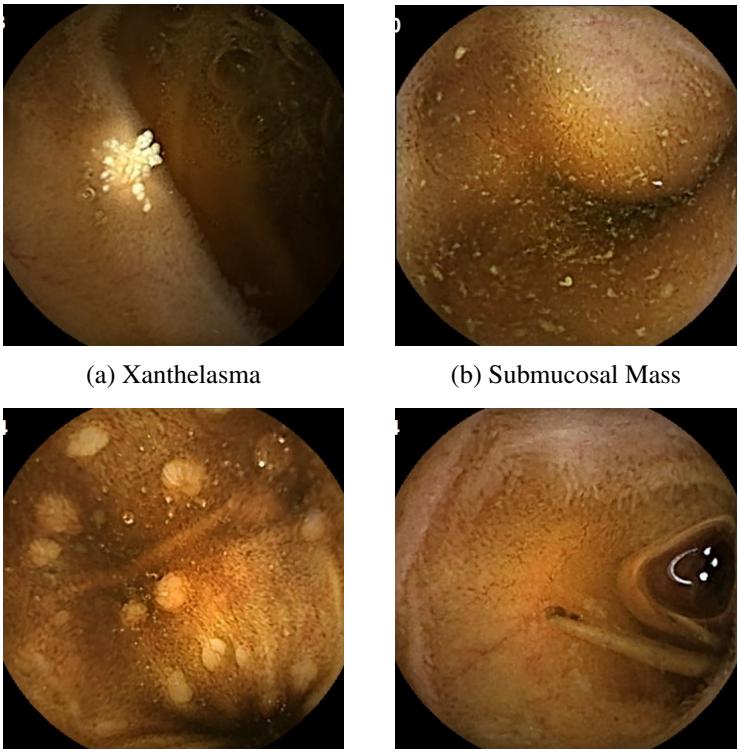
Private Dataset

An extensive private dataset composed of 272 full SB3 videos, each with an average of 20000 frames unlabelled frames, was introduced to complement the GIANA's[1] normal and abnormal content depiction. The downside to the introduction of this new private dataset resides in its very low rate of annotated frames. Thus, this dataset requires an in-depth assessment to ascertain the best approach to the active selection techniques that were employed in this dissertation.

As well as the full videos, we had access to highlighted frames, or findings, concerning the semantic content present in the images or the anatomical position of the endoscope in the tract. The provided dataset counted with a total of 7530 findings with content highlighted by physicians, from which 2351 were annotated in the Rapid Reader[2] VEC proprietary format and 1832 of these had abnormality related annotations. These findings information, although scarce compared to the frame-wise distribution, provided not only real weakly annotations on the frame content, but allowed for an overall analysis on the distribution of the abnormalities in the videos, and served as orientation for the posterior annotation task.

Abnormalities in Study The introduced dataset includes a wider variety of abnormalities compared to the GIANA's[1] content, prominently referencing, vascular lesions, inflammatory lesions,

polyps, lynfagiectasias, bleeding, xanthemasmas, congestions, submucosal masses, nodular lymphoid hyperplasia and parasites. Aside from the most common abnormalities, presented in Section 2.3, the newly considered abnormalities were subjected to another examination, and are shown in Figure 4.4.



(a) Xanthelasma

(b) Submucosal Mass

(c) Nodular Lymphoid Hyperplasia

(d) Parasite

Figure 4.4: Abnormalities found in the private database.

Represented in Figure 4.4a, xanthelasmas are a collection of xanthoma cells, characterised as foamy lipid-laden histiocytic cells [42]. These appear to have very well defined clusters of bright yellowish swollen villi, easily detectable.

Submucosal lesions identification followed the SPICE validation method [43], a technique to distinguish these from benign blunges, which are mere loop angulations of impressions or an adjacent loop. The SPICE takes into account the lumps surrounding mucosa, the height and diameter ratio, the lumen visibility, and its continuous appearance on the video for 10 minutes. As can be visualised in fig 4.4b, these are salient well-defined lumps with a height larger than its diameter, that does not obstruct the lumen.

Nodular lymphoid hyperplasia (NLH), observed in Figure 4.4c, can be characterised by the presence of multiple small nodules, between 2 and 10 mm in diameter. This condition, defined by markedly hyperplastic, mitotically active germinal centres, and well- defined lymphocyte mantles, are found in the lamina propria and/or in the superficial submucosa.

As is shown in Figure 4.4d, parasites were also found in the studied cases, as a possible cause for obscure gastrointestinal bleeding, SB ulceration and infections diseases. They can be categorised into nematodes (roundworms), cestodes (tapeworms), and trematodes (flukes), with the first one being the most commonly found [44, 45].

4.1.2 Datasets Preparation

The two used datasets were applied in the developed strategies, where the GIANA[1] served as a base annotated dataset that represented part of the reality that can be observed in VEC, and the private unlabelled dataset was explored in order to expand the learners perception of this universe. Given the distribution explicit in Table 4.1, and the characterisation in the previous section, this section will introduce the preparation required for each of these datasets before their implementation.

It is worth noting that no preprocessing step was included in this stage, given that previous work presented in Chapter 3, verified no significant improvement in abnormality detection when including preprocessing techniques [20].

GIANA Dataset

In Section 4.1.1, the GIANA[1] was found to have a sufficiently diverse set of well-characterised abnormalities for this set to be considered an eligible starting point for the sequential AL models' evolution. This way, the AL strategies were applied to learners already familiar with some content present in VEC, and could more easily explore the unknown regions of this instance space.

Given the wide variety of abnormalities that can be found in both datasets that are to be included in the methodology, the exploitation of this dataset started by grouping the inflammatory and vascular lesions to learned as abnormalities, and maintain learning coherence with the way the unlabelled dataset will the annotated.

Since the images contained a frame of information of the CE, the outer border text information was cropped, reducing the size of the frames from 576x576 pixels to 512x512. Besides cropping a resize was employed of 224x224 pixels to serve as proper input for the networks. Subsequently, the images underwent data augmentation transformations of 90^{o}, 180^{o}, and 270^{o} rotations, horizontal and vertical flips, which turned our set of 1812 images into 10872. These transformations improve generalisation of the learners, since the resulting images do not fall far from the images naturally captured by the camera in different positions, and allow for the model to be more robust to rotation. This data augmentation step was applied straightaway because, otherwise, it would be repeatedly integrated every time this dataset was incorporated.

Finally, this augmented data set was split into three random batches where 50% was for training, 25% for validation and the final 25% for testing. The augmented GIANA[1] train and validation sets were also included in the updating of the models in certain experimental sets.

Private Dataset

The private dataset was in Rapid Reader[2] format which required a frame-wise extraction, achieved using the software Sensarea[4]. The findings from the videos were retrieved with the PillCam™ reading software Rapid Reader[2].

The videos contained a great extent of the tract beyond the SB. Initially, the first and last finding were believed to correspond to the beginning and end of this extent of the tract, which was the case in most files. Thus, the mean squared error (MSE) between the findings and the video frames connected this correspondence, to reduce the portion of the video that was analysed. However, after a previous criteria correction, it was noted that some findings contained instances of the oesophagus, stomach and colon, which could not be included. For that purpose, these findings were rejected, and the first instance of the duodenum and first image of the cecum served as a reference, the MSE as employed likewise, reducing the video examining to the SB portion effectively.

For the AL exploitation of this set, training, validation and test videos selection approaches should be taken into account to select beforehand the videos which should be excluded from the pool available active selection, and from training. The sorting of the videos suffered alterations with the evolution of the stages of the work. However, a total of 87 videos at were sufficient in the AL selection experimentation, which allowed for the exploitation of the remaining for validation and test sets selection. All the actively selected frames were also resized to 224 by 224 before network input.

Test Sets In a first approach, two videos were put aside to serve as a test set, before any further exploitation. A frame extraction every 200 sequential frames per video amounted to 836 images annotated by members of the team. The chosen number corresponded to a considered reasonable amount of instances to be annotated with enough spacing between them to representatively approximate full-length videos, without querying images that could be too similar. The chosen two clinical cases held the highest amount of findings in comparison with the other videos, and both contained a diverse set of abnormalities.

After the first correction of the annotation and video extraction approaches, this test set was corrected, with a re-extraction of the frames. This alteration was due to the exclusion of oesophagus, stomach and colon content mentioned above, and a criteria correction that was necessary. With the shortening of the videos, the sequential frame extraction followed an interval of 100 frames instead, resulting in a second training set with 1052 images.

In the more recent experimental sets, four other clinical cases were included, fairly equally abundant in abnormality diversity and quantity, and which have not been analysed in previous experimental setups. However, 200 were extracted per video, counting with a total of 1200 data samples. An extension of the test set was applied since the inclusion of more clinical cases could

[4]Sensarea - "http://www.gipsa-lab.grenoble-inp.fr/~pascal.bertolino/software.html"

improve the learner's performance perception. However, the number of samples per video had to be reduced, given the time-consuming task of annotating.

Validation Sets The AL set selection strategies, which will be further explored in a forthcoming section, query a 90 samples-sized batch to be included in the updating of the model at that iteration of the sequential evolution. The validation set selection followed three procedures, presented below.

In the initial approach, the data augmentation techniques previously applied to the GIANA[1] were employed to the entire small AL selected set, resulting in a 540 batch. From here, 100 random samples were selected as validation data. This decision was taken due to the initial unfamiliarity with the video's content overall distribution, and this being considered a possible approach in the AL validation set selection strategies [36].

However, with the work development, a high unbalanced active selection of samples was noticed, given the predominance of normal selected images compared to the abnormal ones. So the validation selection strategy evolved to a second more balanced set, in which from the 540 augmented samples, in the 100 selected, up to 50 would have abnormal content. Hence, there was a training and validation distribution which intended for the training set to have at least 50% of the abnormalities. The only downside to this approach was that in the first iterations, even with data augmentation, in some occurrences, the augmented batch would not have enough abnormal samples to make up a balanced validation set.

In later stages of the work, a fixed balanced set of 400 images manually selected for validation. This approach ought to pick out representative samples of the reality expressed in VEC, to explore the evolution of the networks when trained on the same validation set. The collected frames were extracted from 8 videos which were not included in the active selection component from the start, since there were no findings associated with them. By not having any findings, these videos did not allow for automatic identification of the first and last instance of the SB; Yet these videos allowed for a manual exclusion of frames that corresponded to the oesophagus, stomach and colon extent of the tract, which was part of the selection. Besides the selected frames, the validation set also includes a portion of the findings, given that the videos without findings did not have sufficient abnormalities to balance the set. These findings were excluded from the active selection by the same MSE correspondence presented in 4.1.1.

4.2 Base Model Selection and Initialisation

For the task of detecting abnormalities, CNNs were implemented given their high success rate in the task of VEC abnormalities identification, when compared to other ML techniques, as it was concluded from Chapter 3. Re-purposing trained networks for different applications through fine-tuning is a highly successful approach when one has to handle small datasets. Thus, to serve as a starting point to the AL experimentation, the initial models were obtained through fine-tuning networks previously trained on the ImageNet database [46], on the GIANA[1] dataset.

The base pre-trained models applied for this component were selected based on work previously conducted by the team [20], from which two of the best-succeeded models were selected. The first being the VGG16 with batch normalisation and a secondary DenseNet-121 included in the committee based AL strategies. From there, the original models were acquired from PyTorch's available models[5] already pre-trained on the ImageNet[46] database. These had all their layer entirely fine-tuned, given that it allowed feature extraction from the VEC frames and a more profound adjustment to the learners' weights. The datasets used for training and validation were the augmented GIANA[1] sets presented early in this section, with the required normalisation transformations for training each of the models were employed [47, 48].

4.3 Performance Evaluation

The resulting models, for each stage of its development, had their performance evaluated concerning ground truth classification of the test sets presented in Section 4.1.2.

Their efficiency was reflected on the classification distribution of true positives (TP), the correct identification of abnormalities in the set, compared with the number of false positives (FP) and false negatives (FN) classified. These values allow the estimation of the performance using recall, precision, area under the receiver operating characteristic (AUC) curve and F score. These metrics are widely used for performance evaluation in the reviewed literature (Section 3).

$$precision = \frac{TP}{TP+FP} \tag{4.1}$$

$$recall = \frac{TP}{TP+FN} \tag{4.2}$$

$$FScore = \frac{2 * precision * recall}{precision + recall} \tag{4.3}$$

Precision, as can be seen in equation 4.1, verifies how accurately the positive class is classified since it compares the TP with all the values considered positive. Recall, on the other hand, reflects on the model's capacity to identify the positive class with regard to the whole positive distribution, as is presented in the 4.2 equation. The F Score, represented in equation 4.3, is the harmonic mean of precision and recall, combining both to ascertain how close the model is to the perfect performance.

$$FalsePositiveRate = \frac{FP}{FP+TN} \tag{4.4}$$

The ROC curve (receiver operating characteristic) sizes to define the confidence of the model in its classification by plotting the true positive rate, or recall, versus the false positive rate (equation 4.4) at different classification thresholds. To compute the ROC, one would have to evaluate

[5] PyTorch master documentation — torchvision.models - "pytorch.org/docs/stable/torchvision/models.html"

the model at different classification threshold and plot the resulting curve. However, the area under this curve, the AUC, can express the confidence score of the model. In this work, the AUC score was directly obtained with the Scikit learn AUC score function[6], by comparing the model's classification with the ground truth values. The AUC metric has the advantage that it focuses on how well the predictions are ranked, irrespective to the threshold chosen.

4.4 Active Selection Strategies

To maximise the learner's perception of the full VEC reality with sets of samples that are representative enough to reflect it, AL strategies exploited the unlabelled private dataset. This implementation allowed for a more cost-effective labelling approach to the provided set, given that this effective selection of images imply that only these require annotation as opposed to all the video frames.

All the presented approaches are applied in a Pool-based Sampling query scenario since the query selection strategies employed examine an entire pool of data and compare the informativeness measure of all elements in the pool to query the most eligible elements.

Thus, this section goes into detail about the querying and batch selection strategies that were employed at each iteration to evaluate the learners' knowledge over the VEC universe, and select the most effective batches with its sequential evolution.

4.4.1 Query Strategy

The selected queries used *uncertainty sampling* and *query-by-committee* (QBC) based strategies. These groups of sampling strategies were chosen due to the time consuming, and computationally expensive resources required to apply *expected model change*, *expected error reduction* or *variance reduction* to neural networks using images as features, as it was presented in Section 3.4.2.

As it was introduced in Section 3.4.2, *Uncertainty sampling* strategies are probabilistic based approaches, seeking to measure the model's familiarity with the dataset, by comparing the model's classification probability distribution when faced with each instance of the pool. In this implementation, *Least Confident, Margin* and *Entropy sampling*, were applied at each stage of the VGG16 model's evolution, to select the samples that had a higher chance at clarifying the models' classification difficulties.

QBC strategies, on the other hand, compare the output classification from a committee of learners and evaluate the region of the instance space that should be explored by all elements of the committee. In this case, only the VGG16 and the DenseNet121 were selected for the committee, given the time-consuming task of retraining multiple models. *Vote Entropy, Consensus Entropy* and *Maximum Disagreement* were compared to verify the evolution of the selected committee based on their collective difficulties in classifying the unlabelled set of full videos provided.

[6]Scikit Learn AUC score - "scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html"

Implementing these measures was achieved with the AL Python library modAL[7], which allowed for a direct query selection based on the pretended procedures. This library included all *Uncertainty sampling* strategies, as well as all the *QBC* approaches. Given that this library was constructed for Scikit Learn models, the wrapper Skorch[8], which allowed the incorporation of Pytorch models[5].

All these strategies were experimented with given that no AL technique was ever applied in the study of VEC query selection for abnormalities detection, to the best of our knowledge. Concurrently to the active selection strategies that were employed, random sampling was also applied to serve for comparison with the base sequence AL strategies. Nonetheless, random sampling is also a valid selection strategy for the vast content in the videos since it can easily query a very diverse set.

4.4.2 Batch Selection Strategy

Throughout all the employed strategies, there was a consistent selection of 90 samples per batch. The batch would be composed of 15 images per video, analysing a total of 6 videos per iteration. This distribution was chosen regarding the average number of findings per video being 25, and accounting for a percentage of anatomy related findings. Given the high cost of annotation, 90 samples were considered an adequate amount to annotate per iteration.

This batch approach also serves as a ***Subsampling*** diversity strategy, given that, it imposes the selection from 6 different clinical cases individually, each with a different reality of the instance space at a time, so there is a lower probability for redundant selection amongst the videos. Thus, allowing for a more viable implementation of ***"Top k"*** selection to query the instances.

In a latter stage of the strategies developed, a diversity measure for a ***Greedy Selection*** was also integrated. To this end, the measure was employed in a similar way to the information density strategy introduced in Section 3.4.2.

In this implementation, the selected uncertainty measure was *Least Confident Sampling*, and the same selection of 6 videos per iteration was applied. Only this query strategy was selected since this experimental set merely wanted to test the viability of the implementation of diversity measure.

$$\alpha \cdot uncertainty(x) + \beta \cdot d(x,B) \tag{4.5}$$

Contrary to the first batch implementations, 50 instances with the lowest confidence are reviewed per video. From these 50 only 15 are included in the batch. The batch is successively filled out with the elements that minimise the expression 4.5, where $d(x,B)$ is a distance measure of the element x to all the elements in the batch. In the same way as the information density

[7] modAL library- "modal-python.readthedocs.io/en/latest/index.html"
[8] Skorch Library- "skorch.readthedocs.io/en/stable/index.html"

scheme, α and β are the relative importance of the uncertainty and diversity terms, respectively. In this experiment, the chosen values were $\alpha = 0.7$ and $\beta = 0.3$.

$$sim_{cos}(x,y) = \frac{x \cdot y}{||x||_2 \cdot ||y||_2} \tag{4.6}$$

The average *Cosine Similarity*, presented in 4.6, of each x to the batch was the distance measure applied to the flatted sampled images. This operation used the pairwise cosine similarity function provided by the Scikit Learn library[9]. Given that images are composed of only positive intensity pixels, the resulting distance always lies in the interval of 0 to 1. The same goes for uncertainty outcome, since it corresponds to the maximum posterior probability, as it was explained in 3.4.2.

4.4.3 Annotation

For this task of the AL framework, MSE was used to compare the ground truth findings with the actively selected frames from the full videos and obtain instant annotation of the frames that were highlighted by the specialist.

However, given the large number of unlabelled samples in comparison to the number of findings per video, the active selection tended to collect a significant quantity of non previously annotated frames. As we could not have an expert available full-time to help with the annotation, in a first stage it was decided for the team to pre-annotate the AL selected sets. This process would be orientated by a physician and use the findings of each video as a reference to the abnormal content found in each clinical case. Thus, strict norms for the abnormality distinction were required.

With this in mind, the defined criterion followed the detailed abnormality descriptions exposed in Sections 2.3 and 4.1.1, and included guidelines validated by the Centro Hospitalar do Porto, from which alterations in the mucosa could identify abnormalities regarding the following aspects:

- Vascular alterations would be identified by bluish and reddish colouration of regions in the mucosa, small point-shaped red lesions or bright-red ectasias with capillary dilatation and presenting a vessel appearance;
- Active haemorrhage could be easily identified by the presence of blood in the lumen;
- Polyps could be singled out by their distinct inflamed or eroded projected lesion morphology;
- Alterations of the villi with white or yellow alterations portraited xanthoma cells or lynfagiectasias;
- NLH were simple to identify by its small nodules in the mucosa;
- Parasites were seen as strange worm-like artefacts which are easily spotted;

[9] Scikit Learn Cosine Similarity function - "scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html"

- Submucosal masses were regarded as well-defined lumps with a height larger than its diameter, that do not obstruct the lumen;
- Obstruction of the lumen by the haustral folds, with congested mucosa were considered as edemas;
- Finally, regions of excavated musoca were acknowledged as erosions and ulcerated areas.

The criterion defined, however, were subjected to alterations with the evolution of the work developed, and will be further explored in the next chapter.

4.5 Model Updating

The presented approaches involved the sequential evolution of the models' with the introduction of the newly annotated actively selected sets. In this stage of the work, several updating approaches were employed, based on fine-tuning and transfer learning, which are the two of leading approaches to train the networks on the newly AL acquired labelled data [34].

Hence, this section will present the general experimental setups of the updating strategies that were employed, counting with a sequential fine-tuning approach, and two based on transfer learning. In all applications, the labelled reality provided to the model was solemnly previously selected by the AL metric applied to that sequential evolution.

4.5.1 Fine-tuning

"Fine-tuning all the layers of models" was the first strategy to be applied for the models' update, following the initialisation of the models (Section 4.2). This approach, in all experiments, was applied to the ImageNet pre-trained VGG16 with batch normalisation, and Densenet121 in the case of AL QBC and random selection strategies. The different implementations with this model updating strategy essentially lied in the validation and train set distribution, which will be exposed in the next chapter.

Nonetheless, in all implementations, at each iteration, the newly included annotated set, is augmented and split into training and validation groups. The split training set is appended to the entire labelled data selected by that AL strategy up to that point, and the GIANA[1]'s augmented dataset. The fine-tuning update of the ImageNet pre-trained models is applied to this incremented whole, with the selected validation set. Finally, the entire augmented batch is added to the labelled set at the end of the training.

4.5.2 Transfer Learning

When exploring other typically employed approaches to update ML models with AL strategies, transfer learning is also considered a viable option. This technique takes advantage of the features that the layers from the base pre-trained models learned for previous tasks, and readjusts the classifier for other applications. To that end, two main strategies were developed, the first, Continuous

updating, by iteratively updating the models on the newly acquired dataset, and another one where the original pre-trained models are retrained on the increasing dataset of labelled samples at each iteration. Both these approaches are introduced in this section and have the advantage of saving training time, given the lower number of weight to readjust.

Continuous Update

This approach corresponds to a continuous fine-tuning of the classification layer of the updated model on the labelled set. To this end, the AL selection started by testing the knowledge of the GIANA[1] pre-trained models presented above in Section 5.1. From then on, at each iteration, data augmentations is applied to the AL selected batch, and the most recent version of the models is updated on this set.

Updating on the Increasingly Large Labelled Set

Contrary to the previous approach, the following presented strategy consists of always applying transfer learning to the GIANA[1] pre-trained models presented in Section 5.1. For this purpose, at each iteration, the newly acquired dataset undergoes a data augmentation step and is added to the labelled dataset. Thus, transfer learning is applied to the entire labelled set.

4.6 Summary

Throughout this section, the framework proposed in Figure 4.1 had its tasks examined, and the whole methodology was outlined, considering the reviewed literature in chapter 3.

The presented VEC data underwent an analysis that provided its proper setup for the implemented strategies. This step led to a CNNs initialisation that regarding their performance would enter an AL cycle. The cycle included querying a batch of informative samples, and its educated annotation that would lead to updating the models. A couple of updating strategies were proposed, the latter with two variants. Regarding the networks' sequential evaluation on the selected test sets, the models would prove eligible for the task at hand, and exit the cycle.

Chapter 5

Results and Discussion

The present chapter introduces the experiments that were conducted regarding the evolution of models in the task of VEC abnormality detection with AL selection strategies, following the methods previously presented in Chapter 4. Thus, this segment is organised in the chronological order of the work developed, separated by the six stages it explored, their implementation specifics and the annotation criteria refinements that were necessary to correct the labelling component of the AL strategies. Each stage will expose the experiment, followed by the achieved results and finally, a discussion. In case there were annotation alterations, they are mentioned at the end of the section. Tables 5.1 and 5.2 outline the different components of each of the experimentation sets that will be presented.

All stages evaluate the models' development in performance when classifying the full VEC test sets mentioned in Section 4.1.2. The iterative development of the presented models regards the defined AL strategy responsible for the selection of the dataset, included in the models' training at each iteration. The sets' annotation was performed by an element of the team, following the defined guidelines. The stages evolved with respect to the challenges that were faced with the criteria's approximation to the medical standards.

As it is expressed in the presented tables, throughout all the strategies employed, there was a sequential batch AL selection of 90 images, 15 per video. As a result, the AL sampling would be faster on a reduced pool of samples, and it also served as a subsampling diversity sampling strategy, as it was presented in Section 4.4.2. Whereas in terms of model updating, it remained a 30 epoch training approach, with a batch size of 8, and with cross-entropy loss being computed every training and validation step. Even though there was a slight exploration of the hyperparameters, a learning rate scheduler was maintained to reduce the learning rate 0.1 every 7 epochs of training.

Additionally, all AL strategies were initially applied to the GIANA[1] pre-trained models, of which training process will introduce this chapter. The VGG16 with batch normalisation, and DenseNet121 fine-tuned on this dataset, will be presented as iteration 0 in the results.

Table 5.1: Datasets Selection and Transformations.

	Implementation	**Stage 0**	**Stage 1**	**Stage 2**	**Stage 3**	**Stage 4**	**Stage 5**
		FT	FT	FT TL	FT	TL	FT
Active Selection	**Video Selection**		6 videos per iteration				
	AL Strategies		A	B	B	B	C
	Batch Selection Strategies		K	K	K	K	G
	Final Batch		15 images per video, 90 in total				
Data Augmentation	**Employed Strategies**	BA	BA	BA	BA	BAR	BA
	Application	DG	DN	DN	DN	DCM	DN

FT: Fine-Tuning TL: Transfer-Learning

A: Uncertainty Sampling, Query-by-Committee B: Uncertainty Sampling, Query-by-Committee and Random Sampling C: Least Confident Sampling

K: Top K with Subsampling G: Greedy batch selection with Top K and Subsampling

BA: 90^{o}, 180^{o}, 270^{o} rotations and horizontal, vertical flip BAR: BA + repetition of class in minority

DG: Entire GIANA[1] dataset DN: New labelled Frames DCM: Class in minority of the new labelled frames

Table 5.2: Models' Updating and Testing approaches.

	Implementation	Stage 0	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
		FT	FT	FT TL	FT	TL	FT
Validation Dataset	**Validation Set**	VG	Vr	Vr Vb	Vb	VF	Vb
	Balanced	No	No	No Yes	Yes	Yes	Yes
Training Dataset	**Training Set**	TG	Tb	Tb Trt	Tb	Tls	Tb
	Balanced	No	No	No No	No	Yes	No
Training Parameters	**Used Models**	Aim	Aim	Aim Api	Aim	Base Model	VGG16 BN
	Epochs	30	30	30	30	30	30
	Batch Size	8	8	8	8	8	8
	Optimiser	SGD	SGD	SGD	SGD	Adam*	SGD
	Learning Rate	0.001	0.001	0.001	0.001	V	0.001
	Learning Rate Scheduler	Learning Rate is reduced by 0.1 every 7 epochs of training					
	Momentum	0.9	0.9	0.9	0.9	0	0.9
	Weight Decay	0	0	0	0	0	0
	Early Stop	-	-	E5	E5	E7	E5
Test Set		All Test Sets	T1	T2	T3	T3	T3

* with betas=(0.9, 0.999) FT: Fine-Tuning TL: Transfer-Learning

VG: 2718 images from the augmented GIANA[1] dataset Vr: 100 random images from the new selection Vb: 100 form the labelled set where at least half have abnormalities
VF: Fixed validation set with 400 samples

TG: 5436 images from the augmented GIANA dataset Tb: 440 New Set + Labelled Set + GIANA Trt: 440 New Set Tls: Entire Labelled Set

Aim: VGG16 Batch Norm. and Desnsenet121 pre-trained on the ImageNet Api: VGG16 Batch Norm. and Desnsenet121 pre-trained on the previous iteration

SGD:Standard Gradient Descent V: VGG16: 0.01, Dense: 0.001

E5: Training stops if validation loss non-improved after 5 epochs E7: Training stops if validation loss non-improved after 7 epochs

T1: 2 Clinical Cases - 836 images T2: 2 Clinical Cases - 1052 images T3: 6 Clinical Cases - 1200 images

5.1 Stage 0 - Base Model Initialisation

In this initial stage, the models pre-trained initially on the ImageNet database had all their layers fine-tuned by the GIANA[1] augmented training base dataset presented in 4.1.2. Here 30 epochs were set, with a batch size of 8 and with stochastic gradient descent (SGD) as the optimiser. The momentum was defined to 0.9, and the learning rate to 0.001, adopting a decay of 0.1 every 7 epochs. The loss was computed using the cross-entropy loss function in every training and validation step [49].

Their performance in the test sets of videos from the private data sets varied throughout the implementations, since the alterations in criteria also implied a re-annotation of the selected sets, as it will be explained in the following sections. Nonetheless, in all applications, the base models achieved a recall of 1 due to the originally unbalanced base dataset distribution with approximately $\frac{2}{3}$ of abnormalities compared to healthy frames. This was reflected on and initial $FN \simeq 0$, since the model is biased into classifying all instances as abnormal, of positive. The low FN rate resulted in a high number of FP at the start of every model sequential evolution, as it as will be seen in the results presented in all the coming implementations.

Although the initial results are not satisfactory, the growth of these networks with the AL strategies expects to trigger precisely this lack of performance and expand the models' perception on the diversity of normal and abnormal content that can be found in full VEC.

5.2 Stage 1 - Active Learning and Fine-tuning

In a starting approach, the same fine-tuning applied to the base set was employed using the sequentially labelled sampled sets, as it was introduced in Section 4.5.1.

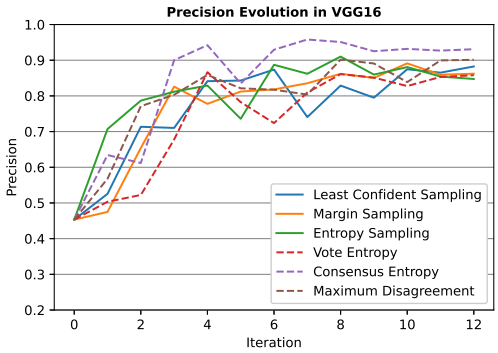
At each iteration, the 90^o, 180^o and 270^o rotation and horizontal and vertical flip data augmentation techniques were implemented to the introduced 90 image batch, given its small size, resulting in 540 new samples. Following the data augmentation, a 100 sized validation set is randomly selected from this distribution, as presented in Section 4.1.2. The other 440 instances are added to the labelled set, that is used for retraining the pre-trained CNNs, alongside the base dataset. After updating the models, the validation set is also included in the labelled dataset for future iterations.

The model update followed the same structure that was introduced in Section 5.1, with a 30 epoch training, on an 8 batch distribution with SGD as the optimiser. The momentum and learning rate evolution were maintained equally maintained, as well as the computation.

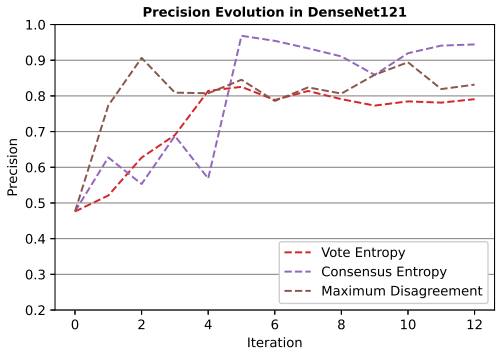
The test set used in this stage was the first mentioned in the private dataset subsection of Section 4.1.2, with a total number of 836 video frames, from which 378 were abnormal instances, and 458 were healthy.

5.2.1 Results

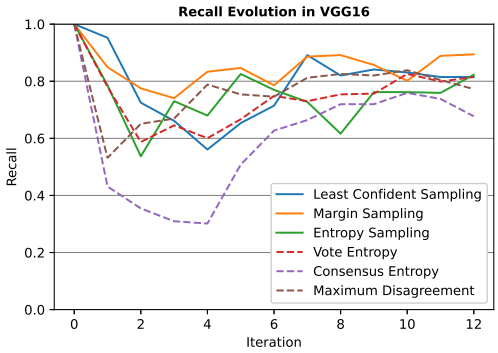
The presented results show the evolution of VGG16 with batch normalisation, and DenseNet121, respectively, for 12 iterations, or until $12 \times 90 = 1089$ new labelled samples are added to models' knowledge. No further sequential updating of the models was applied, due to the identified annotation errors, which will be explained at the end of this section. Figure 5.1 presents the three uncertainty sampling selection strategies in a continuous line, and three QBC strategies can be observed in a dashed line. The models' evolution was independently applied using the batches selected from each strategy, given that the uncertainty sampling measures were only used on the VGG16, and the QBC included both the models.



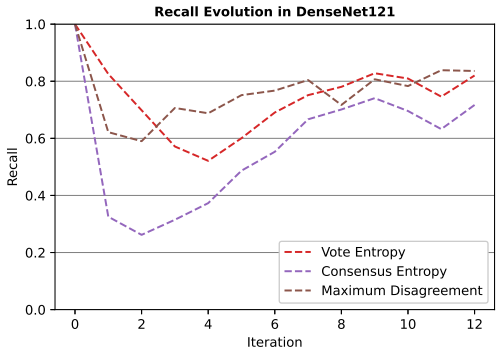
(a) Stage 1 - Precision evolution VGG16.



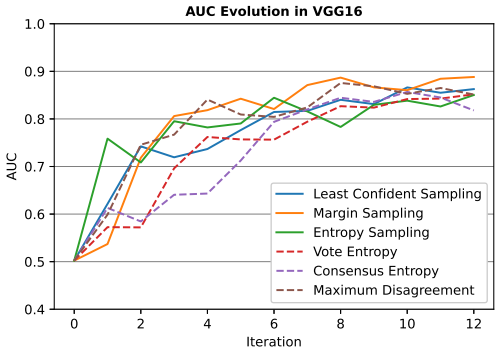
(b) Stage 1 - Precision evolution DenseNet121.



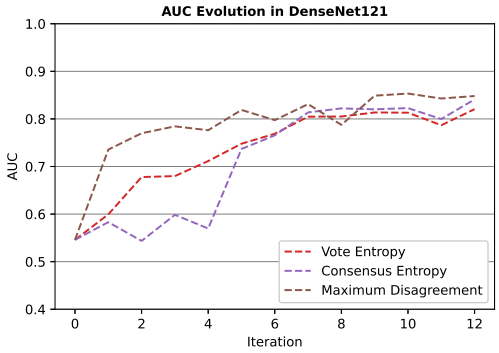
(c) Stage 1 - Recall evolution VGG16.



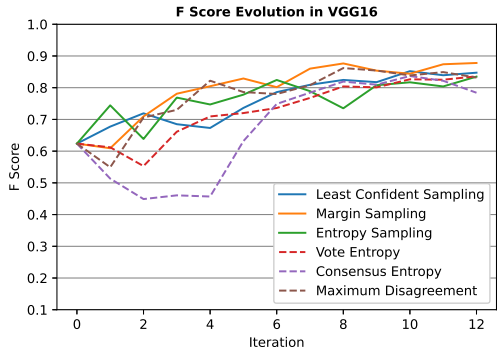
(d) Stage 1 - Recall evolution DenseNet121.



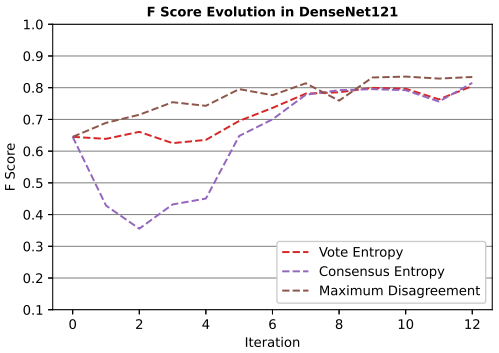
(e) Stage 1 - AUC evolution VGG16.



(f) Stage 1 - AUC evolution DenseNet121.



(g) Stage 1 - F Score evolution VGG16.



(h) Stage 1 - F Score evolution DenseNet121.

Figure 5.1: Results for the first stage.

5.2.2 Discussion

The obtained results suggest that the AL selection strategies allow for a fast evolution of the networks' knowledge over the VEC reality. In Figure 5.1, one can observe that there is a general conversion of all models to improved performances over the test set.

The recall evolution, started at 1 in both models, given the base model's biased behaviour, as explained in Section 5.1 which resulted in a high FP number of 456 samples over the 458 healthy frames, to an FN rate of 0. This justifies the negative evolution of the recall score in both models, starting with a substantial reduction of the performance that is further on predominantly corrected with the development of all AL selection strategies for both models.

The margin sampling strategy in the VGG16 seems to outperform all other selecting strategies from the start, reaching a final AUC of 0.89 and an F Score of 0.88, and especially in recall (Figure 5.1c) it appears to have the most stable evolution of all uncertainty sampling strategies. This measure is selecting the samples where the model is most unsure between the two classes (normal and abnormal). This response seems to trigger precisely the data that can enrich the model the most, particularly in reducing the rate of FN. Nonetheless, all uncertainty sampling strategies are contributing to eh VGG16 evolution expectedly. From the QBC strategies, max disagreement thrives in both models with a consistent confidence high AUC of 0.85 in the VGG and 0.85 in the DenseNet (Figures 5.1e and 5.1f), and F scores of 0.83 in both networks as well (Figures 5.1g and 5.1h), but it still behaves analogously to all uncertainty sampling measures. Consensus entropy seems to diverge the most in its behaviour which is curious, given its similarity to maximum disagreement sampling takes into account the classifiers' class probabilities to query the samples.

Overall, each AL strategies development seems to be stably increasing in Precision (Figures 5.1a and 5.1b) at the final iteration, with the lowest FP number of 12 to 836 in consensus entropy for the DenseNet121 (5.1b). Although the number of FN only reached a minimum of 40 instances with margin sampling, all sampling strategies indicate future recall improvements incrementations to the sets.

In the light these results, the learning processes proved to be very successful in its application for complete video analysis, continuously increasing onto the models' perception. However, technical medical issues were raised that needed improvement, which will be presented in the next section.

5.2.3 Criteria redefinition

Despite the achieved promising improvements with the implementation of the AL strategies in this experiment, after meeting with a specialist, it was pointed out by a physician that the annotation criteria needed a refinement, given the misinterpretation of equivocal frames.

Some annotation errors were identified in ambiguous frames which were considered abnormal when, in reality, they corresponded to images with normal content such as with debris or mucus, which could be easily confused with inflammatory lesions. Moreover, some venous shapes which appeared in the surface of the mucosa were mistaken for vascular lesions. Besides these errors, it was also noted that stomach and colon frames were also being included in the analysis. As mentioned in Section 4.1.2 of the private data set, they also required correction at this point.

With these issues brought up, the classification criteria that was passed to the model did not correspond to the correct clinical SB analysis required for the network to be used as a CAD. Nonetheless, it is worth noting that the models' exceptional evolution in performance with the VEC frame selection strategies, could only be refined by this criteria improvement and objectification for the medical problem at hand.

Thus, an entire relabelling of all the annotating data and retraining was required, correcting these nuanced cases for the following approaches. The training set was also replaced by the second mentioned in 4.1.2 with 1052 frames form the 2 previously selected clinical cases. The relabelling of the test set was also divided by two members of the team for a more robust annotation strategy. This labelling led to some disagreement in a few frames, which were presented to the specialist who clarified our annotation proposals.

5.3 Stage 2 - Active Learning with Annotation Criteria Correction

The correction of the criteria applied in the previous section demanded a re-implementation of the AL strategies, which generated the sampling of different actively selected batches, and its annotation. Hence, this stage applied the newly acquired insights of the more ambiguous cases in the VEC analysis to improve the CNNs' perception of the reality at hand. Before any further development, it is worth noting that the criteria redefinition ensued AL selected image sets with far fewer abnormalities than before, where, initially, each 90-sized batch selected, on average, 36 abnormalities, and now this was reduced to 14 in all AL strategies. This accounts for a reduction of 38% in the average abnormalities selected per batch, given that there was a significant amount of normal content that was misinterpreted. These alterations were also reflected in the abnormal content images identified in the newly labelled test set, which were only present in 151 frames of 1052.

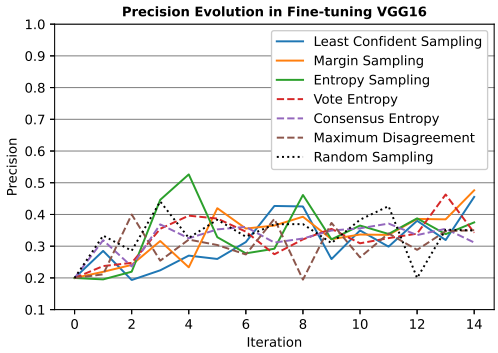
5.3.1 Fine-tuning

After the annotation correction, given the high success of the first approach, the same strategy was applied with the new labelling criteria. Given the misleading results that were achieved in the previous section, the same six AL strategies were applied to both VGG16, and DenseNet121 included only in QBC strategies. In this implementation, a random sampling strategy as also included to contrast with the active selection strategies. The only alteration to the proposed method in Section 5.2 was the insertion of the early stop routine at 5 epochs of non-improved validation loss.

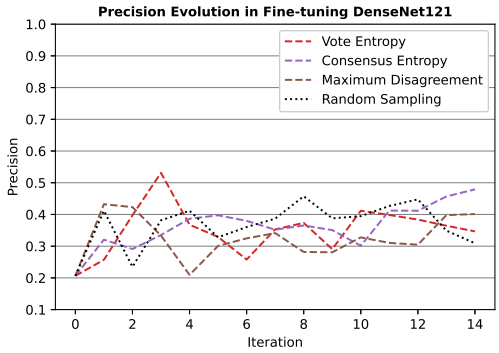
5.3.1.1 Results

The sequential evolution was applied over 14 iterations, in which $14 \times 90 = 1260$ samples were added to models' knowledge at the end of the set. Given the rapid evolution achieved in the first stage (Section 5.2), this annotation effort was estimated to be sufficient, mainly since the same coherence in annotation was maintained in the AL selected sets and the test sets.

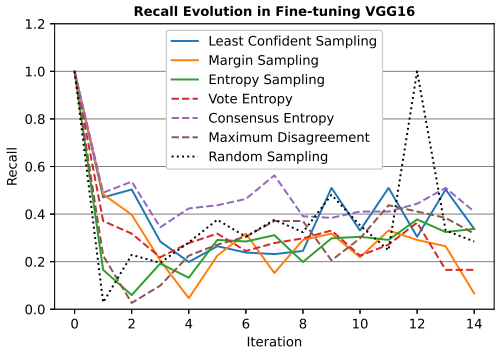
As before, Figure 5.2 presents the three uncertainty sampling selection strategies in a continuous line, and three QBC strategies can be observed in a dashed line, in its separate evolution. The new random sampling strategy can be observed in a dotted dark line.



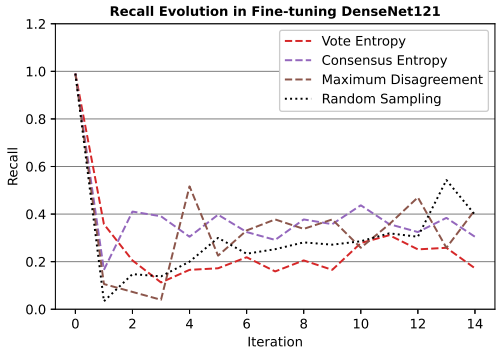
(a) Stage 2 - Precision evolution VGG16.



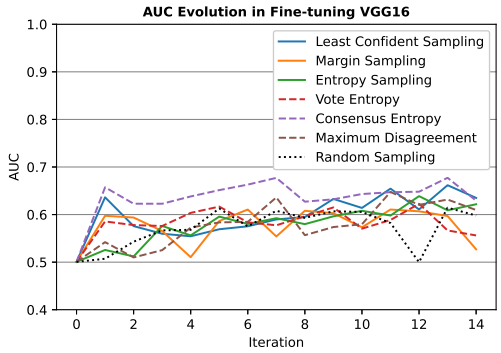
(b) Stage 2 - Precision evolution DenseNet121.



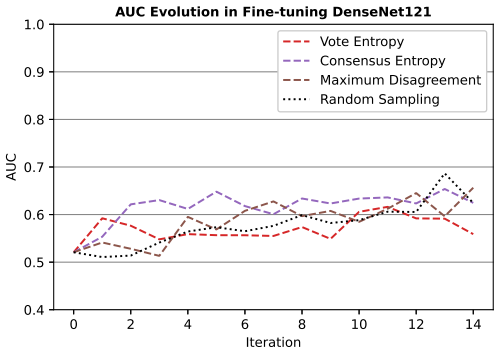
(c) Stage 2 - Recall evolution VGG16.



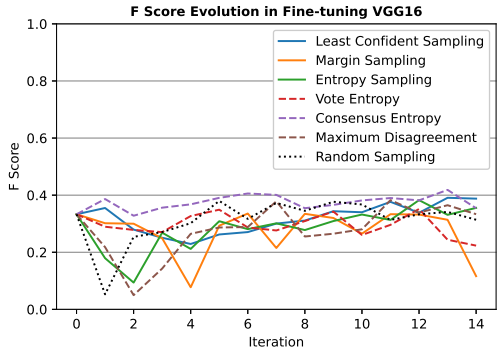
(d) Stage 2 - Recall evolution DenseNet121.



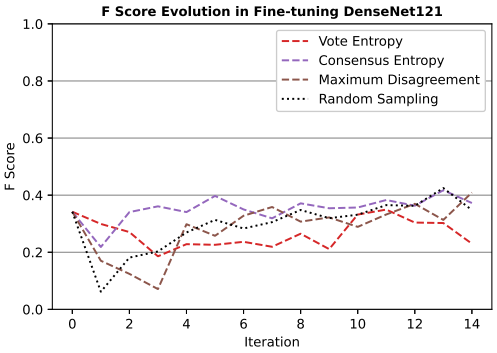
(e) Stage 2 - AUC evolution VGG16.



(f) Stage 2 - AUC evolution DenseNet121.



(g) Stage 2 - F Score evolution VGG16.



(h) Stage 2 - F Score evolution DenseNet121.

Figure 5.2: Results for the second stage with Fine-tuning.

5.3.1.2 Discussion

The obtained results in Figure 5.2 imply that the criteria redefinition had a considerable impact on the complexity of the full video analysis, characterised by a lower improvement rate on the evolution of the networks' perception. With a significantly lower number of ground truth positive instances, the precision values did not transcend 50% (Figures 5.2a and 5.2b), apart from spikes in Entropy iteration 4 (Precision=0.53) and Vote Entropy iteration 3 in the DenseNet121 model (Precision=0.53), which means that the number of FP most often exceeded the number of TP.

The similar behaviour could be noticed in the recall evolution (Figures 5.2c and 5.2d), although some iterations FN rate being exceeded by the TP, namely consensus entropy in 5.2c iteration 2, 7 and 13 (recalls of 0.54, 0.56 and 0.51 respectively) and max entropy in iteration 4 5.2b (recall=0.52). Given that the TP and FN were relative to the ground truth positive distribution, this was more likely to happen since, in precision, the FP was relative to a larger ground truth negative distribution. Overall, the AUC scores (Figures 5.2e and 5.2f) had better results than the F scores (Figures 5.2g and 5.2h) because the latter does not account for the TN distribution. The best achieving AUC with the VGG being 0.68, and 0.65 in the DenseNet both through the Consensus Entropy selection strategy. The best F Scores were also achieved by this selection strategy, reaching a score of 0.42 in both networks.

Overall, this means that the models had more difficulty at identifying the abnormalities, which was plausible given that the line between normal and abnormal content is now more slender, with a large predominance of normal content. Even when annotating, the selected sets were incrementally more nuanced, and it was more challenging to approach the heterogeneity of the content. The selected sets allowed for a relative improvement of this perception, but no further effort was invested because this fine-tuning approach was highly time-consuming.

In this experimental set, margin sampling is far from being the best performing AL strategy, having very few and oscillatory improvement compared to the concurrent strategies. In this case, one can observe that Least Confident sampling has a more steady evolution, and entropy did not fall far behind (Figures 5.2e and 5.2g).

The initial evolution of maximum disagreement batch selected models also performed poorly, but it evolved more rapidly from the fourth iteration on (Figures 5.2g and 5.2h). In this setup, consensus entropy had the best performance in comparison to all QBC strategies. It even exceeded uncertainty sampling strategies, in the VGG16, with a slow but steady evolution.

Random sampling even performed better at times, since it also is a viable sampling strategy with a higher chance to query instances different from each video. Furthermore, given the initial unprepared base set pre-trained models when facing the VEC reality, all samples can improve their behaviour. Regardless, this metric had a very unpredictable evolution with a recall spike in Figure 5.2c which was met with a slope in precision and AUC, Figures 5.2g and 5.2e, meaning that the performance was reset to the original.

The slow evolution of all strategies was believed to result from a higher selection of instances which were more complicated to categorise, thus more challenging to be annotated, and it was believed that noise was being added to the annotation. It is later concluded that it could be due to unbalanced training and validation set.

5.3.2 Transfer Learning

Transfer learning was explored in this implementation due to its advantageous faster updating since, in this case, only the classification layers are fine-tuned dynamising the annotation process. The same uncertainty sampling, QBC and random sampling strategies were applied, since, in previous fine-tuning developments, no strategy had a guaranteed best performance.

In this approach, the updated model, which served as a reference for the AL selection strategies, has their classification layers updated again on the new samples, resulting in the continuous updating strategy proposed in Section 4.5.2. Comparatively with the previous fine-tuning setups, and as can be observed in the table, the current implementation also includes the validation balancing strategy presented in the last chapter (Section 4.5.2). This alteration in validation attempted to compensate for the increased class unbalance verified after the criteria redefinition, and fine-tuning results obtained in Section 5.3.1.1. Nonetheless, this strategy also included an early stop routine when the training reached 5 consecutive epochs with no improvement of the validation loss.

Thus, in this setting, the newly acquired dataset augmented with the same transformations applied in Section 5.2, which enlarges the entire batch without any balancing strategy. The validation

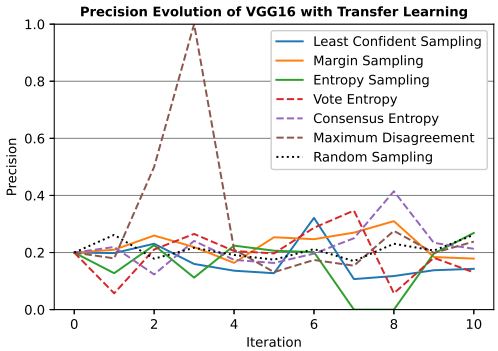
set was selected as presented in the second balanced approach in Section 4.1.2, which also had its implications in the training set. Hence, at each stage of the sequence of transfer learning updating, there is a training set of 440 frames, and 100 validation frames all from the 540 augmented new batch.

Transfer learning with this batch followed the same procedure as before, at 30 epochs with batches of 8, applied to the most recent version of the updated models. SGD is used as the optimiser with a learning rate of 0.001, which would be reduced to 0.1 every 7 epochs. Furthermore, an early stop routine was also employed, where the training would stop after 5 epochs of non-improved validation loss.

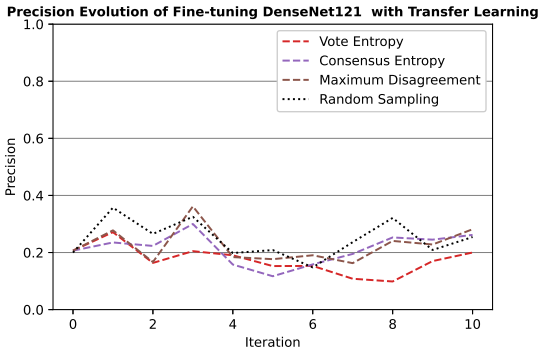
5.3.2.1 Results

The obtained results reflect on a 10 iteration evolution of the models with transfer learning. No further sequential updating was employed given that, as it will be observed in Figure 5.3.2, no coherent improvement was observed; thus, a continued annotation effort would be aimless.

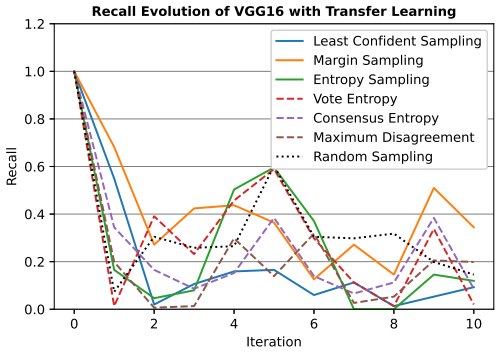
Similarly to previous results, Figure 5.3, the independent evolution of the models were represented, in a continuous line, for the three uncertainty sampling selection strategies, the three QBC strategies in a dashed line, and random sampling strategy can be observed in a dotted dark line.



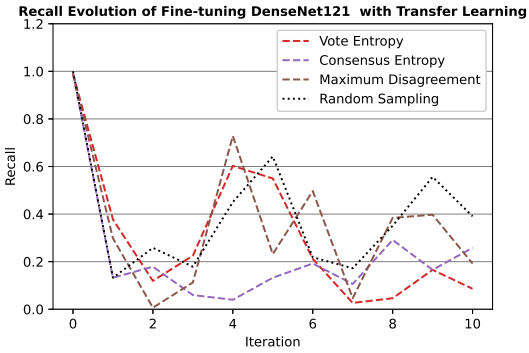
(a) Stage 2 - Precision evolution VGG16.



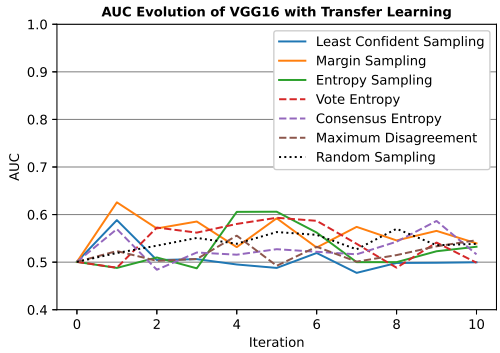
(b) Stage 2 - Precision evolution DenseNet121.



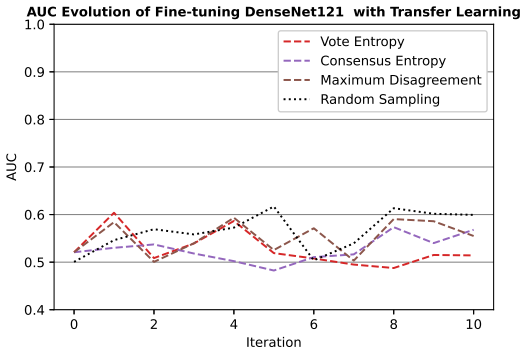
(c) Stage 2 - Recall evolution VGG16.



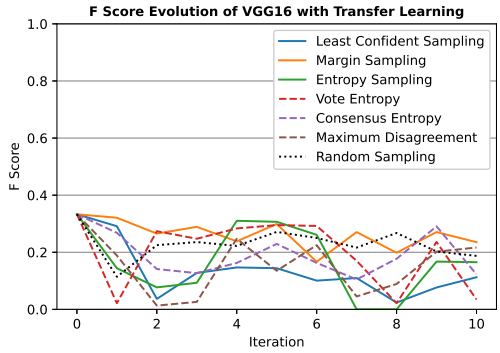
(d) Stage 2 - Recall evolution DenseNet121.



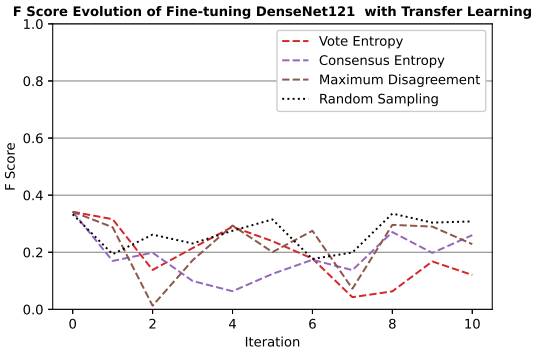
(e) Stage 2 - AUC evolution VGG16.



(f) Stage 2 - AUC evolution DenseNet121.



(g) Stage 2 - F Score evolution VGG16.



(h) Stage 2 - F Score evolution DenseNet121.

Figure 5.3: Results for the second stage with Transfer Learning.

5.3.2.2 Discussion

The obtained results represented in Figure 5.3 suggest that constant updating is not an eligible approach to the problem at hand. The models' inability to converge properly to improved results with the several AL strategies, indicates that the augmented 440 sized batch, is not sufficient to correct the classification layer of the models for its adequate application to the VEC analysis.

Despite the poor performance, it is interesting to observe in Figure 5.3g that the selected batches, in fact, have a different impact on the classifiers. The entropy selected batches, for example, had a wildly oscillating behaviour which meant that the classifier moved even further from the ideal performance in the test set after learning certain batches (iterations 1 to 3). However, in the following iterations, this poor classification was once again corrected by the newly selected batches (iterations 4 to 6). A similar pattern can be observed in the model evolution with maximum disagreement, both in 5.3g and 5.3h, and with vote entropy sampling in 5.3h, which suggest that the clinical cases analysed in iterations 2 were further from the reality in the two VEC selected for the test set. At the forth update it moved closer once again.

Nonetheless, ultimately, the F Score was unimproved for both models with all AL selected batches, with its initial performance reaching 0.33 and 0.34 for the VGG and DenseNet respectively. The AUC in was met with slight improvement, although the evolution of the models was

very oscillatory, with a final maximum improvement of only 0.04, in the VGG with the maximum disagreement selection strategy, over the initial 0.5 AUC. In the DenseNet, the AUC improvement was of 0.09 achieved by the random sampling strategy, regarding the initial 0.50. This evolution meant that either the small batches on themselves were unable to correct the classification of the base pre-trained models and that the features required fine-tuning as well, or that the learning parameters were inadequate for this transfer learning approach, due to their oscillatory behaviour.

5.3.3 Criteria refinement

The increased difficulty in the models' convergence to improved results observed above, suggested that there could be cases in which the ambiguity of some frames could be resulting in contradictory annotations for similar events. Hence, we decided to adjust the criteria for a more strict annotation process.

For that purpose, a test set of 30 ambiguous images was defined to be classified by two elements of the team, alongside a guiding specialist. The discussed guidelines were established after reaching a 90% agreement in the classification of the selected frames. The origin to most of the mistakable cases was associated with the limitation of isolated frame-wise analysis instead of video analysis, since there was no way to distinguish static abnormalities in the mucosa from debris floating in the SB, for example. Thus the interpretation consented of the content became more straightforward, in which only clear evidence of abnormality was considered.

Thus, on top of the initially defined criteria on the identification of abnormalities (Section 4.4.3), and to further fundament the previous adjustments (Section 5.2.3), the added guidelines were the following:

- Erosions and Ulcers would only be considered in case there was no deceiving presence of debris or mucus in the tract;
- Phlebectasias and diminute angiectasias would only be considered if there is no deceiving presence of venosities in the mucosa;
- Angiectasias like red spots, erythemas and blood should have an unmistakable red colouration, no dark or brown content was considered;
- Submucosal masses were considered in case there was clear evidence and it was conferred in the findings of the video, like in the case of Figure 4.4b;
- Lynfagiectasias were taken into account when they could not be mistaken for mucus or small bubbles on the villi;
- Moving images would not be considered unless the content displayed unmistakable evidence of abnormalities;

The criteria refinement once again involved the re-implementation of the work for an updating of the labels, which included the relabelling of the test set. Besides the initial two clinical cases

in the test set, four other clinical cases were included for a more generalised verification of the performance in the stages that followed, as it was mentioned in Section 4.1.2. Two elements of the team divided this task.

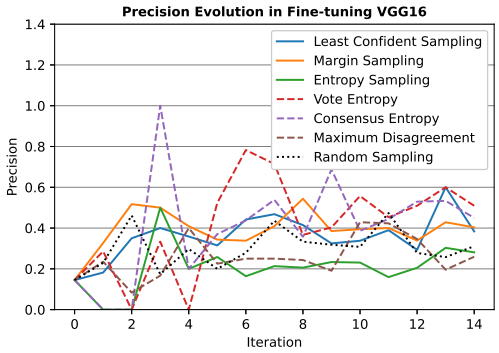
The four introduced clinical cases were expected to broaden the diversity of the testing content; however, the resulting test set still had a high discrepancy of abnormal instances of only 174 to 1026 healthy content.

5.4 Stage 3 - Active Learning and Fine-tuning with Refined Criteria

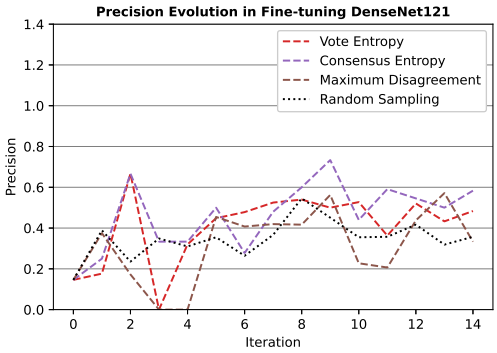
For this third implementation, the full fine-tuning setup applied in Section 5.3.1 was included, with data augmentation of all the newly introduced instances, and the same training and early stop routine likewise. The validation set, however, was balanced as in Section 5.3.2. To verify the criteria refinement progress, all the AL techniques and random sampling were implemented.

5.4.1 Results

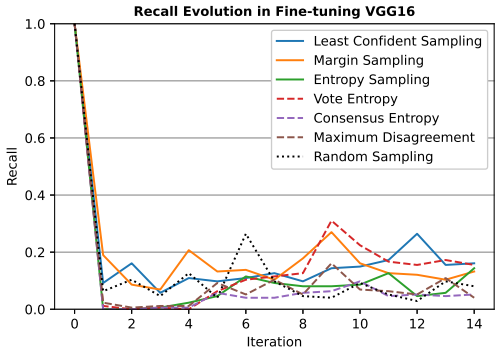
Similarly to stage e in Section 5.3.1, this experimental set took 14 training upgrades of the VGG16 and DenseNet121 models. The VGG16 evolved over the three uncertainty sampling strategies, represented by the continuous lines, and QBC and random sampling strategies were used on the VGG16 and DenseNet121, as dashed and dotted lines respectively.



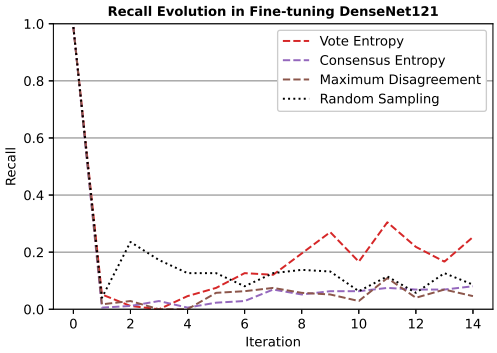
(a) Stage 3 - Precision evolution VGG16.



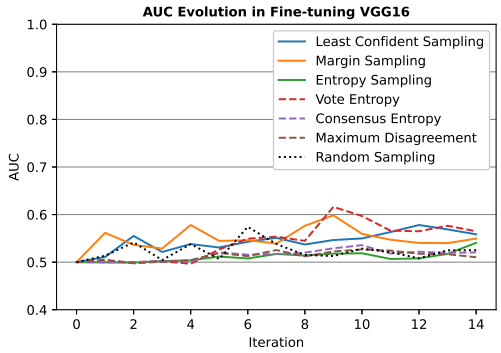
(b) Stage 3 - Precision evolution DenseNet121.



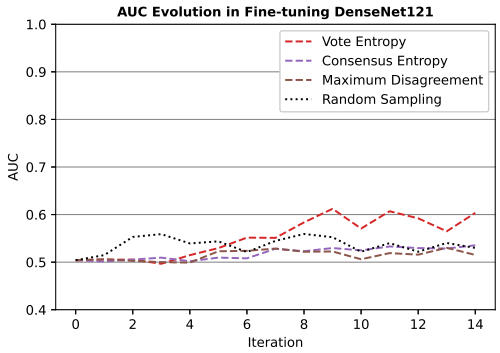
(c) Stage 3 - Recall evolution VGG16.



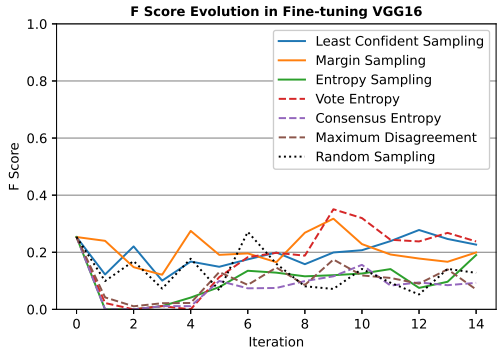
(d) Stage 3 - Recall evolution DenseNet121.



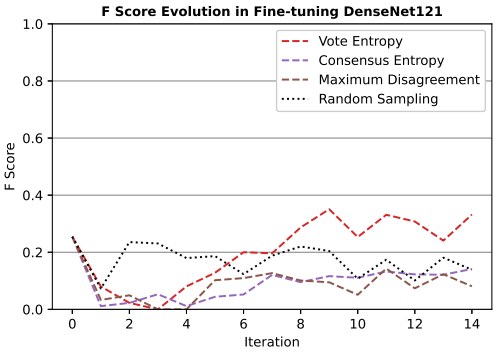
(e) Stage 3 - AUC evolution VGG16.



(f) Stage 3 - AUC evolution DenseNet121.



(g) Stage 3 - F Score evolution VGG16.



(h) Stage 3 - F Score evolution DenseNet121.

Figure 5.4: Results for the third stage with Fine-tuning.

5.4.2 Discussion

Figure 5.4 illustrates the resulting performance of the full fine-tuned models for all the employed selection strategies, with the new criteria and test set. The changes made to the data labelling did not have the expected effect on the models' evolution, which, as can be seen, continued to have a slow improvement in identifying the positive instances or the abnormalities, marked in the recall evolution in Figures 5.4c and 5.4d, with vote entropy selection in the VGG and DenseNet achieving a final recall of 0.16 and 0.25; This resulted from a low reduction of FN instances with the models' evolution. However, when compared to the results in 5.2, precision evolved more promptly, given the low number of FP and equally a low number of TP, with the best performing strategies reaching final values of 0.51 in the case of VGG with vote entropy, and 0.58 in the DenseNet with consensus entropy. Nonetheless, this meant that the high initial FP rate was rapidly reduced in the first iteration since it started with 0.15 precision rate in both models.

In general, where there is a higher increase in TP, there is also a rise of FP, which meant that the validation balancing strategy was depriving the test set of having newly queried abnormalities. Thus, the most successful strategy, vote entropy, probably had a higher selection rate of abnormalities which allowed for the test set to have more abnormalities for the models to learn. Nevertheless, with the increase of the viewed clinical cases in the iterative process, would result

in a rise of abnormal instances in the training set. Even if very few samples are added to the set at each iteration, it would probably result in a slow evolution of the curves like it is observed in Figure 5.4. The final best performance was achieved with Vote Entropy sampling for both models, in which, in the VGG16, the AUC improved from 0.5 to 0.56. The F score in this model, however, initially at 0.25, finished with a score of 0.23, although it had peaked at 0.35 in iteration 9 (Figure 5.4g), which indicates that the training in iteration 9 provided the model with a data distribution which allowed its knowledge to be closer to the test's reality. In the case of DenseNet121, for this same informativeness measure, the AUC evolved from 0.50 to 0.60, and the F Score from 0.25 to 0.33.

With the latter observed results, over more clinical cases of the new test set, it is estimated that slow progress would be noted, especially with the re-defined criteria. A continued annotation and retraining effort, however, would be very time-consuming. Apart from these requirements, the achieved results suggest that balancing the training set as well as the validation set could be a viable solution to the slow convergence of the models, which will be further explored in the following section.

5.5 Stage 4 - Active Learning and Transfer Learning on Increasingly Large Dataset

In this stage, transfer learning was selected for testing balancing the training sets. Since the transfer learning results obtained in Section 5.3.2.2 did not progress as expected, the model upgrading approach underwent a reassessment. Moreover, with the criteria alterations and lack of abnormal image selections, training set balancing strategies were worth being accessed. Thus, this Section presents the stage's entire setup as well as the acquired results and their analysis.

First of all, in this method, a fixed balanced validation dataset was handpicked as presented in Section 4.1.2. As it was explained in Section 4.5.2, the models' update, in this stage, always applies fine-tuning of the classification layers to the base pre-trained models presented in Section 4.2. This sequential updated included all the samples that had been actively selected up until then, and the training dataset balancing was approached in two different experimentation sets. However in both cases, contrary to previous implementations of data augmentation, the techniques of flipping and rotation that were only applied to the class in the minority of the set's distribution, which is most often the abnormal selected frames.

For the selection of a balanced training set, a first approach, would select an equally distributed set of frames from the entire labelled and augmented dataset at each epoch in training, as it is purposed in the AL book by Munro [36]. This approach forecasts an active selection convergency to an even selection of all classes with the model's evolution. Another balancing technique was applied, where the augmented elements in the minority were repeated until the batch set is balanced. Training is then carried on with all the labelled augmented samples.

Before full implementation of all the AL strategies, the best training set selection and hyperparameters were tested and applied to the models random batch selection to observe the consequent

evolution. All the testing was applied to the latter more diverse test set with the most recent annotation guidelines.

5.5.1 Hyperparameter Selection

To experiment with the hyperparameters and training data selection strategies, the sequential networks' performance evolution on the test set was monitored, by employing the random selection strategy, since this set was already pre-selected and annotated. The two balanced training set selection strategies employed were:

- Epoch-wise selection of a random set of balanced data from the entire labelled set distribution;
- Repetition of the set's elements in the minority to balance the whole set of labelled data.

Alongside these strategies, the hyperparameters explored were:

Table 5.3: Hyperparameters explored for Transfer Learning.

Hyperparameter	**Range Values**
Optimiser	SGD, Adam
Learning Rate	0.0001, 0.001, 0.01
Momentum	0, 0.5, 0.9
Weight Decay	0, 0.0001, 0.001

From the obtained curves, and their average performance on the test set, the best performing combination was the following:

Table 5.4: Training approach selected for Transfer Learning.

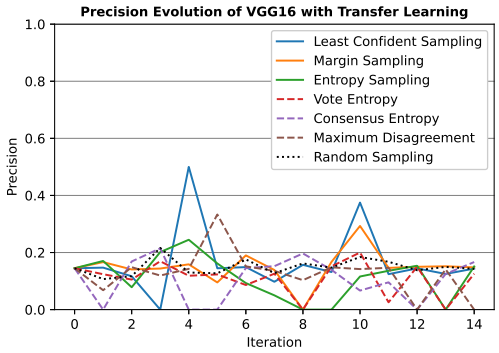
Hyperparameter	**Range Values**
Optimiser	Adam
Learning Rate	Densenet121: 0.001, VGG16: 0.01
Momentum	0
Weight Decay	0
Test Set	Repetition of the elements in minority

Accordingly, the implemented AL selected batches were included in the transfer learning approach presented in this section. Besides the presented training parameters introduced in this section, the same 30 epoch training with batches of 8 samples was employed. The Adam optimiser was applied with the standard betas at $0.9, 0.999$, and the same learning rate scheduler was

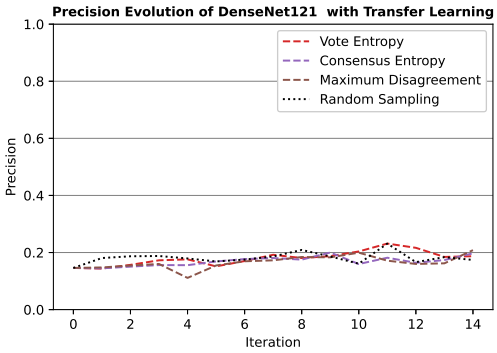
used. This time, however, the early stop routine was applied after 7 epochs of non-improved validation loss which was computed with cross-entropy loss function in every training and validation step.

5.5.2 Results

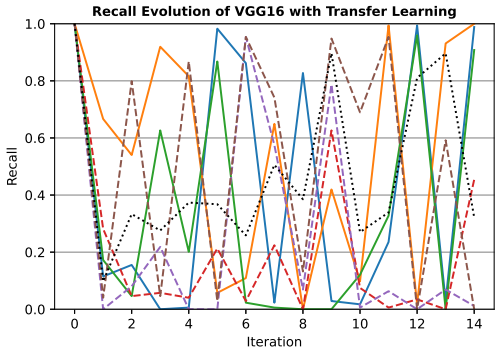
As in the previous result sections, this experimental set took 14 training upgrades of the VGG16 and DenseNet121 models. Three uncertainty sampling strategies, represented by the continuous lines, were applied to the VGG16, and the QBC and random sampling strategies were used on the VGG16 and DenseNet121, as dashed and dotted lines respectively. The VGG16 recall curves caption in 5.5c is equivalent to the presented in other VGG16 evolution curves in 5.6.



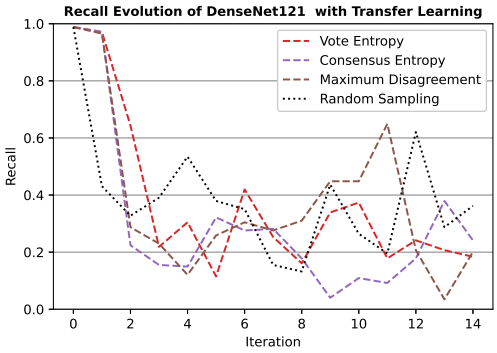
(a) Stage 4 - Precision evolution VGG16.



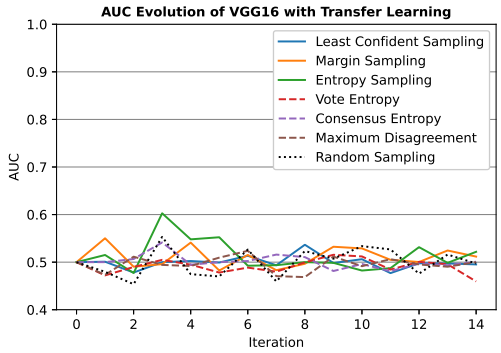
(b) Stage 4 - Precision evolution DenseNet121.



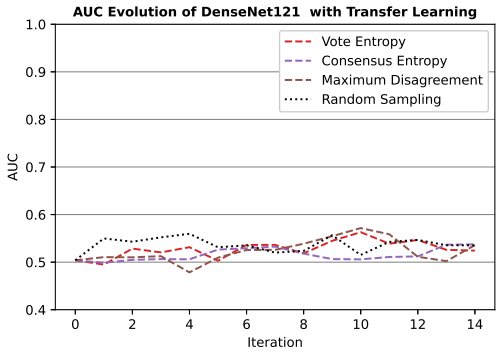
(c) Stage 4 - Recall evolution VGG16.



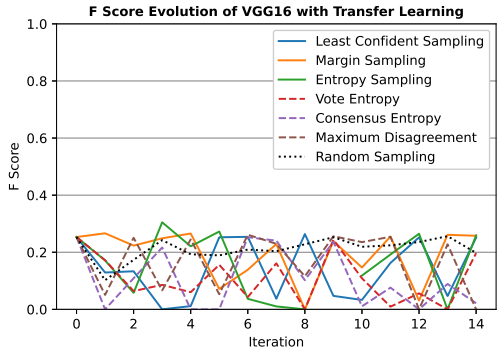
(d) Stage 4 - Recall evolution DenseNet121.



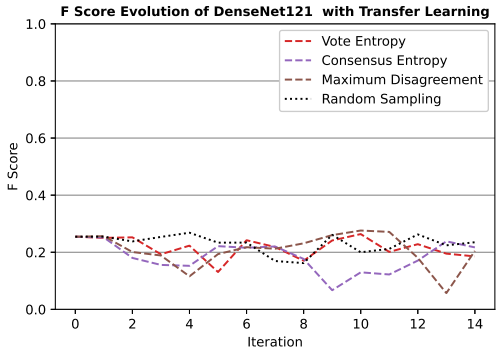
(e) Stage 4 - AUC evolution VGG16.



(f) Stage 4 - AUC evolution DenseNet121.



(g) Stage 4 - F Score evolution VGG16.



(h) Stage 4 - F Score evolution DenseNet121.

Figure 5.5: Results for the forth stage with Transfer Learning.

5.5.3 Discussion

The significantly worse and more disordered results were obtained in this implementation, as can be observed in 5.5. After both transfer learning implementations, although they vary significantly, suggest that transfer learning applied to the base models is not a viable way to approach this problem because the learned features from GIANA[1] are not sufficient to be adapted to the full VEC abnormality detection.

5.6 Stage 5 - Greedy Selection and Fine-tuning

After Section 5.4 achieved the best reliable results, with Least Confident (LC) uncertainty sampling strategy the best performing of the non-QBC, the same approach was applied, but this time, including the greedy batch selection strategy presented in Section 4.4.2. Only LC was employed given that the QBC strategies would require the training of 2 models, and in the full fine-tuning setup this would consume a lot of time.

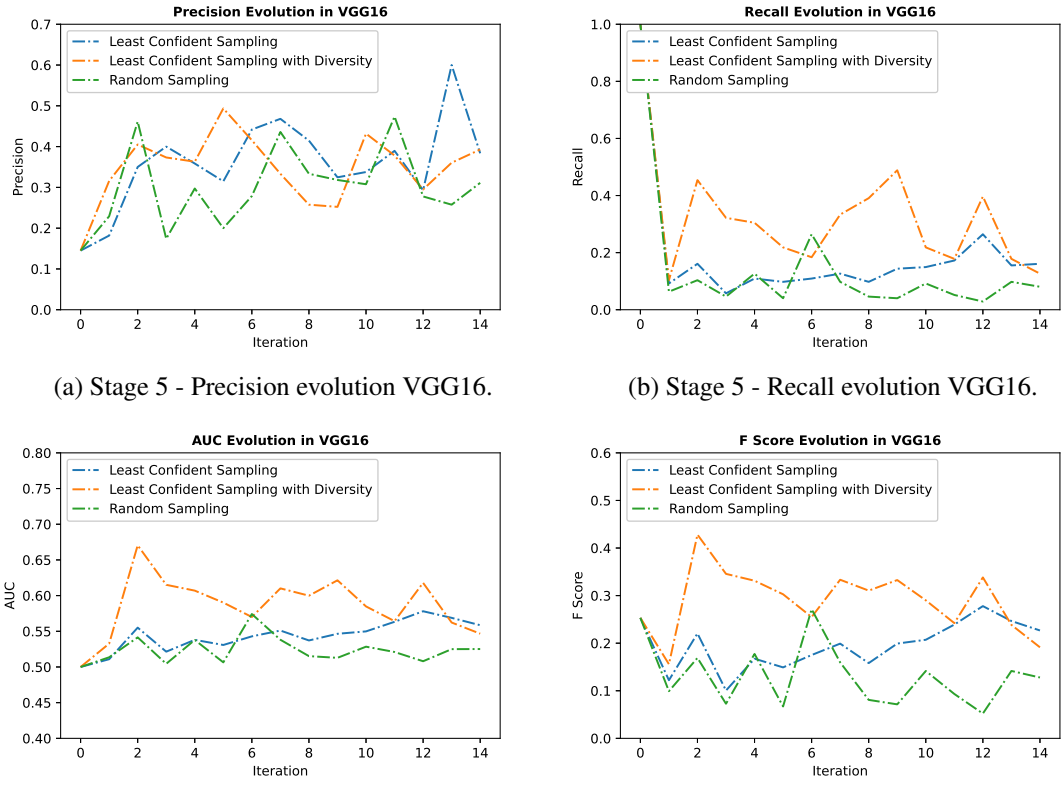
In the proposed approach, a more diverse set of samples would be selected per video to potentially enhance the sampling strategy by querying more distinct images and avoid redundant

labelling. For that purpose, both the simple LC sampling and LC with greedy batch selection applied to the evolution of VGG16 were compared in this final Section, alongside random sampling applied to the same model.

The images' similarity were compared with the ones in the batch through cosine similarity, and this served as the diversity measure for the information density computation. In this application, the weight of the informativeness measure was of 70% and the diversity of 30 %. The same training procedure expressed in Section 5.4 was applied to this implementation.

5.6.1 Results

The here present results show the evolution of the VGG16 with the iterative inclusion of the actively selected datasets. The curves show, in different colours, this evolution with LC sampling, LC with the diversity measure sampling, and random sampling.



(a) Stage 5 - Precision evolution VGG16.

(b) Stage 5 - Recall evolution VGG16.

(c) Stage 5 - AUC evolution VGG16.

(d) Stage 5 - F Score evolution VGG16.

Figure 5.6: Results for the third fifth with Diversity sampling.

5.6.2 Discussion

The obtained results in Figure 5.6 suggest that including the diversity measure resulted in an overall improvement of the selected batches enhancing the learning process from the beginning. The F Score and AUC score evolution of the models (Figures 5.6d and 5.6c) trained on the LC selected

batches, however, evolves more consistently than the new model with the diversity selection complement. The simple informativeness cumulative batch trained model even surpasses the new one in the final iteration, suggesting that the main subsampling diversity strategy employed in all the developed experiments was enough to ensure the selection of different elements in the batch. It is also possible that the cosine distance when applied directly to the images does not serve as a proper similarity measure, or that the information density balancing of 70% uncertainty measure to 30% diversity was not adequate for this application. Nonetheless, the rapid improvement, in the beginning, indicates that this implementation should be further investigated.

5.7 Summary

In this chapter, an initial stage and five different experiments conducted to experiment with AL strategies and model updating were presented. The evolution of the work was presented in its chronological order, including the criteria alterations the annotation process was subjected to.

The employed AL strategies included three uncertainty sampling, three QBC, a random selection and a final diversity sampling strategy. A total of six stages were performed, with essentially three different setups of fine-tuning and transfer learning updating. These setups were applied over various training actively selected sets, with different validation and training sets.

With the approximation of the criteria to the medical standard, a more complex data analysis was required when retraining and labelling the VEC data. This complex analysis reflected on the slower improvement of the learners' capability to perceive the VEC reality.

All things considered, from the achieved results, it was possible to conclude that:

- Transfer Learning updating techniques over models pre-trained on the base data set was not sufficient for the models to properly learn the full VEC content. On the other hand, full fine-tuning the networks on increasingly large selected sets of data achieved a slow increase in performance;
- From the employed stages of the developed work, there was no distinctively best performing active selection strategy;
- An increased annotation effort and expertise is believed to be at the core of more significant improvement in the learning evolution of the models;
- Including a diversity measure alongside the base informativeness measure can improve model evolution results.

Chapter 6

Conclusions and Future Work

This chapter will finalise the development of this dissertation. The strategies experimented with employed a new approach to developing a functional CAD to assist specialists in the exhausting reviewing required to analyse CE resulting videos, involving AL to improve abnormalities detection in full VEC. For that purpose, the present chapter will be divided into two sections, a first summarising of the dissertation's development and its results, and another exploring future work which can be further developed.

6.1 Conclusions

The introduction of CE has allowed for the close inner study of the SB, using a pill-shaped endoscope. When the endoscopy is performed, the screening results in an extensive video that requires a tedious and prone to errors reviewing from specialists. Hence, several approaches to automatic abnormalities detection have been explored to create a CAD to support the medical specialists on this task. The most successful strategies take advantage of ML methods, with DL approaches achieving the best efficiency. These tend to be trained and tested in selected sets of data, which result in a good performance in the set's reality, but are not sufficient when analysing full VEC. This output regards the nature of the selected sets, which fail to represent the diversity in content that found in the full videos. The frame-wise annotation of full VEC, however, is not a viable way to mitigate this problem, since it requires too much effort.

With this in mind, this dissertation explored AL strategies that could take advantage of unlabelled CE frames from full videos and aimed at selecting the most informative ones to optimise the model's learning, reducing the annotation cost. In the presented work, this was used to expand the base knowledge provided by a set of images from a public labelled dataset, and enhance the models' performance in abnormality detection. With the high success of DL on the task of lesion detection presented in the reviewed literature, especially CNNs, a VGG16 and a DenseNet121 pre-trained on ImageNet were selected and fine-tuned on a public VEC dataset to possess some background understanding of the CE reality.

A provided private dataset composed of full videos allowed for the incremental increase of these models' awareness based on the frames that would contribute with the most information on the unseen reality in VEC. For that purpose, several techniques of AL were implemented, exploring different sampling approaches, batch selection, and model updating strategies using fine-tuning and transfer learning.

Including full VEC, required in-depth assessment of their content for their integration in the AL cycle. The work developed includes a comprehensive analysis and preparation of the provided videos to construct datasets suitable for the CNNs' training. This investigation also allowed for establishing medically coherent criteria on the process of frame-wise annotation.

The first stage of the model revealed highly satisfying results for the abnormality detection in VEC, with both the VGG16 and the DenseNet121 converging rapidly into an improved performance with the increase of actively selected sets, for all the AL strategies employed, when applying a full fine-tuning on them. These results proved that the implementation of the AL strategies allowed for a classification improvement with fewer data than if we were to include fully annotated videos, thus reducing the effort of annotation efficaciously. Nonetheless, with the improved performance came an increase in the criteria strictness applied to the annotation process, where specialists noted that the initially defined criteria did not cover all the heterogeneities of the VEC content.

With the criterion closer to the medical perception, the frame annotation became more complex, and the abnormal selected content was significantly reduced and the classification learning process slew down considerably. To tackle these adversities, four other approaches were employed exploiting fine-tuning and transfer learning, with and without data balancing, including diversity measures and exploiting the models' hyperparameters. In the end, we concluded that the best approach would be to fine-tune the entire models for proper feature learning, and that including a diversity measure in the batch selection strategy can increase the informativeness of the whole batch and enhance the performance.

All in all, the main contribution of this dissertation lies in the proven application of AL accelerated the performance of CNNs in VEC abnormality detection; however, the annotation should rely on reliable and objective criteria, for which the medical assistance is vital.

6.2 Future Work

The work developed hereafter is the first, to the best of our knowledge, to investigate the impact of AL strategies on the VEC frames sampling to be applied in DL methods. Despite the proven results that batch sampling of the most informative content from full videos can significantly improve a CNNs' perception of the content present in VEC, objective annotation criteria should be able to cover all the heterogeneities of the VEC reality correctly. For that purpose, an improvement in the annotation strategy can be applied for increased performance of the active selecting strategies, and for the models to be able to approach abnormalities detection in VEC in an efficient manner for their integration in CAD. This could be achieved by including a full-time specialist in the AL

cycle; however, due to their annotation cost, alternative strategies could be explored regarding dealing with noisy oracles.

With the increase of abnormalities in the labelled dataset, a step of characterisation could be included. This CAD integration could involve the identification and segmentation of each of the abnormalities to also assist in the reviewing process.

The advantage of video analysis has over frame-wise analysis is that the first's allows for the visual distinction of static content from stagnant one. Thus, another approach that could be employed would involve a study of the correlation of content between frames, that could improve the distinguishing of abnormalities from normal debris and bubbles.

In the updating of the model, fine-tuning could be further explored to ascertain which layers of the model could be frozen for better exploitation of the features learned in the public set.

Additionally, the full VEC data integration, contrary to the GIANA[1], could benefit from some pre-processing of the frames. The AL strategies queried a variety of batches in which there was also the selection of very noisy and unclear content, which were not observed in a first analysis.

References

[1] Prashanth Rawla, Tagore Sunkara, and Adam Barsouk. Epidemiology of colorectal cancer: Incidence, mortality, survival, and risk factors. *Przeglad Gastroenterologiczny*, 14(2):89–103, 2019. doi:10.5114/pg.2018.81072.

[2] Marcellus Simadibrata and Randy Adiwinata. Precancerous Lesions in Gastrointestinal Tract. *The Indonesian Journal of Gastroenterology, Hepatology, and Digestive Endoscopy*, 18(2):112, 2017. doi:10.24871/1822017112-117.

[3] Juliane Flemming and Silke Cameron. Small bowel capsule endoscopy. *Medicine (United States)*, 97(14), 2018. doi:10.1097/MD.0000000000010148.

[4] Cristiano Spada, Deirdre McNamara, Edward J Despott, Samuel Adler, Brooks D Cash, Ignacio Fernández-Urién, Hrvoje Ivekovic, Martin Keuchel, Mark McAlindon, Jean-Christophe Saurin, Simon Panter, Cristina Bellisario, Silvia Minozzi, Carlo Senore, Cathy Bennett, Michael Bretthauer, Mario Dinis-Ribeiro, Dirk Domagk, Cesare Hassan, Michal F Kaminski, Colin J Rees, Roland Valori, Raf Bisschops, and Matthew D Rutter. Performance measures for small-bowel endoscopy: A European Society of Gastrointestinal Endoscopy (ESGE) Quality Improvement Initiative. *United European gastroenterology journal*, 7(5):614–641, jun 2019. doi:10.1177/2050640619850365.

[5] Miguel Muñoz-Navas. Capsule endoscopy. *World Journal of Gastroenterology : WJG*, 15(13):1584, apr 2009. doi:10.3748/WJG.15.1584.

[6] Anastasios Koulaouzidis, Dimitris K. Iakovidis, Alexandros Karargyris, and John N. Plevris. Optimizing lesion detection in small-bowel capsule endoscopy: From present problems to future solutions. *Expert Review of Gastroenterology and Hepatology*, 9(2):217–235, 2015. doi:10.1586/17474124.2014.952281.

[7] Limamou Gueye, Sule Yildirim-Yayilgan, Faouzi Alaya Cheikh, and Ilangko Balasingham. Automatic detection of colonoscopic anomalies using capsule endoscopy. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 1061–1064. IEEE, sep 2015. doi:10.1109/ICIP.2015.7350962.

[8] Yingju Chen and Jeongkyu Lee. A review of machine-vision-based analysis of wireless capsule endoscopy video. *Diagnostic and Therapeutic Endoscopy*, 2012, 2012. doi:10.1155/2012/418037.

[9] Dimitris K. Iakovidis and Anastasios Koulaouzidis. Software for enhanced video capsule endoscopy: Challenges for essential progress. *Nature Reviews Gastroenterology and Hepatology*, 12(3):172–186, 2015. doi:10.1038/nrgastro.2015.13.

[10] Jonathan Folmsbee, Xulei Liu, Margaret Brandwein-Weber, and Scott Doyle. Active deep learning: Improved training efficiency of convolutional neural networks for tissue classification in oral cavity cancer. In *Proceedings - International Symposium on Biomedical Imaging*, volume 2018-April, pages 770–773. IEEE, apr 2018. doi:10.1109/ISBI.2018.8363686.

[11] Neil Volk and Brian Lacy. Anatomy and Physiology of the Small Bowel. *Gastrointestinal Endoscopy Clinics of North America*, 27(1):1–13, 2017. doi:10.1016/j.giec.2016.08.001.

[12] Badireddy M. Collins JT. Anatomy, abdomen and pelvis, small intestine. Apr 2019.

[13] Jens Brøndum Frøkjaer, Asbjørn Mohr Drewes, and Hans Gregersen. Imaging of the gastrointestinal tract-novel technologies. *World journal of gastroenterology*, 15(2):160–8, jan 2009. doi:10.3748/wjg.15.160.

[14] Michael J Bartel, Mark E Stark, and Frank J Lukens. Clinical Review of Small-Bowel Endoscopic Imaging. *Gastroenterology & hepatology*, 10(11):718–726, nov 2014.

[15] Hyun Joo Song and Ki-Nam Shim. Current status and future perspectives of capsule endoscopy. *Intestinal research*, 14(1):21–9, jan 2016. doi:10.5217/ir.2016.14.1.21.

[16] Romain Leenhardt, Cynthia Li, Anastasios Koulaouzidis, Flaminia Cavallaro, Franck Cholet, Rami Eliakim, Ignacio Fernandez-Urien, Uri Kopylov, Mark McAlindon, Artur Németh, John N Plevris, Gabriel Rahmi, Emanuele Rondonotti, Jean-Christophe Saurin, Gian Eugenio Tontini, Ervin Toth, Diana Yung, Philippe Marteau, and Xavier Dray. Nomenclature and semantic description of vascular lesions in small bowel capsule endoscopy: an international Delphi consensus statement. *Endoscopy international open*, 7(3):E372–E379, mar 2019. doi:10.1055/a-0761-9742.

[17] Elizabeth Bollinger, Daniel Raines, and Patrick Saitta. Distribution of bleeding gastrointestinal angioectasias in a Western population. 18(43):6235–6239, 2012. doi:10.3748/wjg.v18.i43.6235.

[18] K Mönkemüller, AV Safatle-Ribeiro, C Olano, and LC Fry. Unusual Findings in the Small Bowel. *Video Journal and Encyclopedia of GI Endoscopy*, 1(1):286–288, jun 2013. doi:10.1016/S2212-0971(13)70125-0.

[19] Kyeong Ok Kim and Byung Ik Jang. Lessons from Korean Capsule Endoscopy Multicenter Studies Lessons from Korean Capsule Endoscopy Multicenter Studies. (September 2012), 2014. doi:10.5946/ce.2012.45.3.290.

[20] Maria Teresa Gomes Silva. Abnormalities Detection on Videos of Endoscopic Capsules (VEC). *Tese de Mestrado, Faculty of Engineering of the University of Porto*, 2019.

[21] Patel Janakkumar Baldevbhai. Color Image Segmentation for Medical Images using L*a*b* Color Space. *IOSR Journal of Electronics and Communication Engineering*, 1(2):24–45, 2012. doi:10.9790/2834-0122445.

[22] Stéphane Vignes and Jérôme Bellanger. Primary intestinal lymphangiectasia (Waldmann ' s disease). 8:1–8, 2008. doi:10.1186/1750-1172-3-5.

[23] Guo Bing Pan, Guo Zheng Yan, Xin Shuai Song, and Xiang Ling Qiu. Bleeding detection from wireless capsule endoscopy images using improved euler distance in CIELab. *Journal of Shanghai Jiaotong University (Science)*, 15(2):218–223, 2010. doi:10.1007/s12204-010-9716-z.

[24] Cadman L. Leggett and Kenneth K. Wang. Computer-aided diagnosis in GI endoscopy: looking into the future. *Gastrointestinal Endoscopy*, 84(5):842–844, nov 2016. doi:10.1016/j.gie.2016.07.045.

[25] Zhen Ding, Huiying Shi, Hao Zhang, Lingjun Meng, Mengke Fan, Chaoqun Han, Kun Zhang, Fanhua Ming, Xiaoping Xie, Hao Liu, Jun Liu, Rong Lin, and Xiaohua Hou. Gastroenterologist-Level Identification of Small-Bowel Diseases and Normal Variants by Capsule Endoscopy Using a Deep-Learning Model. *Gastroenterology*, 157(4):1044–1054.e5, 2019. doi:10.1053/j.gastro.2019.06.025.

[26] Vinu Sankar Sadasivan and Chandra Sekhar Seelamantula. High accuracy patch-level classification of wireless capsule endoscopy images using a convolutional neural network. *Proceedings - International Symposium on Biomedical Imaging*, 2019-April(Isbi):96–99, 2019. doi:10.1109/ISBI.2019.8759324.

[27] Michael Vasilakakis, Anastasios Koulaouzidis, Diana E. Yung, John N. Plevris, Ervin Toth, and Dimitris K. Iakovidis. Follow-up on: optimizing lesion detection in small bowel capsule endoscopy and beyond: from present problems to future solutions. *Expert Review of Gastroenterology and Hepatology*, 13(2):129–141, 2019. doi:10.1080/17474124.2019.1553616.

[28] Muhammad Sharif, Muhammad Attique Khan, Muhammad Rashid, Mussarat Yasmin, Farhat Afza, and Urcun John Tanik. Deep CNN and geometric features-based gastrointestinal tract diseases detection and classification from wireless capsule endoscopy images. *Journal of Experimental and Theoretical Artificial Intelligence*, 00(00):1–23, 2019. doi:10.1080/0952813X.2019.1572657.

[29] Michael D. Vasilakakis, Dimitris Diamantis, Evaggelos Spyrou, Anastasios Koulaouzidis, and Dimtris K. Iakovidis. Weakly supervised multilabel classification for semantic interpretation of endoscopy video frames. *Evolving Systems*, 0(0):0, 2018. doi:10.1007/s12530-018-9236-x.

[30] Baopu Li and Max Q.H. Meng. Ulcer recognition in capsule endoscopy images by texture features. *Proceedings of the World Congress on Intelligent Control and Automation (WCICA)*, pages 234–239, 2008. doi:10.1109/WCICA.2008.4592930.

[31] Qian Wang, Ning Pan, Wei Xiong, Heng Lu, Nan Li, and Xijing Zou. Reduction of bubble-like frames using a RSS filter in wireless capsule endoscopy video. *Optics & Laser Technology*, 110:152–157, feb 2019. doi:10.1016/J.OPTLASTEC.2018.08.051.

[32] Dimitris K. Iakovidis, Spiros V. Georgakopoulos, Michael Vasilakakis, Anastasios Koulaouzidis, and Vassilis P. Plagianakos. Detecting and Locating Gastrointestinal Anomalies Using Deep Learning and Iterative Cluster Unification. *IEEE Transactions on Medical Imaging*, 37(10):2196–2210, 2018. doi:10.1109/TMI.2018.2837002.

[33] Yun Gu, Mali Shen, Jie Yang, and Guang Zhong Yang. Reliable Label-Efficient Learning for Biomedical Image Recognition. *IEEE transactions on bio-medical engineering*, 66(9):2423–2432, 2019. doi:10.1109/TBME.2018.2889915.

[34] Samuel Budd, Emma C Robinson, and Bernhard Kainz. A Survey on Active Learning and Human-in-the-Loop Deep Learning for Medical Image Analysis. 2019. arXiv:1910.02923.

[35] Burr Settles. Active learning literature survey. 07 2010.

[36] Robert Munro. Human-in-the-loop machine learning, 2020.

[37] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(4):623–656, 1948. doi:10.1002/j.1538-7305.1948.tb00917.x.

[38] Asim Smailagic, Pedro Costa, Hae Young Noh, Devesh Walawalkar, Kartik Khandelwal, Adrian Galdran, Mostafa Mirshekari, Jonathon Fagert, Susu Xu, Pei Zhang, and Aurelio Campilho. Medal: Accurate and robust deep active learning for medical image analysis. *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, pages 481–488, 2019. doi:10.1109/ICMLA.2018.00078.

[39] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, pages 1–12, 2018. arXiv:1708.00489.

[40] Zuobing Xu, Ram Akella, and Yi Zhang. Incorporating diversity and density in active learning for relevance feedback. volume 4425 LNCS, pages 246–257. Springer Verlag, 2007. doi:10.1007/978-3-540-71496-5_24.

[41] Tianxu He, Shukui Zhang, Jie Xin, Pengpeng Zhao, Jian Wu, Xuefeng Xian, Chunhua Li, and Zhiming Cui. An active learning approach with uncertainty, representativeness, and diversity. *Scientific World Journal*, 2014, 2014. doi:10.1155/2014/827586.

[42] Signe Ledou Nielsen, Peter Ingeholm, Susanne Holck, and Ian Talbot. Xanthomatosis of the gastrointestinal tract with focus on small bowel involvement. *Journal of Clinical Pathology*, 60(10):1164–1166, 2007. doi:10.1136/jcp.2005.035261.

[43] Jaime Pereira Rodrigues, Rolando Pinho, Adélia Rodrigues, Joana Silva, Ana Ponte, Mafalda Sousa, and João Carvalho. Validation of SPICE, a method to differenciate small bowel submucosal lesions from innocent bulges on capsule endoscopy. *Revista Espanola de Enfermedades Digestivas*, 109(2):106–113, 2017. doi:10.17235/reed.2017.4629/2016.

[44] M^{a} Teresa Galiano-de Sanchez. Obscure Gastrointestinal Bleedingl. In *Atlas of Capsule Endoscopy 2*, pages 274–284. IEEE, apr 2012.

[45] Martin Keuchel, Nageshwar Reddy, and Fritz Hagenmuller. Infectious Diseases of the Small Bowel. In *Atlas of Capsule Endoscopy 2*, pages 274–284. IEEE, apr 2012.

[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[47] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:2261–2269, 8 2016. URL: http://arxiv.org/abs/1608.06993.

[48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 9 2014. URL: http://arxiv.org/abs/1409.1556.

[49] PyTorch. Pytorch master documentation — torchvision.models.