UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Industrial

Diseño y desarrollo de un sistema automático para la predicción de la colitis ulcerosa aplicando técnicas de aprendizaje profundo débilmente supervisado sobre Whole-Slide Images.

Trabajo Fin de Máster

Máster Universitario en Ingeniería Biomédica

AUTOR/A: Meseguer Esbrí, Pablo

Tutor/a: Naranjo Ornedo, Valeriana

Cotutor/a: Amor del Amor, María Rocío del

CURSO ACADÉMICO: 2021/2022

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ESCOLA TÈCNICA SUPERIOR ENGINYERIA INDUSTRIAL VALÈNCIA

BIOMEDICAL ENGINEERING MASTER THESIS

DESIGN AND DEVELOPMENT OF AN AUTOMATIC SYSTEM FOR THE ULCERATIVE COLITIS PREDICTION WITH WEAKLY SUPERVISED BASED DEEP LEARNING TECHNIQUES IN WHOLE SLIDE IMAGES

AUTHOR: PABLO MESEGUER ESBRI

SUPERVISOR: VALERY NARANJO ORNEDO

CO-SUPERVISOR: ROCÍO DEL AMOR DEL AMOR

Academic year: 2021-22

Greetings

To my family for always being by my side supporting me, to Rocío for her invaluable help during all the project, to Valery for the opportunities and to the rest of the CVBLab colleagues for making this stage a better experience.

Abstract

Ulcerative colitis (UC) is a chronic inflammatory bowel disease (IBD) affecting the colon with an incidence of around 20 cases per 100.000 inhabitants. The aim of treatment in UC is achieving inflammatory remission and histological evaluation takes a crucial role in the follow-up to get histological remission (HR), the main target of treatment nowadays. A novel index has been proposed to evaluate the ulcerative colitis activity in biopsies and it is based on the presence or absence of neutrophils in different regions of the tissue. To get a clinical diagnosis in UC, pathologists used to analyze under the microscopes the biopsies obtained from the rectum and the colon sigmoid during endoscopic procedures. However, with the advances in digital pathology biopsies can be digitized in whole-slide images (WSIs) offering the possibility of being shared across clinical centres to be analyzed by different pathologists. This procedure reduces the subjectivity associated to the pathologist in the grading of the biopsy. The implementation of computer-aided diagnosis systems (CAD) can help to overcome the commented handicaps. For this reason, the present project aims to develop a predictive algorithm based on weakly supervised deep learning techniques for the prediction of histological remission in WSI. Specifically, we intend to implement an algorithm based on multiple instance learning (MIL) and apply constraints to the convolutional neural network to force it to learn how to locate neutrophils, key cells for the prediction of histological remission. For that purpose, a database of WSI corresponding to patients with an ulcerative colitis diagnosis will be available along with pixel and image-level annotation of each biopsy.

Keywords: Weakly supervised deep learning, ulcerative colitis, histological imaging, histological remission.

Resumen

La colitis ulcerosa (UC) es una enfermedad intestinal inflamatoria crónica que afecta al intestino grueso con una incidencia cercana a los 20 casos por 100.000 habitantes. Para el tratamiento de la UC es fundamental eliminar la inflamación intestinal y para su seguimiento, la evaluación histológica juega un papel fundamental puesto que la remisión histológica (HR) es el principal indicador de enfermedad. Para cuantificar la presencia o ausencia de actividad de colitis ulcerosa se ha propuesto un índice histológico que define la HR a partir de la presencia o ausencia de neutrófilos en diferentes compartimentos celulares. Para obtener un diagnóstico clínico de esta enfermedad, los patólogos solían analizar bajo ek microscopio las biopsias obtenidas del recto y del colon sigmoide durante los procedimientos endoscópicos. Sin embargo, con los avances de la patología digital las biopsias pueden ser digitalizadas en *whole-slide images* (WSI) ofreciendo la posibilidad de ser compartidas entre centros clínicos para ser analizadas por diferentes patólogos. Este procedimiento reduce la subjetividad asociada al patólogo durante la gradación de la biopsia. La implementación de sistemas de diagnóstico asistidos por ordenador (CAD) puede ayudar a solucionar esta problemática. El objetivo de este TMF reside en el desarrollo de un algoritmo predictivo basado en técnicas de aprendizaje profundo débilmente supervisado para la predicción de la remisión histológica en WSI.En concreto, se pretende implementar un algoritmo basado en multiple instance learning (MIL) y aplicar restricciones a la red neuronal convolucional para forzar a que aprenda como localizar los neutrófilos, células clave para la predicción de la remisión histológica. Para ello, se contará con una base de datos de WSI correspondientes a pacientes con colitis ulcerosa junto con la anotación a nivel de pixel y de imagen de cada una de ellas.

Palabras clave: Aprendizaje profundo debilmente supervisado, colitis ulcerosa, imagen histológica, remisión histológica

Resum

La colitis ulcerosa (UC) és una malaltia intestinal inflamatòria crònica que afecta l'intestí gros amb una incidència pròxima als 20 casos per 100.000 habitants. Per al tractament de la UC és fonamental eliminar la inflamació intestinal i per al seu seguiment, l'avaluació histològica juga un paper fonamental perquè la remissió histològica (HR) és el principal indicador de la malaltia. Per a quantificar la presència o absència d'activitat de colitis ulcerosa s'ha proposat un índex histològic que defineix la HR a partir de la presència o absència de neutròfils en diferents compartiments cel·lulars. Per obtenir un diagnòstic clínic d'aquesta patologia, els patòlegs solien analitzar sota els microscopis les biòpsies obtingudes del recte i del còlon sigmoide durant els procediments endoscòpics. Tanmateix, amb els avancos de la patologia digital les biòpsies poden ser digitalitzades en *whole-slide images* (WSI) oferint la possibilitat de ser compartides entre centres clínics oferint per a ser analitzades per diferents patòlegs. Aquest procediment redueix la subjectivitat associada al patòleg en la qualificació de la biòpsia. La implementació de sistemes de diagnòstic assistits per ordinador (CAD) pot ajudar a solucionar aquesta problemàtica. L'objectiu d'aquest projecte resideix en el desenvolupament d'un algoritme predictiu basat en tècniques d'aprenentatge profund feblement supervisat per a la predicció de la remissió histològica en WSI.En concret, es pretén implementar un algoritmee basat en Multiple Instance Learning (MIL) i aplicar restriccions a la xarxa neuronal convolucional per a forçar al fet que aprenga a localitzar els neutròfils, cèl·lules clau per a la predicció de la remissió histològica. Per a això, es comptarà amb una base de dades de WSI corresponents a pacients amb colitis ulcerosa juntament amb l'anotació a nivell de píxel i d'imatge de cadascuna d'elles.

Paraules clau: aprenentatge feblement supervisat, colitis ulcerosa, imatge histològica, remissió histol'gica

List of acronyms

AI	Artificial Intelligence
AC	Attention Constraints
CAD	Computer-Aided Diagnosis
CAM	Class Activation Maps
CNN	Convolutional Neural Network
DL	Deep Learning
GAP	Global Average Pooling
HR	Histological Remission
IBD	Inflammatory Bowel Disease
MIL	Multiple Instance Learning
PHRI	PICaSSO Histologic Remission Index
UC	Ulcerative Colitis
WSL	Weakly Supervised Learning
WSI	Whole-Slide Image

Contents

List of Figures

List of Tables

Part I

Memory

Chapter 1

Introduction

Contents

1.1 Motivation and description of the problem

Ulcerative Colitis (UC) is an inflammatory bowel disease (IBD) located in the colon with a rising incidence surrounding 20 cases per 100.000 inhabitants. UC is characterized by a relapsing-remitting cycle, which means that a patient diagnosed with UC could stand in large periods without symptoms (remitting phase) followed by others with more disease activity (relapsing phase) [39].

UC can be classified as proctitis, left-sided colitis, or extensive colitis, depending on the extension of the colon affected. The probability of suffering negative relapsing phases is related to the portion of the colon damaged with a higher level of active inflammation [33]. The risk factors for suffering UC are yet to be clarified. It has been observed that patients with a history of UC in the family have a higher probability of suffering the disease, however the predictive capability of genomics is still limited [4]. The most common signs and symptoms for patients in relapsing phase include blood presence in the stool, incontinence, and weight loss.

The aim of treatment in UC has evolved in recent years and nowadays focuses on achieving histological remission (HR). HR is also called mucosal healing and has a stronger association with favorable clinical outcomes than remission at the endoscopic level. Patients with a UC diagnosis undergo clinical procedures such as colonoscopies, an intervention during which endoscopists extract biopsies from the colon for further analysis. Pathologists used to analyze these specimens under the microscope to look for the characteristic patterns of the disease.

With the advancements in digital pathology, biopsies are digitalized into whole-slide images (WSI). WSIs are giga-pixel images that can be shared across medical centers and studied by different pathologists. However, the most interesting point is to use WSI in computer-aided diagnosis (CAD) systems that offer the possibility of implementing automatic algorithms to perform image analysis and prediction.

Considering the clinical context, this project pretends to develop an automatic algorithm to predict histological remission and activity in WSI from patients with a UC diagnosis. The deep learning algorithm is based on weakly supervised learning (WSL) techniques and specifically in multiple instance learning (MIL). It incorporates novelties such as attention constraints to asses the network learning process and improve the results. For that purpose, we use an extensive private database that includes WSIs of UC patients together with pixel-level annotation for some biopsies and image-level annotation for all of them.

1.2 Project framework

This work is part of the research project called *"Development of Artificial Intelligence using i-scan videos and digital histological images"* coordinated by the University of Birmingham. This project is also known as *PICASSO* due to the development of a novel endoscopic score for ulcerative colitis called Paddington International virtual ChromoendoScopy ScOre (PICaSSO) [12].

The data used in this project was obtained under the study "A multicenter, international validation study of the i-scan endoscopic scoring system and a new histologic scoring system to define subtle mucosal inflammation in ulcerative colitis" (i-scan) approved by the West Midlands Research Ethics Committee (17/WM/0223).

The main goal of PICASSO is to develop an integrated system for the management of ulcerative colitis (UC) patients that incorporates both endoscopic procedures and histological analysis. For that purpose, a novel score (PICaSSO) for endoscopic grading of UC was developed and it showed better interobserver agreement than other endoscopic scoring indexes [12]. They also developed the PICaSSO Histologic Remission Index (PHRI) which is a simplified and neutrophil-based scoring system for monitoring mucosal healing at the histological level [8]. Its main power in comparison with previous histologic remission indexes is the fact of being more simple and neutrophil-based which reduces the subjectivity associated with pathologists. This also makes the score suitable for being used in deep learning-based computer-aided diagnostic (CAD) systems.

The UPV, specifically the Computer Vision and Behavior Analysis Lab (CVBLab), participates in this project with the design and development of automatic algorithms for the analysis of WSI of patients with a UC diagnosis. These algorithms are based on deep learning techniques, and their goal is to classify the obtained WSI in histological remission or activity according to the novel, simplified, and neutrophil-based scoring system called PHRI.

1.3 Project goals

The main goal of this project is to design and develop artificial intelligence-based algorithms for the automatic analysis and prediction of histological remission in WSIs of patients with ulcerative colitis (UC). A series of secondary objectives must be addressed to achieve the main goal:

1. Put together the database of histological images. The high-resolution WSIs are coming from an international multicenter study so they have to be stored together handling handicaps with their huge file size ranging from 500MB to 3GB. The biopsies were obtained and digitalized following similar procedures in multiple centers and afterward shared with the corresponding annotations of the pathologist.
2. Carry out a literature review to study the techniques in the state-of-the-art (SoA) for the classification of WSI with deep learning algorithms.

3. Carry out the preprocessing of the database. This process includes detecting the WSI with missing ground truth and binarizing the pixel-level annotation to differentiate the regions with and without neutrophils. It also includes binarizing the ground truth of PHRI, which ranges from 0 to 4, into histological remission and UC activity with a cutoff equal to 0.
4. Carry out the patient-level partition of the database. These partitions are conditioned by the availability of the pixel-level annotation in certain WSI.
5. Design and develop a classifying algorithm based on deep learning for the prediction of HR and UC activity in WSI from patients with a UC diagnosis.
6. Implementation of constraints to the deep convolutional neural networks to improve the classification performance of the algorithms in the classification of histological images.
7. Perform multiple ablation studies in the validation dataset to determine the best combination of hyperparameters that leads to better results.
8. Compare the external validation results with other techniques.
9. Propose future lines for the improvement of the problem resolution after taking into account the problems and limitations faced during the project.

1.4 Document structure

In Chapter 2, ulcerative colitis is described in terms of epidemiology, scoring, and management. A brief review of the CAD systems implemented for UC management, which focus on endoscopy, are also presented alongside the novel index for histologic scoring in UC called PHRI. This chapter also introduces weakly supervised and multiple instance learning and its applications in histopathology image analysis.

In Chapter 3, the materials used in the project are detailed with a special focus on the dataset containing hundreds of whole-slide images from patients with a UC diagnosis. The dataset also contains pixel-level annotations of neutrophils for certain biopsies.

Chapter 4 first includes the methodology for WSI preprocessing and a description of the data partitions carried out. It also contains the development of classification algorithms to predict histological remission in WSI.

Results and discussion for the different experiments carried out are presented in Chapter 5. This section not only includes the external validation metrics in the test set, it also incorporates the ablation experiments results, such as Class Activation Maps (CAM) that improve network interpretability.

Finally, Chapter 6 is composed of the project's conclusions and some notes about limitations and future lines of investigation in this field. It concludes with the contribution to the scientific knowledge.

Chapter 2

Theoretical framework

Contents

2.1 Ulcerative Colitis

This section pretends to introduce the reader to ulcerative colitis. It firstly includes some details about the epidemiology, management and scoring of the disease and afterward presents some computer-aided diagnosis systems application in ulcerative colitis.

2.1.1 Definition, epidemiology and management

Ulcerative colitis (UC) is a chronic inflammatory bowel disease (IBD) characterized by a relapsing and remitting mucosal inflammation course that mainly affects the colon [39]. Being a relapsing-remitting disease means that a patient diagnosed with UC could stand in large periods of less severe symptomatology (remitting phase) followed by others where the disease symptoms get worse (relapsing phase). This is a crucial point because a patient with a UC diagnosis could not present active inflammation at the endoscopic or histologic level, so it is necessary to differentiate between active UC and remitting inflammation or remission.

UC can be distinguished in proctitis, left-sided colitis and extensive colitis according to the Montreal Classification [33]. As it is presented in Figure 2.1, the classification depends on the affected portion of the colon, which can vary from just the rectum in proctitis to the whole colon in extensive colitis. Relapsing probability and disease severity are highly correlated to the portion of the colon with active inflammation [33].

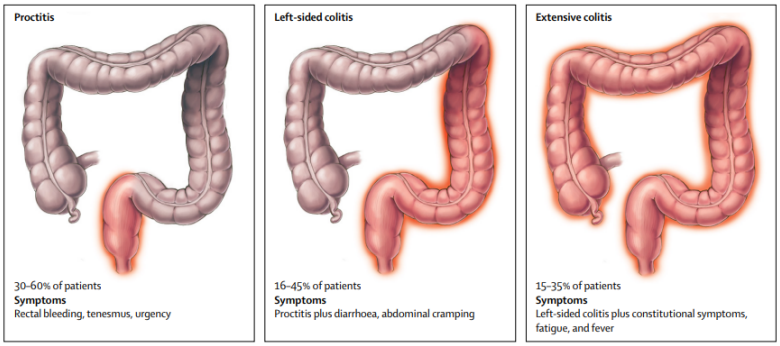

Figure 2.1: Different ulcerative colitis extensions according to the Montreal Classification: proctitis, left-sided colitis and extensive colitis. Obtained from [39].

Ulcerative Colitis incidence (i.e. the number of new observed cases in a population over a certain period of time) is rising rapidly worldwide and ranges between 19.2 in Canada and 24.3 cases per 100.000 inhabitants in Europe [39]. Both regions are also among the territories with the highest prevalence (i.e. number of patients with the condition in a population) with 248 and 505 cases per 100.000 inhabitants, respectively. The incidence value gives an idea about the expansion speed of the disease because it measures the new cases while the prevalence informs about the total number of patients. These epidemiology statistics translate to the clinical practice with an increased number of patients to manage which corresponds to a larger amount of data and medical images to be analyzed by clinicians. That means that computer-aided diagnosis (CAD) systems can take an important role in disease management with a special focus on artificial intelligence (AI) methods that have proven to perform well in the medical image classification problem.

One of the most common clinical presentations of UC is the presence of blood in the stool. Patients with more severe disease can present a large range of symptoms that includes incontinence, fatigue, abdominal pain, and weight loss. Among the risk factors, it has been found that patients with a history of inflammatory bowel disease in the family have a higher probability of developing ulcerative colitis. However, it has also been demonstrated that the predictive capacity of genetics is still limited for efficient clinical use.

Management of UC patients has evolved considerably in the last few years due to the development of novel exploratory techniques. Management was initially focused on inducing and maintaining clinical remission, which means that the aim of treatment was to prevent patient disability and colorectal cancer. The next protocol implemented for disease assessment was endoscopy. Endoscopy remission or mucosal healing became the gold standard for disease prognosis because endoscopy evaluation does not depend on the highly subjective clinical symptoms [3]. Ulcerative colitis has recently included histologic remission as the aim of treatment. In [27], authors found that UC patients in histological remission present a significantly lower relative risk of suffering clinical relapse in comparison with patients with active histological inflammation.

As management can be carried out at the endoscopic and histologic levels, multiple tissue characteristics are disrupted when the inflammatory activity increases. Among the endoscopic features, we find the modification of the vascular pattern and the bleeding of the tissue. In the same vein, the histological patterns that can be observed in biopsies with active colitis include crypt architecture, mucin depletion, and inflammatory infiltration with cells like eosinophils and neutrophils [18]. In Figure 2.2, the relation between the view at the endoscopic level and the histopathological examination is presented for the healthy (upper row) and the active (lower row) tissue.

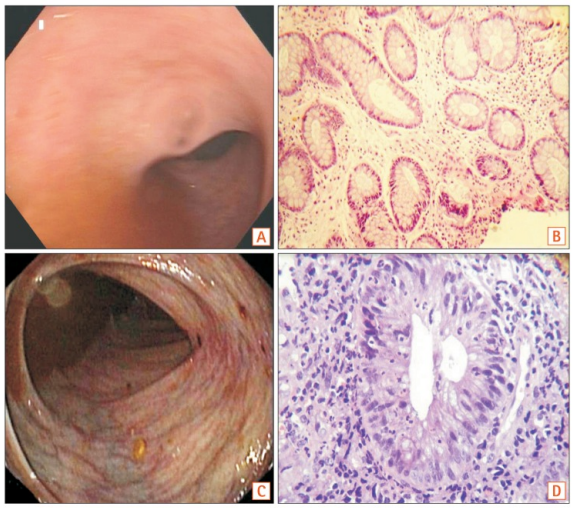

Figure 2.2: (A) Healthy colon tissue. (B) Normal histopathological examination. (C) Endoscopic image with active inflammation. (D). Histopathological examination with evidence of cryptitis. Obtained from [23].

2.1.2 *Scoring on Ulcerative Colitis*

Here, some indices for both endoscopic and histologic evaluation are presented with a special focus on a recently developed index that will be the focal point of the deep learning algorithm.

- **Endoscopic evaluation**: Endoscopic scoring in ulcerative colitis presents a handicap related to the subjective evaluation of the endoscopic findings which considerably determine the final evaluation depending on the clinician's perspective.
  - *Mayo Score*: This score focuses on four keys which include stool frequency, rectal bleeding, findings in endoscopic procedures, and the physical assessment. It grades each of them on a scale from 0 to 3. This index was developed in 1987 and was first used in a clinical trial to evaluate the activity of a new therapy against UC [29].
  - *UCEIS*: The Ulcerative Colitis Endoscopic Index of Severity (UCEIS) was developed back in 2011 and it was more focused on endoscopic assessment than the Mayo score. This scoring system considered up to ten descriptors and finally selected three of them to define the final score. These are the vascular pattern, bleeding, and erosions/ulcers and are graded with the most severe lesion. [38]
- **Histologic evaluation**: In a similar line, histological indices also present a certain component of subjectivity related to the scorer because certain patterns could have different interpretations. This fact complicates its applicability to automatic CAD systems.
  - *Robarts Histologic Index (RHI)*: RHI score is calculated as a linear combination of different components such as ulcerations, inflammatory infiltration, and neutrophils presence. Each of this components of the score that can be graded from 0 to 4 each one [21].
  - *Nancy Histologic Index (NHI)*: NHI grading system follows a decision tree (see Figure 2.3) that outputs a five grade scale to obtain the final score. The components that it considers are ulceration, acute inflammatory cell presence, and chronic infiltrate. Notice that the components are fairly the same in comparison to the RHI score, but NHI avoids grading each of them in a 0-to-4 sub-scale. [19]

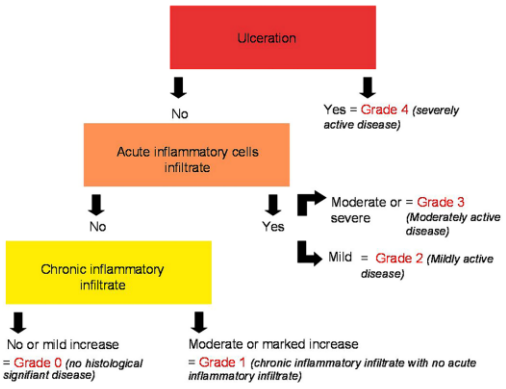

Figure 2.3: The NHI scoring system follows a decision tree. Obtained from [**sha**].

However, this work will mainly focus on a recently developed histological score for ulcerative colitis grading. The PICaSSO Histologic Remission Index (PHRI) is a novel simplified histologic index to determine mucosal healing or disease activity in patients with a UC diagnosis [8].

Note that one of the components that RHI and NHI consider is the neutrophil infiltration in the mucosal tissue. Neutrophils are one of the three main types of white blood cells along with basophils and eosinophils. Each of these types presents its own shape of nucleus and cytoplasm. Here it stands out the characteristic multi-lobed nuclei in the neutrophils as shown in Figure 2.4a. In the other hand, eosinophils (see Figure 2.4b) contain a red cytoplasm and basophils (see Figure 2.4c) are very difficult to be separated due to the nature of the nuclei [31]. PHRI aimed to develop a 'neutrophil-only' evaluation of specimen biopsies for assessing histological remission.

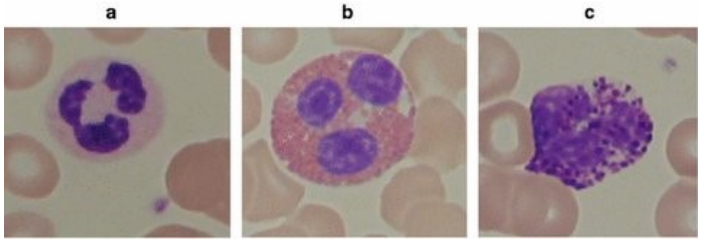

Figure 2.4: Different types of white blood cells. (a) neutrophils, (b) eosinophils and (c) basophils. Obtained from [31].

PHRI analyzes the neutrophils infiltration in four different compartments of the biopsy: lamina propria, surface epithelium, cryptal epithelium, and cryptal lumen. The lamina propria is the connective tissue located around the crypts and normally contains a variable number of plasma cells such as neutrophils and eosinophils depending on the inflammation activity. The surface intestinal epithelium is composed of a single layer of epithelial cells that act as border between the tissue and the external medium. Finally, the crypts are tube-like glands composed of a delimiting cryptal epithelium that surrounds the cryptal lumen. Visual examples for each of the four compartment from a zoomed-in WSI are presented in Figure 2.5.

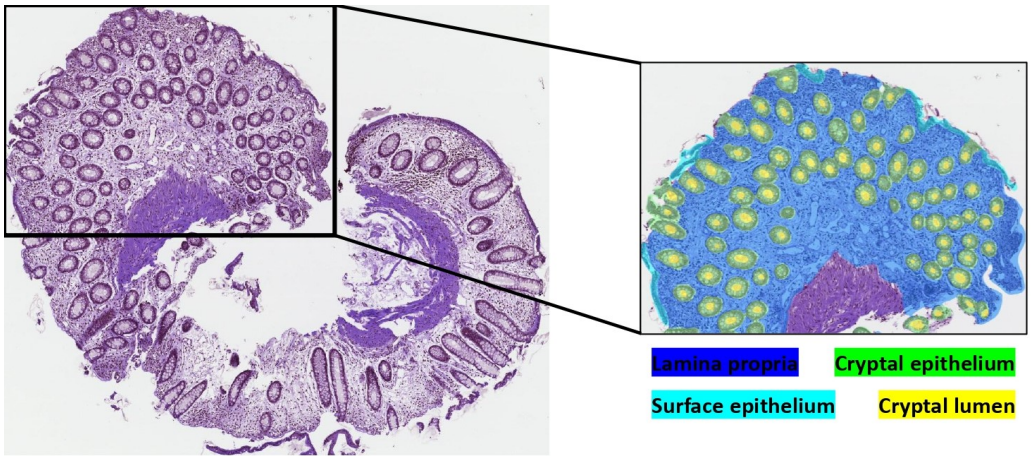

Figure 2.5: Different compartments in the biopsy. (a) WSI, (b) Zoomed-in WSI. Own elaboration.

The pixel-level annotations shown in Figure 2.5 indicating the four regions of interest were performed by experienced pathologists of the PICaSSO project. These are only used in this work to present the four compartments that are considered in the PHRI scoring system to get the grade of each biopsy. They are not contained in any of the datasets presented in the following chapters.

The PHRI scoring systems consists in the observation and analysis of the four regions of interest in the biopsy. In each of this four compartments, the neutrophil infiltration due to presence of ulcerative colitis activity is evaluated. The PHRI grade for each single biopsy is obtained by adding the number of regions with significant histological findings. Its scoring protocol is presented in Table 2.1. Here, its simplicity in comparison with other histological indices, such as NHI that follow a decision tree, is denoted. Note that the minimum score for the index is zero when there is absence of neutrophils while the maximum is four when the neutrophils infiltration is present in all the compartments.

Table 2.1: PICaSSO Histologic Remission Index (PHRI) to predict histological remission. Obtained from [8]

Histologic finding	Score
Neutrophil infiltration in lamina propria	
Absent (No)	0
Present (Yes)	1
Neutrophil infiltration in epithelium	
Absent (No)	0
Present (Yes)	
- Surface epithelium	1
- Cryptal epithelium	1
- Crypt lumen	1
Total Score = sum of all above (maximum 4)	

As you may note, PHRI grades each biopsy on a five-grade scale (0 to 4). The risk of suffering exacerbation episodes is related to the grade of the biopsy. However, the authors proposed to define histological remission according to PHRI with a cut-off equal to zero . That means that patients with a PHRI equal to zero are considered in histological remission, while others with PHRI greater than zero present UC activity. This cut-off proved to stratify well the patients between low and high risk of adverse outcomes. That means that biopsies with PHRI equal to zero are considered as HR and in the other case biopsies contain UC activity [8].

The idea behind PHRI development was to design a simple grading system for ulcerative colitis to reduce inter-observer variability and subjectivity when the pathologists analyze the sample under the microscope. The score is aligned with endoscopy and proved to correlate well with clinical outcomes. Moreover, the inter-rater agreement among pathologist was excellent and surpassed other histological scores [8]. At the same time, the simplicity that characterizes PHRI makes it suitable to be applied in an AI-based system.

2.1.3 CAD systems in Ulcerative Colitis

Due to the recent advances in digital computing and medical technology, computer-aided diagnosis (CAD) systems have emerged as a supporting tool for the decision-making of clinicians. CADs combine multi-modal image processing, extensive data analysis, pattern recognition, and artificial intelligence (AI) among others to assist the diagnostic process [14]. Artificial intelligence is widely used in these systems due to the capability of its algorithms to discover relevant information from large amounts of data. One use case for CAD is real-time clinical decision support, for example, to perform live video analysis during endoscopic explorations. Model interpretability is crucial when implementing AI-based algorithms in medicine. The vast majority of deep learning methods work like a black box which means that the deep model gets its output (i,e, classification label) from the input data with an opaque behavior [9].

As mentioned before, the incidence of ulcerative colitis is rising worldwide which is translating into an increased workload in the hospitals. This fact boosts the implementation of CAD systems to assess clinicians. The first approaches in ulcerative colitis focused on endoscopic imaging and video. These were designed to be used during the endoscopic explorations and their aim was to predict endoscopic remission. In [36], a deep neural network called DNUC was built to evaluate endoscopic images of patients with UC and predict endoscopic remission defined by the UCEIS score. The work [25] also analyzed individual endoscopic images with a convolutional neural network (CNN) based on the GoogleNet architecture [35]. In comparison with the other project, they used the Mayo endoscopic score as ground truth to define endoscopic severity between remission, moderate or severe. At the same time, both CADs were designed to classify individual video frames so a global prediction of the patient video was not obtained.

Focusing on the histopathology field, advances in computational histology and research in specific CAD systems have been exhaustively explored. This is because histopathological images are the gold standard for cancer diagnosis, which focuses the attention of researchers and clinicians because of their clinical and social impact. The most relevant advance in this field is the digitalization of histological tissue samples in Whole-Slide Images (WSI). WSI are giga-pixel images of dyed biopsy slides that in common clinical practices were observed under the microscope, but computational pathology allows their digitalization. The main handicap that WSIs present for AI-based applications is their huge complexity due to the immense number of pixels, but at the same time it is possible to analyze them at different resolutions depending on the desired magnification [24].

Despite histological assessment in ulcerative colitis being critical, only one work has focused on analyzing whole-slide images (WSI) to predict UC [40]. It is focused on eosinophils, white blood cells that have been associated with the disease because it accumulates in mucosal tissue and can be observed in slices under the microscope. They designed a deep learning algorithm that measures the eosinophils density in the tissue of sigmoid colon biopsies. In this case, the histological scores used to assess disease were the Goebes score and Robarts histopathology Index (RHI). So, to the best of the author's knowledge, no previous study has analyzed WSIs to predict ulcerative colitis activity or remission.

2.2 Weakly Supervised Learning (WSL)

This section introduces the concept of weakly supervised learning, a novel branch of deep learning in which focuses the algorithms developed in this project.

2.2.1 Deep Learning

Deep learning (DL) falls within the field of machine learning and artificial intelligence. AI consists of the development of automatic systems capable of carrying out tasks that were previously reserved for humans. In the same vein, machine learning aims to design mathematical models with the ability to make predictions or decisions after analyzing the input data in an automatic training phase.

Within the field of deep learning, deep algorithms have been used in computer vision tasks such as image classification, object detection, or semantic segmentation. For that purpose, deep models are built with a stack of processing layers and activation functions with trainable parameters that are optimized to get an output that differs very little from the expected. The training process is characterized by back-propagation that consists of adjusting the model weights according to the learning rate to reduce the network error.

The ground truth is the target variable, what we want the model to predict. Deep learning-based algorithms are classified into three types depending on the ground truth availability. Supervised learning when the ground truth is available for all the data; unsupervised learning when the ground truth is unknown and semi-supervised learning when only a part of the data contains the ground truth while the rest is not annotated.

Network interpretability is crucial in the deep learning field, specifically in the medical field. For that purpose, class activation maps (CAMs) were designed to determine the relevant features of the image that the CNN is considering to make a prediction. The most popular technique for CAM obtaining is Grad-CAM (see Figure 2.6) introduced in [30] and based in the gradients of any target flowing into the last convolutional layer.

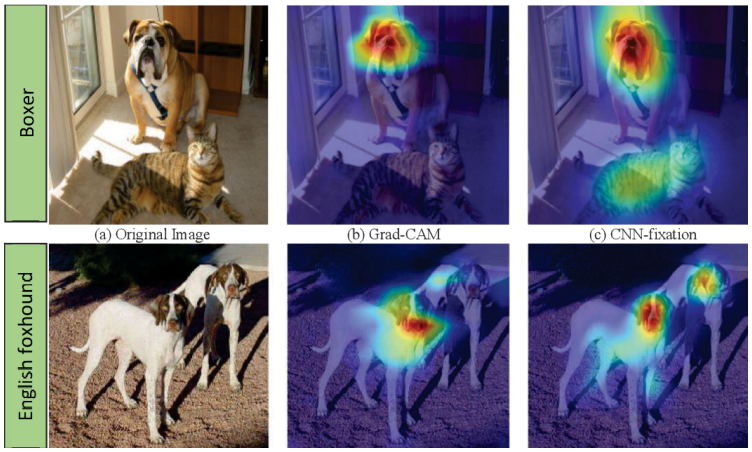

Figure 2.6: CAMs examples for the prediction of different dog breeds with (a) Grad-CAM and (b) CNN-Fixation. Obtained from [22].

2.2.2 *Definition of WSL*

One of the most recent fields of research within Deep Learning (DL) is weakly supervised learning. This particular form of DL is based on weak supervision which consists of labeling large amounts of data with noisy, limited, or imprecise ground truths [42]. An example of weak supervision within the field of computer vision could be video prediction. As a video is composed of a sequence of frames, it is less expensive in terms of time labeling just the whole video and not each individual frame.

Multiple Instance Learning (MIL) is a technique commonly used in WSL. In this particular form of deep learning, the training data is ordered in bags that contain a set of instances. A bag is considered positive when one or more of its instances is positive, a negative label is assigned in the opposite case. So, the goal of the trained model is to recognize the most representative instances and assign a bag-level label [7].

In Figure 2.7 can be seen the typical MIL approach where a bag contains instances of different classes and a decision boundary at the instance level has to be found to classify the global bags. For example, in the histology field a WSI (bag) labeled as cancerous will contain at least one patch (instance) which features correspond to cancerous tissue.

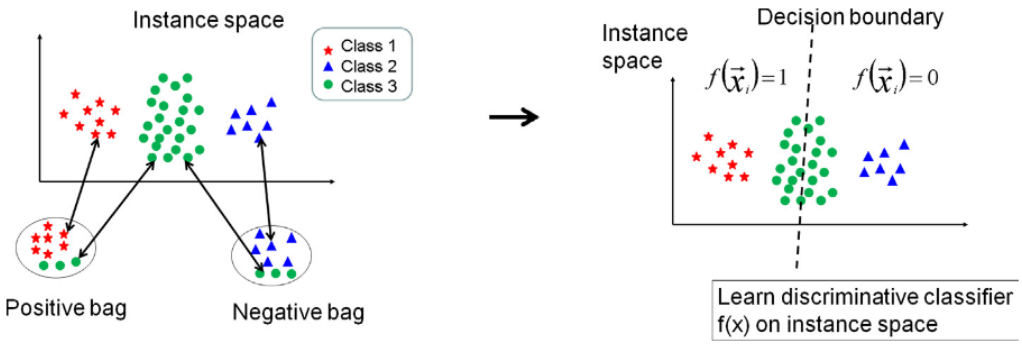

Figure 2.7: Multiple Instance Learning (MIL) framework. Obtained from [1].

Usually, MIL approaches are divided into two steps. The first one obtains the feature vectors of a single bag's instances. The backbone of a CNN extracts these 1D vectors. The second one aggregates the feature vectors to get the bag-level feature vector representing the whole bag.

In Figure 2.8, different aggregation techniques for obtaining the bag-level representation vector are presented. The most simple ones are batch global average pooling (BGAP) and batch global max pooling (BGMP), which consider the mean of the instance features or the maximum among them. MIL attention pooling introduces a trainable aggregation function that gives each instance a different weight to find the most significant ones.

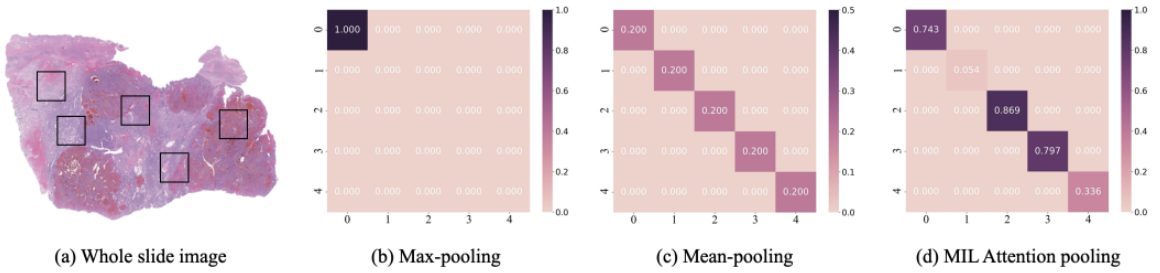

Figure 2.8: Difference between (b) max pooling, (c) mean pooling and (d) MIL attention pooling of instance patches from (a) the WSI. Obtained from [32].

2.2.3 Weakly Supervised Learning in Histopathology

Weakly supervised learning settings have been widely applied in the histopathology image analysis field. Specifically, multiple instance learning frameworks are suitable due to the nature of the data. As WSI are giga-pixel images, they can not be passed through CNNs at their original size due to computation restrictions. As image downsampling would cause a drop in image resolution, approaches have focused on cropping the WSI in multiple patches. In this way, a MIL framework can be simulated considering WSIs as the bags and patches or tiles as the instance.

Here, we present some novel works in this field with aim of comparing the results of our proposed methodology with other in the state-of-the-art. The work in [13] introduced the attention-based MIL pooling (ABMIL). It consists of the first trainable MIL pooling operator and was designed to be adaptive to the data in a certain task. It uses an attention mechanism to obtain the weighted average of instances where the weights must sum 1 to be invariant to bag size. It is thought to be explainable because instances with a higher weight should contain the relevant features for the classification.

In DSMIL [15], authors proposed a dual-stream MIL framework combined with a self-supervised contrastive loss. The proposed framework in this work performs feature concatenation by obtaining the instance features and different image augmentations. However, the feature extraction here is performed but a CNN that has been trained with a contrastive loss. Specifically, this loss allows us to learn robust representation without needing manual pixel-level labels. The method proposed in [17] is called CLAM-SB and combines clustering attention in a MIL framework. The attention mechanism is used to identify tiles that contain high diagnostic values to classify the WSI. They also exploit the instance-level features by performing a constrained clustering over them to refine the feature space. Finally, the algorithm MIL-RNN performs the feature extraction of the instance features with a CNN as several works do. However, its innovation consists in training a recurrent neural network (RNN) to perform the aggregation of this feature to obtain the bag-level aggregated vector [2].

Chapter 3

Materials

Contents

3.1 Database of Whole Slide Images

The present section starts with the database description focusing on the acquisition protocol of the whole-slide images and its pixel-level annotations. It also includes the software and hardware that has been used in the present project.

3.1.1 Description of the database

The database of histological images used in this project was obtained from a collaboration of multiple medical centers around the world that carried out an international multicenter real-life prospective study about endoscopy and histology in Ulcerative Colitis (UC) [11]. During the corresponding endoscopic procedure, more than 600 digitized biopsies were obtained from patients with an ulcerative colitis diagnosis.

These slides were first used to validate the PHRI, a novel score for histological remission evaluation in UC. The clinical protocol consisted of the extraction of two biopsies from each patient. They were obtained from two different sections of the colon, one from the sigmoid and another from the rectum, during the endoscopic procedure. However, biopsies from a single patient are analyzed independently because active inflammation could not be present in both of them.

After the surgical removal of the colon tissue, samples must go through histologic procedures to be observed under the microscope by the pathologists. These procedures include the fixation of the tissue to preserve the cell morphology, the tissue embedding to gain consistency, its sectioning with the microtome in slim slices, and the dyeing of these slices, commonly with hematoxylin and eosin (HE) stain. Afterward, the HE tissue samples are digitalized in Whole Slide Images (WSI) using specialized scanning machines that have emerged with the advances in digital pathology. An abstract of the image acquisition protocol is presented in Figure 3.1.

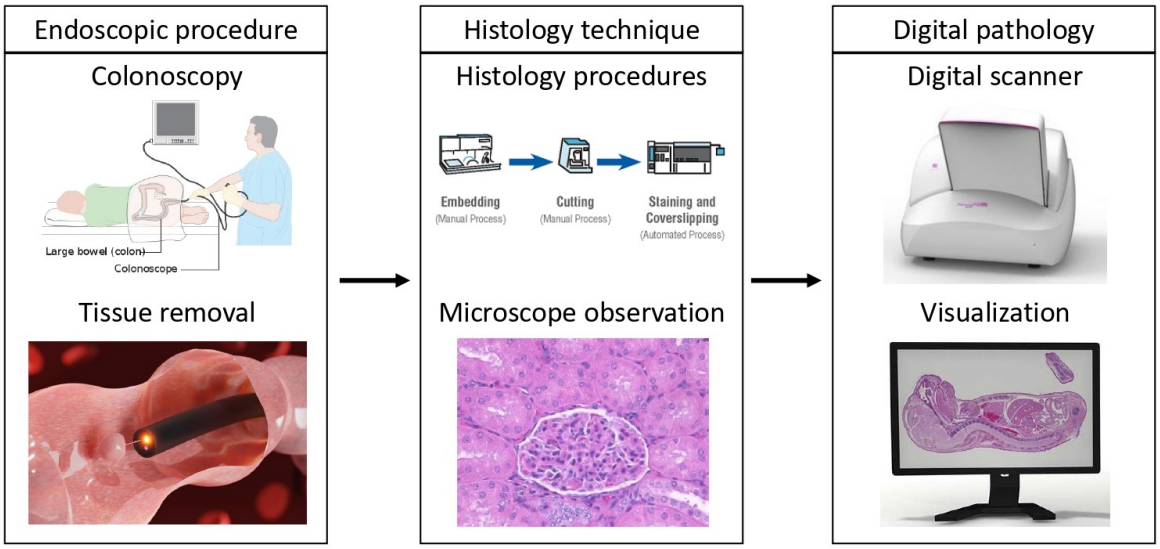

Figure 3.1: Description of the database acquisition protocol. Own elaboration.

Whole Slide Images (WSI) are high-resolution giga-pixel images of digitized tissue samples that suppose an additional challenge in deep learning systems due to memory and computation limitations. Although all the biopsies followed a similar process of extraction and digitalization, the fact of disposing of a multi-center database also allows us to validate the model across images from different hospitals that could present variability in terms of the staining colors or the biopsy orientation.

As noted previously in Table 2.1, PHRI is a histological remission index for ulcerative colitis disease assessment. It focuses on detecting neutrophils in four different regions of the tissue sections: lamina propria, cryptal epithelium, cryptal lumen, and surface epithelium. To score a single biopsy, the pathologist examines carefully these compartments and obtains the global score of the biopsy by adding the number of regions with significant neutrophil presence.

In Figure 3.2, a large WSI with a high level of UC activity is presented. The scoring for this biopsy is PHRI equal to 4. This means that there is the presence of neutrophils in each of the four structures of interest in the biopsy. In this figure, the right column includes four different regions of the WSI where the neutrophils are present and indicated with a black mark. Note that neutrophils are very small and considerably similar to other surrounding cells in the tissue. In descending order, (a) lamina propria, (b) surface epithelium, (c) cryptal epithelium, and (d) cryptal lumen are presented to provide context from the different regions of interest. All WSI are annotated with PHRI at the image-level score that will be used as ground truth (GT) for model training and evaluation.

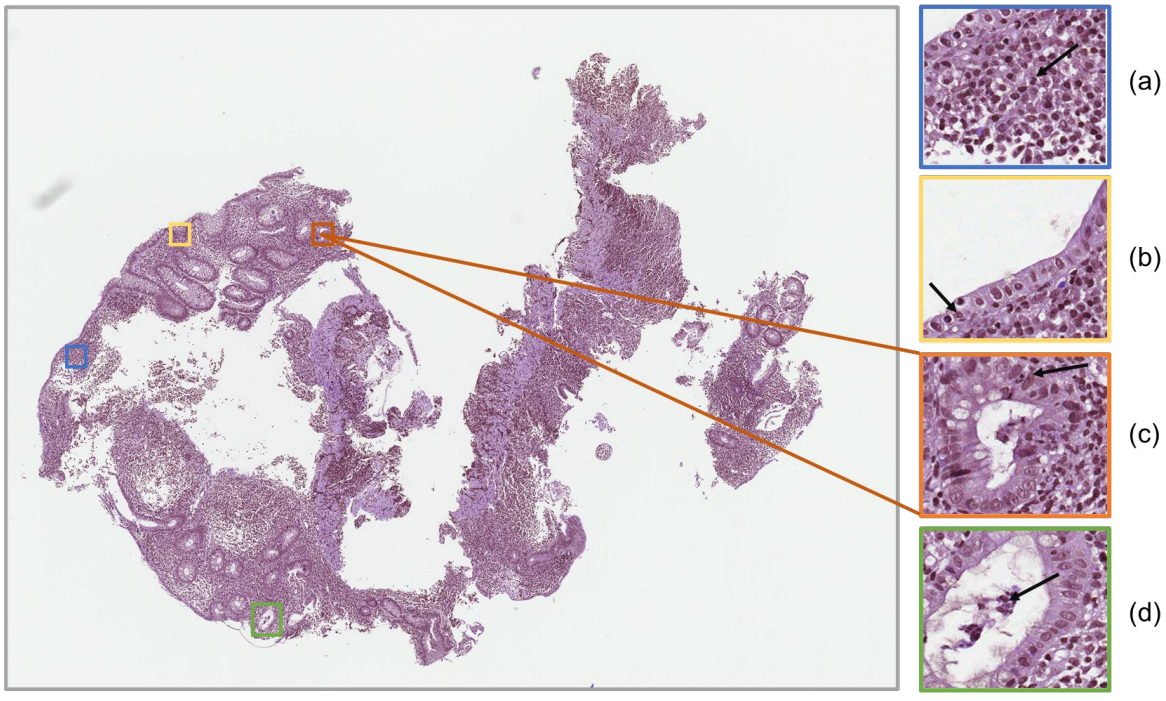

Figure 3.2: A Whole Slide Image of a biopsy with ulcerative colitis activity and its four structures of interest. The right column patches correspond to (a) lamina propria, (b) surface epithelium, (c) cryptal epithelium and (d) cryptal lumen. Own elaboration.

3.1.2 WSI annotations

Detailed image annotations play a crucial role in the development of supervised deep learning algorithms, but often getting them in large datasets is time-consuming or even unapproachable. Specifically in the medical image field where the data availability is limited and in digital histopathology where the image complexity is huge.

For this purpose, the pathologists associated with the PICaSSO project were asked to perform pixel-level annotation only in 120 WSI of the complete database using an in-house software called *MicroDraw*. This software was specifically adapted by CVBLab engineers to perform WSI annotations and offers plenty of flexibility to the annotator.

These annotations consisted in indicating with different colors the four compartments of interest in the biopsy. At the same time, pathologist also annotate in the more precise way the position of the neutrophils in case the biopsy present histological inflammation. WSI in histological remission does not contain this information due to the absence of neutrophils.

The aim of requesting these concrete annotations was twofold. On the one hand, information on localized tissue allowed us to perform the whole slide image cropping in a more specific way saving only the patches that got more than 20% of annotation and tissue. This will play a crucial role as it will be explained in the very next section because it reduces the computational cost of training the model. On the other hand, knowing the neutrophil localization will allow us to impose constraints on the network in a novel way as will be explained in Section 4. Finally, some annotation examples are presented in Figure 3.3.

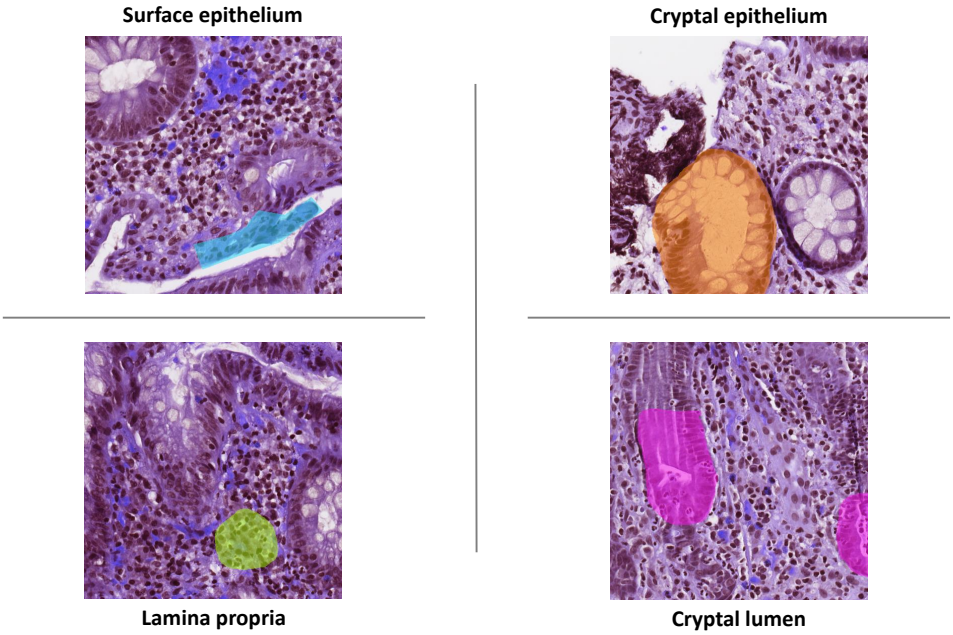

Figure 3.3: Annotations of neutrophils in each of the four compartments of interest. Own elaboration.

3.2 Software

The software used for the implementation of the deep learning algorithm was the programming environment of Python 3.7. Python is a high-level, general-purpose, interpreted, interactive and object-oriented programming language. It can incorporate modules and packages that can be used for solving specific tasks. The most used package in the present project is Keras. This module is an open-code library designed for the high-level implementation of deep learning algorithms and it contains the functionalities such as layers, activation functions, and optimizers to implement convolutional neural networks.

As the training phase of the deep learning process is a high computational cost process, it has been launched in a remote server with high computational performance. For the connection with this server, the MobaXterm software has been used because it allows the creation of an SSH (Secure SHell) connection. This communication protocol facilitates the connection between the working computer and the server which runs the training of the deep models.

Another key software employed is MATLAB, an application developed by *Mathoworks*. This program includes a multi-paradigm programming language and numeric computing environment that can be used to solve scientific and engineering problems. In this project, MATLAB has been employed to carry out the processing of the images and the statistical analysis of the results.

3.3 Hardware

The main core of the project has been carried out in a computer composed of an Intel Core i7-7700HQ processor with 64-bit Windows 10 as an operating system. It also contains an 8GB RAM memory, an external storage hard drive disk (HHD) of 1TB, and an NVIDIA GTX1050-4GB graphics card. The deep model training has been carried out in an NVIDIA DGX A100 (now on, DGX) system owned by the CVBLab. The DGX is one of the most powerful hardware resources for artificial intelligence development allowing unprecedented calculus density and performance for deep training and interference. It contains up to 8 GPU NVIDIA A100 with 640GB GPU memory.

As WSI are files with a huge size ranging from 500MB to 2GB, it has also been a NAS Synology DS918 server with a storing capacity of 32TB. The workflow of the project consists in storing the data and the generated code with the own computer in the NAS server and establishing an SSH connection between the NAS and the DGX where the different scripts will run to perform training and inference with the deep models.

Chapter 4

Methodology

Contents

With the aim of introducing the section in an ordered manner, a flow chart including all the methodology is presented in Figure 4.1. It includes all the methodology processes involved in the development of the algorithms for the ulcerative colitis prediction in whole-slide images. More details on the of the methodology are presented following the flowchart.

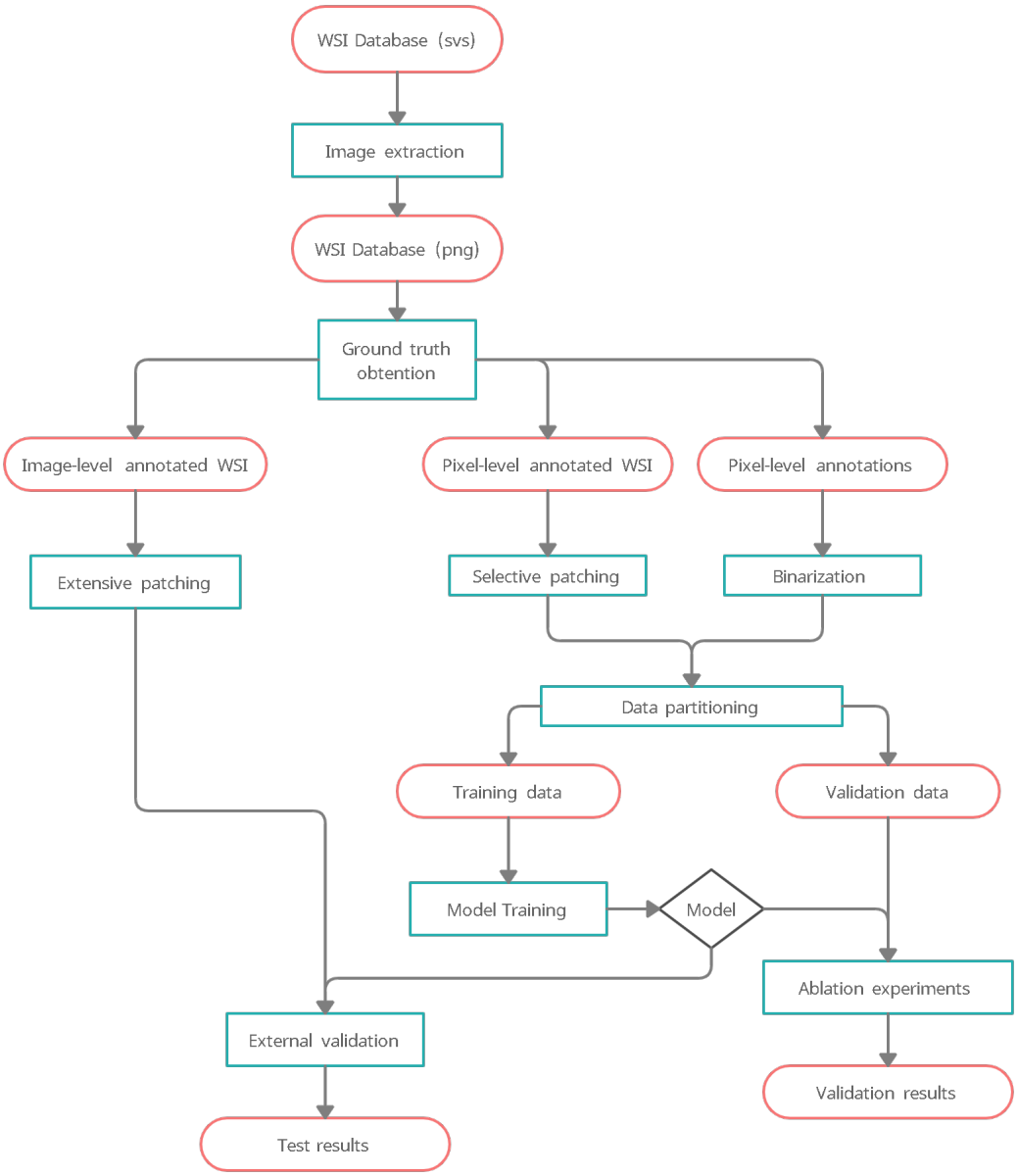

Figure 4.1: Flow chart of the proposed methodology. Own elaboration.

As presented in the previous section, the database for project development consists in a large dataset containing whole-slide images. WSI are giga-pixel images obtained after the digitalization of biopsies in specialized scanners. WSI are contained in .svs files that allow to visualize them in specific software such as Amperio Imagescope. This file must be extracted to .png files that allow the visualization of the images by standard software. Some of the files are attached with a pixel-level annotation of the WSI. This annotation make by the pathologist determines which images are going to be used for training and validating the models. WSI with annotations will be patched in the preprocessing stage in a more selective way saving only the patches with a certain area of annotation. Once the preprocessing stage is completed, we proceed to perform the data partition of the annotated images between training and validation data. This partition is performed following a 70/30 rate.

Afterward, the model development includes all the programming stages for the creation of the deep-learning based algorithm for the prediction of UC in histologic images. This algorithms based on Weakly Supervised Learning are trained using the selected patches of biopsy images and its annotations. Once the model is obtained, the validation is performed using also the pixel-level annotation to evaluate the efectiveness of this information. An extensive series of ablation experiments are carried out to determine the optimal configuration of the framework. Different experiments are evaluated and compared using the traditional metrics in classifications problem to obtain the figures of merit.

Finally, we aimed to evaluate the performance and robustness of our model in an extensive database. For that purpose, we used a set of WSI without pixel-level annotation to perform the external validation of the algorithm. These slides are not patched selectively using the annotation, so they will be composed of a larger number of patches after the preprocessing stage because the entire WSI is included. Note that annotations are not necessary to predict the test images. The test results are evaluated in terms of multi-centre prediction performance and neutrophil detection efficiency.

4.1 WSI preprocessing

Whole-slide images (WSI) are digital biopsies of tissue sections. WSI can include some redundant information because different slices obtained in the microtome are usually placed in the same crystal that is going to be digitalized in the specialized scanners. As multiple instance learning in histopathology is constrained in terms of computational cost, it makes sense to select one of the slices to be processed by the deep model which will help to reduce the size of the images and the number of patches.

However, other preprocessing steps such as image cropping and background elimination are necessary to reduce the computational cost and the noise of the data. All of the WSIs in the database are preprocessed following the next steps. First, the WSI is downsampled to 20x resolution. The magnification of the downsampling process conditions the final size of the image and the portion of the region that will be observed in each patch. Afterward, the downsampled WSI is cropped in patches of 512x512 pixel size with an overlap rate equal to 50%, which means that adjacent patches share half of the information.

In WSIs, tissue is usually not present all around the image. As background patches do not provide relevant information about the biopsy, it is advisable to discard these patches that do not include tissue. For that purpose, the Otsu threshold (i.e. a method for automatic image thresholding) is applied across the red channel of the RGB image to separate the tissue from the background. With this in mind, the patches with a ratio lower than 20% of the area corresponding to tissue will not be considered and will not pass through the deep model.

The WSI preprocessing stage is presented as a graphical abstract in Figure 4.2. Here the grid equal-size grid represents the patching process of the WSI. At the same time, the biopsy tissue is overlayed with a purple color that indicates the presence of tissue and separates it from the background. In this project, color normalization was not considered because one of the aims of the project was to prove the model's robustness against images coming from different centers as data comes from an international real-time study.

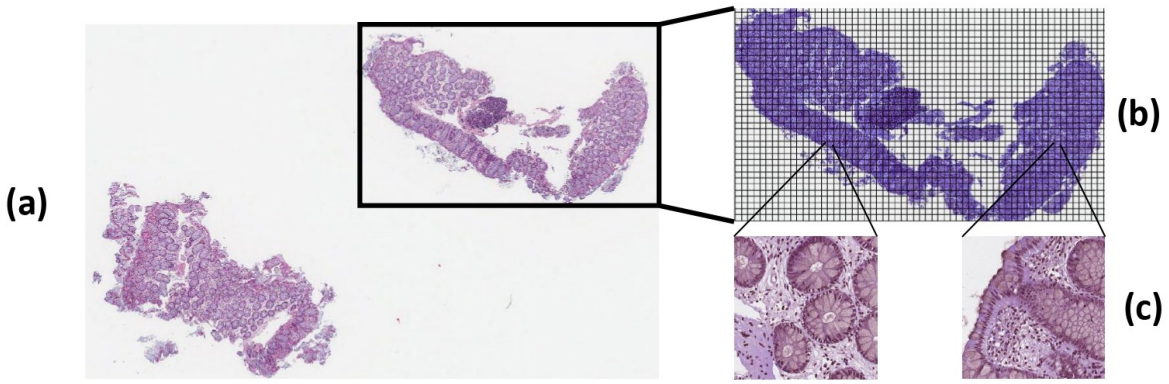

Figure 4.2: Graphical abstract of WSI preprocessing. (a) WSI, (b) cropped WSI and (c) obtained patches. Own elaboration.

4.2 Data partitioning

As noted before, the acquisition protocol consisted in the extraction of two biopsies for each patient. With this in mind, patient-level partitions were carried out to create the three canonical datasets (training, validation and test) in machine learning. 'Patient-level partitions' means that both WSIs from a single patient are contained in the same set, although them are analyzed independently. It is absolutely necessary to consider this partition technique because having images of a single patient in different sets could not represent the real capabilities of the model due to the presence of repeated patterns. This problem does not arise in a reduced number of patients that only contain one biopsy due to lack of availability.

Partitions in three subsets were determined by the availability of the pixel-level annotation of certain WSI. Only WSI with this data associated were used for training and validating the model. As a reminder, these images include detailed information of the neutrophils location. The annotations also indicate the four regions of interest and include regions without neutrophils infiltration for a balanced contribution. This set of images contains a lower number of patches due to only patches with at least 20% of annotation are considered. This fact considerably reduces the computational cost of the model training without limiting the external validation.

Only 130 of 529 WSIs (24.57%) have been used for model development. The remaining 100 histological images constituted the test set for the external validation of the model. Among the relevant features of this set, we find that it includes WSIs from centres that are not present in the training set. This fact allows us to test the model robustness across a multicentre dataset where image properties such as color staining could vary. It is also important to consider that this set of images does not have annotation, so the entire biopsy is patched. Because of this fact, the number of patches per WSI in this set will be bigger in comparison with the training and validation sets.

In Table 4.1 it is presented that the number of patches is considerably larger in the test set for external evaluation in comparison with the training and validation set. In this same table, the number of patches is indicated with the mean and the standard deviation of patches. The third row indicates the percentage of WSI in histological remission (HR) in each of the sets.

Table 4.1: Database description. Amount of whole-slide images in each set (first row) indicating the percentage between the WSI with pixel-level annotations , number of patches (third row) and percentage of slides in HR (fourth row).

	Training	**Validation**	**Test**	**Total**
WSI	84 (64.6%)	46 (35.4%)	399	**529**
patches	61.1 ± 54.2	58.2 ± 36.4	450.2 ± 230.8	**354.3 ± 185.8**
WSI in HR	41 (49.1%)	28 (60.87%)	244 (61.15%)	**313 (59.17%)**

4.3 LCMIL: Location Constraints Multiple Instance Learning

The proposed method consists in an end-to-end multiple instance learning algorithm to perform whole-slide image (WSI) classification. The novel MIL formulation, which includes attention constraints, allows to detect neutrophils in WSI from patients with a ulcerative colitis diagnosis and classify them in histological remission or active inflammation according to PHRI, a neutrophil-based scoring system for histological evaluation in UC. The overall framework is presented in Figure 4.3 and it will be deeply in the following sections of the document.

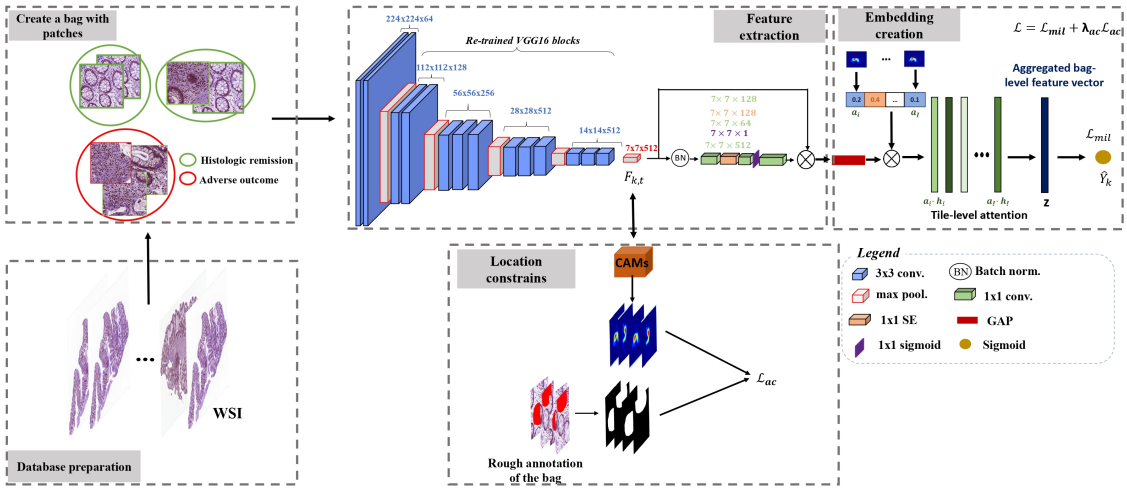

Figure 4.3: Graphical abstract of the LCMIL framework for WSI prediction. Own elaboration.

4.3.1 Problem formulation

In MIL tasks, the training dataset is composed of bags, where each bag contains a set of instances. However, MIL models aim to predict the image-level label, not the patch-level classification. In the histopathology image field, a bag corresponds to one WSI and the instances are each of its patches. A positive label is assigned to a bag if it has at least one positive instance. That means that if a WSI includes a region that presents characteristic features from the positive class (i.e. the presence of neutrophils in our problem), this label will be assigned even if other regions includes tissue for the negative class.

The dataset formulation in MIL tasks is denoted by $\mathcal{S} = (X_k, Y_k)$ with $k = \{1, 2, 3, \ldots, N\}$, where X_k denotes the k-th input bag and $Y_k \in [0, 1]$ refers to the bag-level label assigned to the k-th input bag. As mentioned before, $Y_k = 0$ refers to HR and $Y_k = 1$ refers to UC activity. Our training set is expanded with the pixel-level annotation that indicate the location of the neutrophils being finally denoted as $\mathcal{S} = (X_k, Y_k, A_k)$ with $k = \{1, 2, 3, \ldots, N\}$, where A_k is the rough estimation of neutrophils located in image X_k. Being $a(i, j)_{k,t}$ the pixel (i, j) in the t-th patch from the bag k-th, $a(i, j)_{k,t} = 1$ if it corresponds to a pixel that is located around a neutrophil, whereas $a(i, j)_{k,t} = 0$, otherwise. A visual representation of the problem formulation is presented in Figure 4.4.

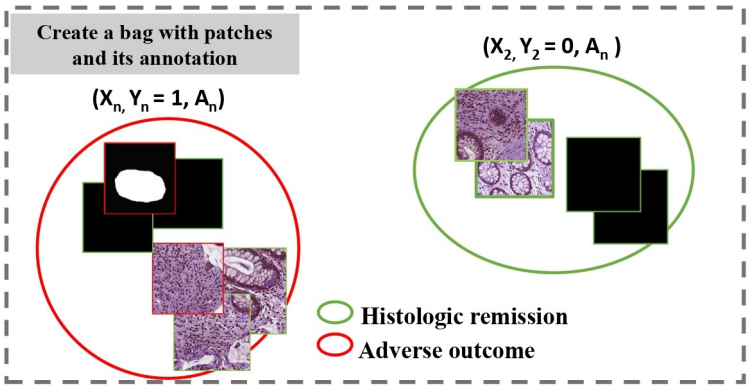

Figure 4.4: Graphical representation of the MIL formulation in our problem. Own elaboration.

4.3.2 CNN backbone with refinement module

To solve the present problem, the proposed CNN backbone to perform feature extraction is the VGG16 with the SeaNet module. This CNN architecture was chosen because it demonstrated to perform better than other widely used network architectures for histological image classification [5]. VGG networks were firstly introduced in [34] for the ImageNet classification problem where authors investigated the effect of the convolutional network depth in its performance. The SeaNet module consists basically in "Squeeze-and-Excitation" (SE) block that recalibrates features responses by modelling the dependencies across channels [10].

As denoted in Figure 4.5, the input image size for VGG16 is 224x224x3 pixels. Each image is passed through a stack of 3x3 convolutional layers and max-pooling layers that perform spatial pooling over a 2x2 pixel window and a stride set to 2. The network has a depth of four convolutional blocks in which the number of channels is duplicated in each of the first three blocks from 64 to 512 channels.

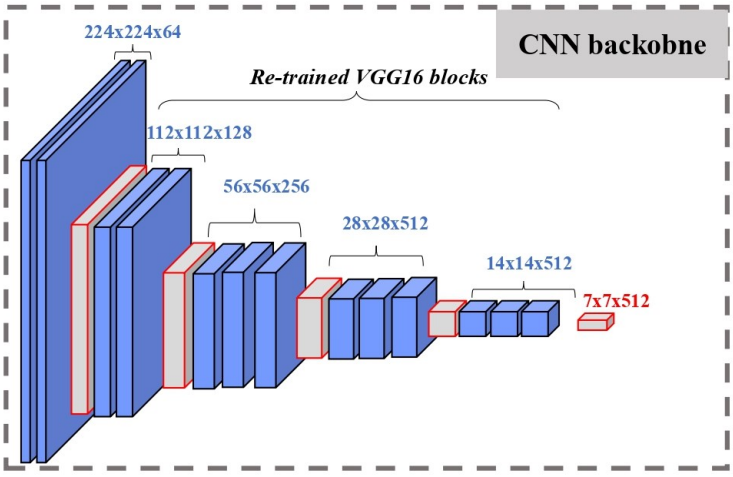

Figure 4.5: Detailed configuration of CNN backbone following the VGG16 top-model architecture. Note that colors and forms follow the same legend than in Figure 4.3. Own elaboration.

Note that only the top-model from the VGG16 due to the incorporation of a refinement module in the feature extraction stage that will be detailed later. The weights of the VGG16 are initialized in the ImageNet dataset, so transfer learning is performed by freezing the first convolutional block to facilitate convergence and improve model performance.

For each separated histological patch, a 7x7x512 embedded map is provided by the VGG16 feature extractor that will be passed through a feature refinement module. The aim of this attention module is to encourage the network to focus on the most discriminative features that will help to improve the classification performance of the model.

The 'Squeeze and Excitation' architecture (SE) proposed in [10] consists of 1x1 convolutions with a decreasing and an increasing number of channels to output an embedded map with the same size that the input one. After a global average pooling (GAP) operation in the input maps, a sigmoid activation and the activation of each patch are recalibrated to obtain refined feature maps that focus on the most discriminative features of each image. Its architecture is presented in Figure 4.6.

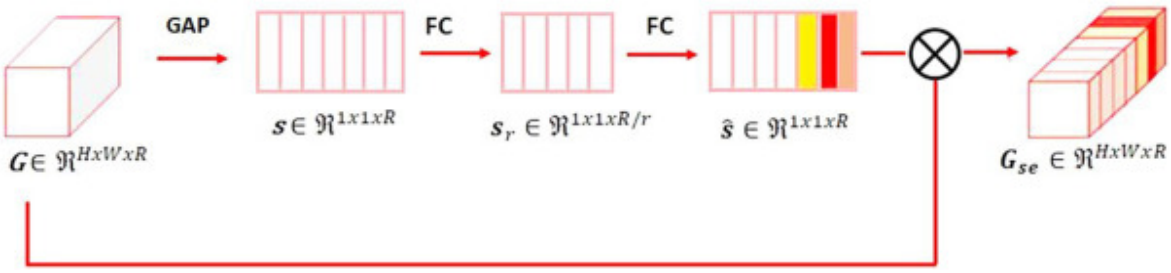

Figure 4.6: Detailed configuration of the SeaNet refinement module. GAP refers to Global Average Pooling, FC refers to fully connected. Obtained from [5].

Finally, a projection head module is included to map the refined feature maps to an embedding vector with a length equal to 512. This vector would allow us to perform the image classification at the single instance level. As explained in [5], global average pooling (GAP) was chosen to obtain the embedded vector while reducing the complexity of the model.

4.3.3 Attention Constraints

Deep learning algorithms are commonly data-hungry, that means they need large amounts of labeled data to obtain acceptable results. For that purpose, imposing constraints to the output of neural networks pretends to improve their performance while reducing the need of labeled data. The most popular form of imposing constraints is to introduce new terms to the loss function. One crucial point to consider in constrained optimization is the weighted importance of each term to the loss functions that will determine the the network convergence [20]. For example, soft-constraints have been widely explored in incremental learning settings. In [16] an additional term, known as knowledge distillation loss, is added to the cross-entropy loss function with the aim of reducing catastrophic forgetting when retraining models with unseen classes.

Due to the nature of the PHRI score, neutrophils identification is crucial for UC activity prediction in WSI. For that purpose, the first approach to solve this problem consisted in designing a combination of two models that were trained separately to perform WSI classification. The first model was composed of the same CNN backbone for feature extraction with the aim of performing neutrophil classification at patch-level. The second model followed a multiple instance learning paradigm by combining the patch-level features extracted by the previous model to obtain an image-level prediction of the WSI. [8]

As training two independent models is not optimal, we included an attention constraint (AC) module to enforce the identification of neutrophils in single patches in an end-to-end MIL framework for the classification of WSI. The AC modules is implement as a neutrophil-based area constraint that restricts the expansion of positive samples during training. This module helps the network not only to detect neutrophils but also to ignore other similar cells that could also be present in the tissue samples. Its graphical representation is showed in Figure 4.7 and detailed as follows.

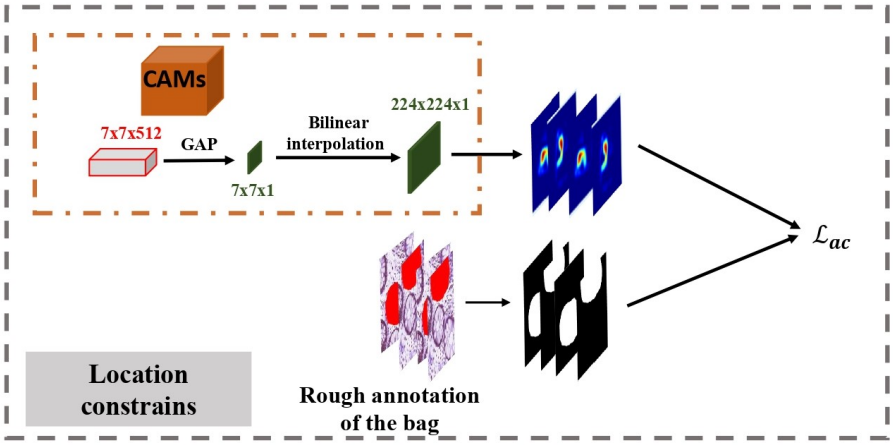

Figure 4.7: Detailed configuration of the attention constraints module. Own elaboration.

The AC module is constructed as follows. A Global-aggregation layer through the channels of a single instance is implemented to obtain an activation map representing the distribution of the features extracted from each of the instances belonging to a given bag. The aim of this layer is to summarize the information from all spatial locations in the feature-embedded map $F_{k,t} \in \mathbb{R}^{H \times W \times C}$ obtained by the feature extractor backbone before the feature refinement submodule. It obtains one representative map $\rho \in \mathbb{R}^{H \times W}$ of each instance, in similar way than Class Activation Maps (CAMs) are obtained for CNN interpretability. Therefore, $\rho \in \mathbb{R}^{H \times W}$ is defined as follows:

$$\rho(i,j)_{k,t} = \frac{1}{C} \sum_{c \in C} F_{k,t}(i,j,c) \tag{4.1}$$

Note that $H \times W$ are the dimensions of the embedded map and C is the number of filters, which do not correspond with the size of the annotations A_k (224^2). Therefore, a bilinear interpolation is performed to the activation maps ρ to make them comparable to pixel-level

annotation that has been previously binarized to only indicate the presence of neutrophils. Then, the resized activation maps ρ are transformed into $\rho_s = \phi(\rho)$, where ϕ is the sigmoid activation function. The aim of the sigmoid activation function is to range the map activation function into [0-1] that allows to define an area constraint L_lc with the L_2 (also called mean square error) penalty measured between the binary annotations and ρ_s.

$$\mathcal{L}_{lc} = \sum_{k,t} I(Y_k = 1 \; and \; a(ij)_{k,t} > 0) \; ((a_{k,t} - \phi(\rho_{k,t}))^2) \tag{4.2}$$

4.3.4 MIL attention embedding

The aim of the MIL attention embedding is to obtain a representation from the whole bag by aggregating the embedded feature vectors extracted by the CNN backbone. To take advantage from the attention constraints module, we propose a novel weighted average of the instance to obtain the final feature vector.

For that purpose, we obtain a single score from each pseudo-CAM by aggregating its amount of information that is used as an instance weight for MIL embedding. This value is thought to be higher when the AC module performs neutrophil detection in a single instance, so the proposed embedding will consider a weighted aggregation of the features with a higher contribution of this instance.

As the aggregation method has to be invariant to the number of patches, weights are normalized to sum 1 with softmax operation. Note that these wights are updated every epoch beacuse the overall framework is trained end-to-end. A visual abstract of the proposed MIL embedding is presented in Figure 4.8.

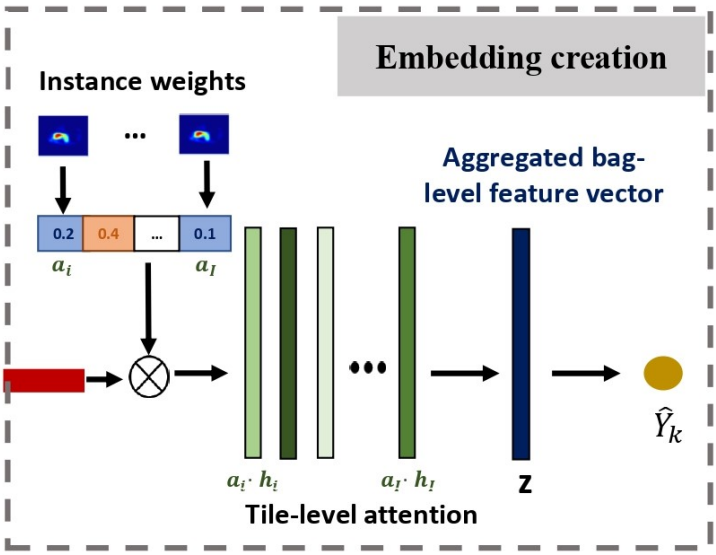

Figure 4.8: Detailed configuration of the MIL embedding. Own elaboration.

Therefore, the embedded feature vector per bag is obtained as $Z_k = \sum_{t \in I_n} a_t \cdot \mathbf{h}_t$ being a_t the wieght of each instance and h_t its 1D feature vector. So, a_t is defined as:

$$a_t = \frac{exp\{\sum \rho(i,j)/S\}}{\sum_{I_n} exp\{\sum \rho(i,j)/S\}} \tag{4.3}$$

where $S = H \cdot W$.

The proposed attention embedding increases the variability between instances of a positive bag which helps the model to recognize positive instance that contains neutrophils. If there is no activation corresponding to neutrophils in the map ($\rho_{k,t}$), the value of a_t will be low and therefore, the embedding features $\mathbf{h}_t$ corresponding to this instance will have a smaller weight in the final prediction. In the case of negatives bags, the attention weights values should be similar over all the patches in the bag so they will contribute equally.

The loss function used to optimize the end-to-end MIL approach is the cross-entropy cost function:

$$\mathcal{L}_{mil} = \sum_k (I(Y_k = 1) log\hat{Y_k} + I(Y_k = 0) log(1 - \hat{Y_k}) \tag{4.4}$$

where $I(\centerdot)$ is an indicator function.

Finally, the global loss of the framework is composed by the MIL loss and a weighted contribution of the attention constraints loss:

$$\mathcal{L} = L_{mil} + \lambda_{ac} L_{ac} \tag{4.5}$$

where λ_{ac} is the weight of the constrained loss.

Chapter 5

Experiments and Results

Contents

5.1 Evaluation metrics

For the comparison of the different methods and results that will be presented in this chapter, we selected the performance metrics widely used in classification problems. Once we obtain the model prediction and dispose of the ground truth (GT) for each whole-slide images, the four possible combinations of results are presented in the confusion matrix of Figure 5.1.

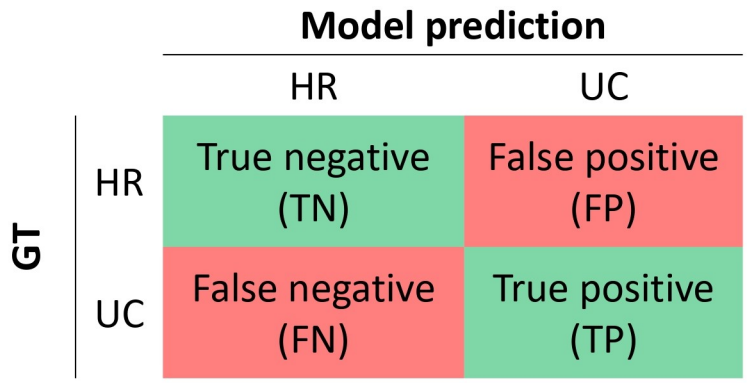

Figure 5.1: Confusion matrix for the proposed problem. As a reminder, UC refers to ulcerative colitis activity and HR to histological remission. Own elaboration.

The seven metrics for evaluating the classification performance of the different models are:

- **Sensitivity (SN)** or **recall** measures the probability of the model to identify positive cases. It is calculated as the rate of positives (P) cases (in this case, WSI with UC activity) correctly classified by the model:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.1}$$

- **Specificity (SPC)** measures the probability of the model to identify negative cases. It is calculated as the rate of negative (N) cases (in this case, WSI in HR) correctly classified:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{5.2}$$

- **Positive predictive value (PPV)** or **precision** measures the probability of a WSI with a positive prediction to be in UC activity. It is calculated as the rate of positive predictions (T) correctly classified:

$$PPV = \frac{\text{TP}}{TP + FP} \tag{5.3}$$

- **Negative Predictive Value (NPV)** measures the probability of a WSI with a negative prediction to be in HR. It is calculated as the rate of negative prediction (N) correctly classified:

$$NPV = \frac{\text{FP}}{FP + FN} \tag{5.4}$$

- **Accuracy (ACC)** measures the ability of the model to return a correct prediction of the WSI. It is calculated as the rate of accurate predictions in the whole set:

$$ACC = \frac{\text{TP} + \text{TP}}{TP + TN + FP + FN} \tag{5.5}$$

- **F1 Score (F1S)** is a measure of the test's accuracy calculated as the harmonic mean of the precision (PPV) and recall.

$$F1S = \frac{2 * \text{PPV} * \text{SN}}{\text{PPV} + \text{SN}} \tag{5.6}$$

- **Area under the ROC curve (AUC)**: It measures the area under the Reciever Operating Characteristic (ROC) curve and it represents the diagnostic ability of a binary classifier. AUC values close to 1 mean that the model has a high probability of returning correct predictions while AUC values near 0.5 mean that it makes random predictions.

5.2 Ablation experiments

In the deep learning context, an 'ablation experiment' refers to a procedure where certain components of the framework configuration are removed or modified. This changes affect the network performance allowing us to understand properly its behaviour and optimize the framework.

In this section, we present a comprehensive evaluation of several components of the proposed Location Constrained MIL (LCMIL) to support the final implementation of the experimental setting. The different experiments are carried out in the validation cohort that includes pixel-level annotations as has been presented in Chapter 3.

This is a crucial point in our work because these annotations are needed for the optimization of the novel loss of attention constraints. However, in the test set, they are not required because the desired output of the network is the final MIL classification of a single bag.

5.2.1 *Weight of location constrain loss*

As mentioned earlier, the optimization of the weight of auxiliary loss functions is crucial to achieve convergence in soft constrained problems. A balance should be found between lower weights that do not contribute considerably with model training and higher values that can disturb networking optimization.

Here we optimize both the weight importance and the cost function for the novel attention constrained loss added to the MIL framework. For that purpose, we evaluate the L1 and L2 (see Equation 5.7) norms, also called absolute-value or Euclidean norm respectively. Both are regression losses that in this problem calculates the difference between the pseudo-CAM obtained by the location constrained module and the pixel-level annotation indicating the presence of neutrophils.

$$L2 = \sum_{i=1}^{D} (x_i - y_i)^2 \tag{5.7}$$

where x_i is the pseudo-CAM , y_i the annotation of the patch and D its number of pixels.

After an initial grid search, the values explored in the different experiments for the weight of the constrained loss are $\lambda_{ac} = \{0.1, 0.1, 1, 1, 5\}$. Note that the contribution of the MIL cross-entropy loss is not modified across the different experiments.

The results of the ablation experiments in terms of accuracy are presented in Figure 5.2. They show a slightly improvement of the L2 norm in comparison with L1 across all the explored weights. With that in mind, we selected $\lambda_{ac} = 1$ which led to a performance improvement around 5 a 10%. These results also show that the importance of performing ablation experiments in soft-constrained optimization where the algorithms performance is remarkably modified when the weighted contribution changes.

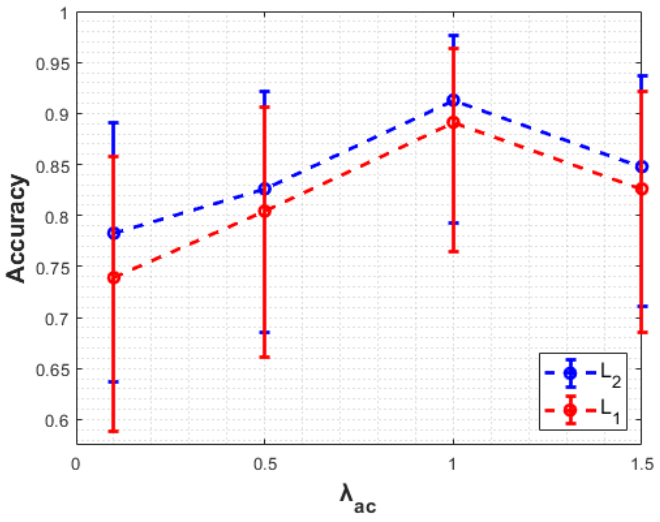

Figure 5.2: Ablation studies on the hyperparameters for λ_{ac} are performed for bag-level accuracy on validation set. Confidence intervals are shown at 95%.

5.2.2 MIL aggregation ablation

As it has been mentioned in Chapter 4, different approaches have been used to perform instance level aggregation in MIL problems. Here we aim to compare the proposed methodology highlighted by the weights coming from the attention constraints with other aggregations methods proposed in the literature and previously introduced in Figure 2.8.

The ablation experiment consisted on modifying the projection head module that performs batch global aggregations to obtain the aggregated bag-level feature vector. Experiments are compared with the figures of merit in the validation set that have been presented in Figure 5.1. Our model (LCMIL) clearly outperforms the other methodologies by a margin of accuracy around 2 and 7%.

Table 5.1: Comparison of the different aggregation methods on the validation set.

	BGAP	**BGMP**	**Attention [13]**	**LCMIL**
SN	0.9643	0.9643	0.8889	**0.9643**
SPC	0.6667	0.7778	0.7778	**0.8333**
PPV	0.8182	0.8710	0.8571	**0.9000**
NPV	0.9231	0.9333	0.8235	**0.9375**
F1S	0.8852	0.9153	0.8727	**0.9310**
ACC	0.8478	0.8913	0.8444	**0.9130**
AUC	0.8155	0.8710	0.8333	**0.8988**

One point to note, is the considerable increase in specificity for the proposed methodology in comparison with the three other aggregations. The upgrade in terms of specifity ranges between 5.5% for BGMP and 17% for BGAP. This translates in a reduced number of WSI in histological remission being predicted as active. This is due to the presence of similar cells to the neutrophils all around the tissue that can be confound with these cell outgrowing the number of positive instance. This fact remarks the usefulness of the attention module and the λ_{ac} loss term of our novel approach to effectively detect neutrophils clearly helping to improve the classification performance.

As remarked in previous ablation experiments, the MIL weight aggregation has proven to be a crucial factor for performance improved, so it is important to understand their distribution. As neutrophils are small cells present only in certain regions of the biopsy, an intuitive reasoning tells us that only a limited number of patches should be consider relevant for determining a positive bag with Ulcerative Colitis activity.

Aggregations functions function such as BGAP do not take advantage of this fact due to all the instance contribute equally to obtain the final embedding. Attention embeddings such as [13] weighted the instance importance, but weights do not differ enough (i.e. are distributed similarly) to fit our problem. So, we obtained the instance weights of our proposed embedding from thee and compare them with [13] weights in Figure 5.3. Remark that the bag comprises around 80 biopsies and only a 15% of them are labeles as positive taking into acount the pixel-level annotations performed by the pathologist.

As observed, MIL attention [13] fails to recognize the most informative instance because the vast majority of weights are distributed around similar values. However, the proposed embedding highlights a reduced number of instances (around 25%) with a higher values while the other stay fairly close to zeros. This facts outlines the capacity of LCMIL of determining the most relevant features for neutrophils localization at the patch level which results highly beneficial to perform bag-level classification.

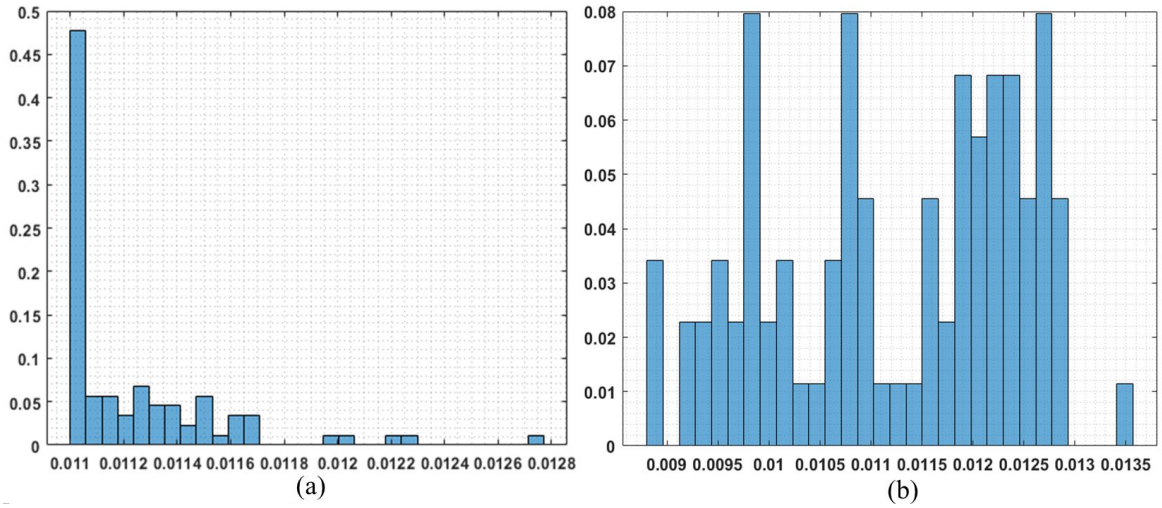

Figure 5.3: Distribution of embedding weights across the instances that comprise a WSI. (a) Proposed attention embeddings. (b) Attention weights proposed in [13].

5.2.3 Network interpretability

Network interpretability is crucial for understanding the behaviour of the model. Moreover, machine learning based systems that work as black boxes (i.e. only outputs the model prediction/function) are thought to not match with clinicians when working as clinical decision support systems (CDSS).

For that purpose, we obtained the Class Activation Maps (CAMs) from the low dimensionality feature maps obtained by the feature extractor with the Grad-Cam methodology proposed in [30]. Some examples for different regions of the biopsy (lamina propria and surface epithelium) in Figure 5.4.

Here we present the pixel-level annotations performed by the pathologist indicating the neutrophils localization and we compare the CAMs obtained by our proposed model based on attention constraints with the MIL Attention [13]. As observed, MIL Attention struggles to recognize the relevant features of the patch while λ_{lc} term focuses in determining cells that correspond with neutrophils.

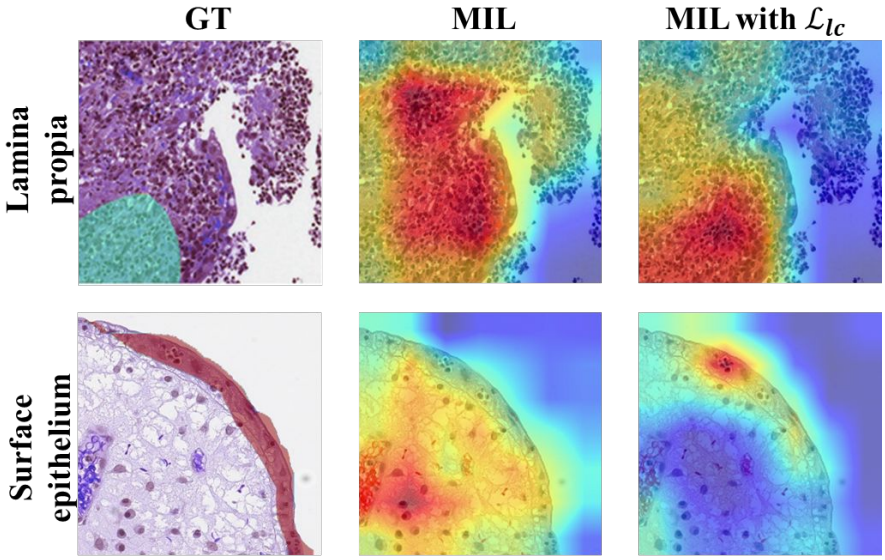

Figure 5.4: Class activation maps (CAMs) of some regions where neuthrophils are found. First column: original images with pathologist annotation; second column: CAMs obtained using the normal MIL Attention model; third column: CAMs using the proposed location constraints.

5.3 Results and discussion

After training and validating the proposed methodology in a subset of pixel-level annotated WSI, we fully exploit these crucial information, as proved in the ablation experiments, while retraining the model with jointly set composed of the training and validation data. Once the final model is obtained, we perform an extensive external validation of the algorithm in a larger set of WSI.

It is important to remember that the WSI in the test set are not annotated so they are patched completely which translates in a increased number of patches as remarked in Table 4.1. Note that for the test cohort only the global bag label are available. The test metrics to evaluate the model performance in subset cohort of 100 WSI are presented in Table 5.2. The comparison with other state of the arts results for multiple instance learning are obtained from [5].

Table 5.2: Comparison of the different MIL techniques in the test cohort.

	ABMIL [13]	**DSMIL** [15]	**CLAM-SB** [17]	**MIL-RNN** [2]	**LCMIL**
SN	0.9583	0.8293	0.9302	0.8667	**0.9583**
SPC	0.6923	0.7288	0.8033	0.7797	**0.9615**
PPV	0.7419	0.6800	0.7692	0.7500	**0.9583**
NPV	0.9473	0.8600	0.9423	0.8846	**0.9615**
F1S	0.8393	0.7473	0.8421	0.8041	**0.9583**
ACC	0.8200	0.7700	0.8558	0.8173	**0.9600**
AUC	0.8253	0.7546	0.8321	0.8009	**0.9599**

The figures of merits shows the bag-level classification performance of our proposed LCMIL algorithm. As presented in Table 5.2, our method clearly outperforms the other SoA approaches for MIL classification. Note that the performance upgrade is considerably higher in terms of specificity. This is due to what has been mentioned early, non-specific approaches tend to confuse neutrophils with other similar small cells what mistakenly outgrows the regions with detected inflammation. This handicap is solved with our specific approach where attention loss constraints collaborate with the MIL cross-entropy to accurately classify the WSI at bag-level while detecting the neutrophils.

As has been mentioned in previous sections, one of the most important features of the available database is its multi-center origin. Although all the slides have been extracted under the protocols of the same international project, image acquisition techniques can vary from one hospital to another, for example, in terms of specimen orientation in the glass crystal or inconsistent staining.

For that purpose, we perform an even greater external validation of the LCMIL algorithm with up to 399 whole-slide images coming from seven different centers which slides were not present in the training set. Centers have been named from C1 to C7 for privacy concerns. The figures of merit for the prediction of the WSI from the different centers are presented in Table 5.3. Note that the first row of the table (N) indicates the number of slides in each hospitals.

Table 5.3: Test results in the full cohort to perform external validation of the LCMIL algorithm.

	C1	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**All**
N	91	85	38	77	28	41	39	**399**
SN	0.9091	0.8182	0.9375	0.7083	1	1	0.9412	**0.9032**
SPC	0.8621	0.7703	0.9546	0.9057	0.7647	0.9333	1	**0.8525**
PPV	0.7895	0.3462	0.9375	0.7727	0.7333	0.9630	1	**0.7955**
NPV	0.9434	0.9661	0.9546	0.8667	0.9583	0.9583	0.9583	**0.9583**
F1S	0.9583	0.8293	0.9302	0.8727	1	1	0.7143	**0.9328**
ACC	0.8791	0.7765	0.9474	0.8442	0.8571	0.9756	0.9487	**0.8722**

As it is observed in the present table, the final algorithm performs worse in center C2 with a minimum of 77.65% accuracy and a low 34.62% predictive positive value so further evaluation of the slides in the set would be needed to evaluate its variability.However, metrics are pretty consistent across the different centers reaching an accuracy higher than 85% in five of them and a solid 93.28% F1 Score in the complete set. This results shows the generalization capability of the model across slides with different sources of variability. It also demonstrates that stain color normalization, a technique to deal with different staining color intensities, is not needed thanks to the demonstrated model robustness.

The proposed LCMIL algorithm takes help of the attention constraints module to locate the neutrophils. Remember that this small cells are the key determinant for grading the PHRI, the histological score used to determine histological remission (HR) or ulcerative colitis activity with a cut-off equal to 0. PHRI grades each biopsy in a 0 to 4 scale after evaluating the presence of neutrophils in four key biopsy compartments such as lamina propria, surface epithelium, crypt abscess and cryptal epithelium.

So, it makes sense to think that the number of neutrophils will be higher in the WSI with a higher PHRI grade. The model specificity, i.e. the rate of WSI in HR accurately predicted, is 85.25% while the algorithm sensibility, i.e. the rate of WSI with UC activity correctly predicted, reaches up to 90.32%. By the way, the sensibility can be analyzed deeper because it includes the four different grades of PHRI (1-4) in which the presence of UC acitivty is considered.

For that purpose, we decompose the results of the biopsies with a PHRI greater than zero in Table 5.4. Note that T indicates the number of WSI with a PHRI equal to each corresponding grade while TP refers to the total of WSI accurately predicted for this grade. By the way, SN* measures the partial sensitivity of the algorithm measures as TP/T. The partial sensibility increases around 10% for the slides with a 3-4 PHRI grade in comparison with slides with less presence of UC activity ranged in Here it is demonstrated that the LCMIL efficiently detects neutrophils which clearly helps the model to accurately classify slides with a higher PHRI grade.

Table 5.4: LCMIL performance for different PHRI grades

	TP	**T**	**SN***
PHRI = 1	38	44	0.8636
PHRI = 2	23	29	0.7931
PHRI = 3	34	36	0.9444
PHRI = 4	45	46	0.9782
PHRI > 0	140	155	0.9032

Chapter 6

Conclusions and future lines

Contents

6.1 Conclusions

Ulcerative colitis (UC) incidence is rising worldwide and it is expected to suppose a challenge to clinical centers due to the increased workload and the difficulties in management. As has been presented during the project, multiple scoring systems have been proposed for grading UC in both the endoscopic and histologic areas. The last clinical guidelines proposed to consider histologic remission as the aim of treatment and a novel, simplified and neutrophil-based index (PHRI) has been proposed to evaluate HR. For that purpose, the TFM pretended to develop an automatic deep learning-based algorithm for the prediction of HR in whole-slide images.

An extensive database of WSI coming from patients with a UC diagnosis was available to validate the PHRI's usefulness as an index with AI applicability and to develop the present project. Among the relevant features of the dataset, we should highlight the multicenter and international origin of the images. This is supposed an additional handicap due to variability in terms of acquisition techniques and color staining, but at the same time allowed us to validate the robustness of the model to predict histological remission.

The dataset also included pixel-level annotation for some WSI which were crucial for the model development in multiple ways. First of all, they indicated the localization of neutrophils and key cells for grading UC according to PHRI as presented in Table 2.1. This labeled data allowed us to reduce the computational cost of the model training without affecting its performance because they are only used in the training phase and not for WSI prediction. Remember that the proposed patient-level partitions were also conditioned by the annotation's availability for each WSI.

Before the development of the algorithms, an extensive literature review was performed to discover the novel techniques used for UC implementation in computer-aided decision (CAD) systems and the classification of WSI. This process revealed that weakly supervised learning (WSL) and more specifically multiple instance learning (MIL) were the state-of-the-art techniques for WSI prediction. In the same vein, deep learning-based for UC prediction only have been explored with endoscopic data, so our work is the first to aim for the prediction of UC with histologic images.

As PHRI can be considered as a neutrophil-based scoring system, for the development of the deep learning models we pretended to take advance of this feature instead of only considering a MIL approach due to its limitations to solve high complexity problems. First of all, a two-step algorithm based on instance-level neutrophil classification and the bag-level prediction was proposed. However, this approach was not ideal due to the presence of cells looking like neutrophils that disturbed the first task of the model. Pixel-level data is hard to obtain due to the time invested in the annotation, but it could provide relevant information to upgrade the performance of the algorithms. In this project, we take maximum advantage of pixel-level annotations by implementing a novel approach based on attention constraints integrated into an end-to-end framework.

The proposed algorithm was called 'Location Constrained Multiple Instance Learning' (LCMIL) and it is based on MIL aggregation and attention constraints. Model development took advantage of previous work carried out by other researchers at CVBLab. The selection of the SeaNet with VGG16 for the CNN backbone was supported by previous publications that demonstrated performance improvements in comparison with other techniques [5]. While the CNN backbone performs feature extraction, the bag-level embedding was obtained weighting the importance of certain instances. For that purpose, the AC module forced the CNN to identify the neutrophils by implementing a soft constrained term to the MIL loss. This was based on the L2 norm between the rough annotation of the database and pseudo-CAMs obtained by the UC module by aggregating and resizing the feature maps obtained by the CNN extractor. The compound loss was optimized end-to-end to perform WSI prediction classification.

In the last chapter, we aimed not only to present the results but also a further explanation of how was the model constructed. For that purpose, a series of explanatory ablation experiments are firstly presented. Here we highlight the presentation of Class Activation Maps (CAMs) that are useful in computer vision problems. While AI-based algorithms usually work as a black box, these systems are not suitable in the medical context where clinicians need information about the decision-making process of the algorithms.

The presented CAMs showed that the novel approach based on constraints improved the WSI prediction ability thanks to a more accurate recognition of neutrophils at the instance level. The instance weights for feature aggregation are also relevant because they proved to mimic the diagnostic process of clinicians that only consider a few instances where the neutrophils are located.

The most significant conclusion about the results is the model's ability to classify WSI from different clinical centers. Performance metrics across all centers showed the model robustness to variability due to color staining or acquisition procedures. This supposes a relevant contribution to the work due to model development commonly uses data coming from one medical center that could not be representative of the population.

Finally, LCMIL showed improved sensitivity for the detection of WSI with a PHRI equal to or higher than 3 demonstrating the effectiveness of the attention module. It also showed a performance improvement in comparison with other state-of-the-art techniques for MIL approaches. In conclusion, we reached the goal of the project that aimed to design an automatic deep learning-based system for histological remission sing with histologic images. As no other approaches have used WSI for UC diagnosis, we presented for publication in clinical journals the algorithms and results obtained in this project.

6.2 Future lines

One of the main handicaps of the present project is that the proposed LCMIL algorithm is only capable of predicting histological remission versus UC activity. However, the PHRI scoring system grades each biopsy on a 5-grade scale which includes HR remission (PHRI = 0) and four different grades (1-4) to evaluate UC activity. In future lines, improved algorithms could be designed to automatically evaluate each WSI in the full range of PHRI. For that purpose, it makes sense to take help from the nature of the index itself which considers four different compartments to identify neutrophils. For this purpose, a hybrid algorithm of segmentation and WSI classification could make sense. In the first instance, the model would segment the different regions of interest and afterward apply the LCMIL which detects neutrophils and performs bag-level classification to grade each of the regions in the WSI. It would be interesting to train an end-to-end algorithm capable of jointly performing both segmentation and classification tasks.

Detecting UC activity in WSI or grading PHRI can reduce the workload of the pathologist and asses them in the clinical decision process. However, it remains difficult to predict the clinical outcome of the patients after only analyzing histological slides or endoscopic videos. With clinical outcome, we refer to different events that can occur to patients with a UC diagnosis during the relapse phase. These events include surgery, a change of treatment, or hospitalization. As we mentioned, the last two are referred to relapsing episodes where UC symptoms exacerbate and clinical intervention is needed. Surgery intervention usually refers to colectomy which is also used to treat other colon diseases such as cancer and diverticulitis. It consists in removing a small part of the bowel where the inflammation is present or cancer has rapidly evolved to stop its propagation to other sections of the colon. It has been found that the incidence of colectomy in UC is being reduced thanks to early diagnosis and improved treatments [37] while the early and late postoperative complications of colectomy should not be underestimated. [28]

As it has been mentioned before, PHRI design and validation as a simplified score for UC management with applicability in AI-based systems is part of a bigger international project called PICaSSO. The aim of the project is to develop an integrated systems with endoscopic videos analysis and histological evaluation to finally asses a the clinical outcome of the patient. However, at this point it would make more sense to combine endoscopic and histological data. In the deep learning context, late fusion of multimodal information consists on combining the extracted features from different data sources and it has been proposed for emotion classification in videos [26]. Exploring the feature fusion for endoscopic and histological should improve the diagnostic capability of the outcome predictor algorithm. Among the handicaps to deal in the feature fusion, we find the different feature size in the latent space (512 for histology and 2048 for endoscopy) due to the different CNN backbone architectures (VGG16 vs ResNet) used to extract them.

6.3 Contributions

The scientific and clinical research carried out in collaboration of the CVBLab engineers and clinicians from United Kingdom and Italy has led to publications in journals. The first paper introduced the PHRI score, the novel simplified scoring system for ulcerative colitis grading [8].

Here a deep learning algorithm was constructed to prove the simplicity of the index and its applicability AI-based systems. The proposed framework consisted in a two-step algorithm that was first trained to predict neutrophil presence at the instance level. Once the feature extractor was trained, a MIL aggregation embedding based on MIL attention [13] was implemented to obtain the aggregated bag-level feature vector and the final prediction of the WSI. The accuracy for the WSI prediction in this work was around 86%.

This work has also been presented in different medical congresses for the dissemination of the findings. In [41], we presented an validation of the previous algorithm in a larger multicentre dataset composed of 95 WSI.

The proposed LCMIL algorithm was also accepted for publication in the journal *Computer Methods and Programs in Biomedicine* [6]. Here we introduce Location Constrained Multiple Instance Learning algorithm used in this project for the prediction of histologic remission in WSI from patients with a UC diagnosis. It also includes the comparison with the state-of-the-art methods and a presentation of the ablation experiments explained deeper in this thesis.

Bibliography

[1] Jaume Amores. "Multiple instance classification: Review, taxonomy and comparative study". In: *Artificial intelligence* 201 (2013), pp. 81–105.

[2] Gabriele Campanella et al. "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images". In: *Nature medicine* 25.8 (2019), pp. 1301–1309.

[3] Britt Christensen and David T Rubin. "Understanding endoscopic disease activity in IBD: how to incorporate it into practice". In: *Current gastroenterology reports* 18.1 (2016), pp. 1–11.

[4] UK IBD Genetics Consortium, Wellcome Trust Case Control Consortium 2, et al. "Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region". In: *Nature genetics* 41.12 (2009), p. 1330.

[5] Rocıo Del Amor et al. "An attention-based weakly supervised framework for spitzoid melanocytic lesion diagnosis in whole slide images". In: *Artificial intelligence in medicine* 121 (2021), p. 102197.

[6] Rocıo Del Amor et al. "Constrained Multiple Instance Learning for Ulcerative Colitis prediction using Histological Images". In: *Computer Methods and Programs in Biomedicine* (2022), p. 107012.

[7] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. "Solving the multiple instance problem with axis-parallel rectangles". In: *Artificial intelligence* 89.1-2 (1997), pp. 31–71.

[8] Xianyong Gui et al. "PICaSSO Histologic Remission Index (PHRI) in ulcerative colitis: development of a novel simplified histological score for monitoring mucosal healing and predicting clinical outcomes and its applicability in an artificial intelligence system". In: *Gut* 71.5 (2022), pp. 889–898.

[9] Jianxing He et al. “The practical implementation of artificial intelligence technologies in medicine”. In: *Nature medicine* 25.1 (2019), pp. 30–36.

[10] Jie Hu, Li Shen, and Gang Sun. “Squeeze-and-excitation networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7132–7141.

[11] Marietta Iacucci et al. “An international multicenter real-life prospective study of electronic chromoendoscopy score PICaSSO in Ulcerative Colitis”. In: *Gastroenterology* 160.5 (2021), pp. 1558–1569.

[12] Marietta Iacucci et al. “Development and reliability of the new endoscopic virtual chromoendoscopy score: the PICaSSO (Paddington International Virtual ChromoendoScopy ScOre) in ulcerative colitis”. In: *Gastrointestinal endoscopy* 86.6 (2017), pp. 1118–1127.

[13] Maximilian Ilse, Jakub Tomczak, and Max Welling. “Attention-based deep multiple instance learning”. In: *International conference on machine learning*. PMLR. 2018, pp. 2127–2136.

[14] Howard Lee and Yi-Ping Phoebe Chen. “Image based computer aided diagnosis system for cancer detection”. In: *Expert Systems with Applications* 42.12 (2015), pp. 5356–5365.

[15] Bin Li, Yin Li, and Kevin W Eliceiri. “Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 14318–14328.

[16] Zhizhong Li and Derek Hoiem. “Learning without forgetting”. In: *IEEE transactions on pattern analysis and machine intelligence* 40.12 (2017), pp. 2935–2947.

[17] Ming Y Lu et al. “Data-efficient and weakly supervised computational pathology on whole-slide images”. In: *Nature biomedical engineering* 5.6 (2021), pp. 555–570.

[18] F Magro et al. “European consensus on the histopathology of inflammatory bowel disease”. In: *Journal of Crohn’s and Colitis* 7.10 (2013), pp. 827–851.

[19] Aude Marchal-Bressenot et al. “Development and validation of the Nancy histological index for UC”. In: *Gut* 66.1 (2017), pp. 43–49.

[20] Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. “Imposing hard constraints on deep networks: Promises and limitations”. In: *arXiv preprint arXiv:1706.02025* (2017).

[21] Mahmoud H Mosli et al. “Development and validation of a histological index for UC”. In: *Gut* 66.1 (2017), pp. 50–58.

[22] Mohammed Bany Muhammad and Mohammed Yeasin. “Eigen-cam: Class activation map using principal components”. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2020, pp. 1–7.

[23] Vikram Narang et al. “Association of endoscopic and histological remission with clinical course in patients of ulcerative colitis”. In: *Intestinal research* 16.1 (2018), p. 55.

[24] Muhammad Khalid Khan Niazi, Anil V Parwani, and Metin N Gurcan. “Digital pathology and artificial intelligence”. In: *The lancet oncology* 20.5 (2019), e253–e261.

[25] Tsuyoshi Ozawa et al. “Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis”. In: *Gastrointestinal endoscopy* 89.2 (2019), pp. 416–421.

[26] Yagya Raj Pandeya and Joonwhoan Lee. “Deep learning-based late fusion of multimodal information for emotion classification of music video”. In: *Multimedia Tools and Applications* 80.2 (2021), pp. 2887–2905.

[27] Sunhee Park et al. “Histological disease activity as a predictor of clinical relapse among patients with ulcerative colitis: systematic review and meta-analysis”. In: *Official journal of the American College of Gastroenterology| ACG* 111.12 (2016), pp. 1692–1701.

[28] L Peyrin-Biroulet et al. “Systematic review: outcomes and post-operative complications following colectomy for ulcerative colitis”. In: *Alimentary pharmacology & therapeutics* 44.8 (2016), pp. 807–816.

[29] Kenneth W Schroeder, William J Tremaine, and Duane M Ilstrup. “Coated oral 5-aminosalicylic acid therapy for mildly to moderately active ulcerative colitis”. In: *New England Journal of Medicine* 317.26 (1987), pp. 1625–1629.

[30] Ramprasaath R Selvaraju et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[31] Ahmed Ismail Shahin et al. “A novel white blood cells segmentation algorithm based on adaptive neutrosophic similarity score”. In: *Health information science and systems* 6.1 (2018), pp. 1–12.

[32] Zhuchen Shao et al. “Transmil: Transformer based correlated multiple instance learning for whole slide image classification”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 2136–2147.

[33] Mark S Silverberg et al. “Toward an integrated clinical, molecular and serological classification of inflammatory bowel disease: report of a Working Party of the 2005 Montreal

World Congress of Gastroenterology". In: *Canadian journal of gastroenterology* 19.Suppl A (2005), 5A–36A.

[34] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[35] Christian Szegedy et al. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[36] Kento Takenaka et al. "Development and validation of a deep neural network for accurate evaluation of endoscopic images from patients with ulcerative colitis". In: *Gastroenterology* 158.8 (2020), pp. 2150–2157.

[37] Laura E Targownik et al. "The epidemiology of colectomy in ulcerative colitis: results from a population-based cohort". In: *Official journal of the American College of Gastroenterology| ACG* 107.8 (2012), pp. 1228–1235.

[38] Simon PL Travis et al. "Developing an instrument to assess the endoscopic severity of ulcerative colitis: the Ulcerative Colitis Endoscopic Index of Severity (UCEIS)". In: *Gut* 61.4 (2012), pp. 535–542.

[39] Ryan Ungaro et al. "Ulcerative colitis". In: *The Lancet* 389.10080 (2017), pp. 1756–1770. ISSN: 0140-6736. DOI: https://doi.org/10.1016/S0140-6736(16)32126-2. URL: https://www.sciencedirect.com/science/article/pii/S0140673616321262.

[40] Niels Vande Casteele et al. "Utilizing Deep Learning to Analyze Whole Slide Images of Colonic Biopsies for Associations Between Eosinophil Density and Clinicopathologic Features in Active Ulcerative Colitis". In: *Inflammatory Bowel Diseases* 28.4 (2022), pp. 539–546.

[41] V Villanacci et al. "OP15 A new simplified histology artificial intelligence system for accurate assessment of remission in Ulcerative Colitis". In: *Journal of Crohn's and Colitis* 16.Supplement_1 (2022), pp. i015–i017.

[42] Zhi-Hua Zhou. "A brief introduction to weakly supervised learning". In: *National science review* 5.1 (2018), pp. 44–53.

Part II

Budget

Chapter 7

Budget

Contents

7.1 Aim

This section aims to provide a financial evaluation of the economic costs to carry out the present project based on the development of deep learning models for ulcerative colitis prediction in whole-slide images.

7.2 Partial budgets

The total budget of the project has been divided into three partial budgets: personnel, software, and hardware costs. This partition allows us to explain each of the three partial budgets in more detail.

7.2.1 Personnel costs

This section pretends to take into account the human resources that have been necessary to carry out the project. The development of this project has involved three researchers:

- Dª Valery Naranjo Ornedo, full university professor.
- Dª Rodío del Amor del Amor, Ph.D. student.
- D Pablo Meseguer Esbri, masters student.

The student can be considered a junior biomedical engineer and has been in charge of the development of the bulk of the project and the writing of this report. As the master's thesis has a cost of 24 ECTS, the estimated time spent by the student on the project development is 600 hours.

The work has been mentored by a full professor of the 'Universitat Politècnica de València'. Her tasks consisted mentoring of the project and the review of the final manuscript. At the same time, a Ph.D. student supervised the development of all the parts of the project. The breakdown of the personnel costs is presented in Table 7.1.

Description	**Duration (h)**	**Unitary cost (€/h)**	**Total cost (€)**
University professor	30	30	900
Ph.D student	50	20	1000
Student	600	12	7200
TOTAL			**9100 €**

Table 7.1: Breakdowns of personnel costs

7.2.2 Hardware costs

In the following, the hardware costs are detailed including the ones associated with the personal computer and the specific hardware resources for deep model training.

The laptop used for carrying out the core of the project including the programming and writing of the manuscript is a Lenovo Legion Y520. It includes an i7 Intel Core®processor and a graphic card NVIDIA GTX 1050. As its computing power is still limited, for training the neural networks an NVIDIA DGX A100 system, a property of the CVBLab, has been used. Its global cost is around 200.000€, but only one of its eight GPUs has been occupied for this project so its partial cost is 25.000€. The breakdown of the hardware costs is presented in Table 7.2.

Description	Units (uds)	Unitary cost (€/uds)	Useful life (months)	Use time (months)	Total cost (€)
LENOVO LEGION Y520	1	999	72	8	111
NVIDIA DGX A100 (1 GPU)	1	25000	120	3	625
TOTAL					**736€**

Table 7.2: Breakdown of hardware costs.

7.2.3 Software costs

In the following, the software costs are presented and they include the licenses for the computerized systems and programming environments.

The MobaXTerm software was used to create a secure SSH connection between the student's computer and the DGX for model training in the high-capacity GPUs. Matlab (Mathworks ®) and its R2022.a version has also been employed to perform statistical analysis, among others. The breakdown of the software costs is presented in Table 7.3.

Some free software resources have also been used such as Overleaf for the writing of the document and the Keras library for deep learning model development in Python. This has not been included in the following table.

Description	Units (uds)	Unitary cost (€/uds)	Useful life (months)	Use time (months)	Total cost (€)
Matlab R2022.a	1	800	12	6	400
MobaXTerm Professional	1	60	12	6	30
TOTAL					**430€**

Table 7.3: Breakdown of software costs.

7.3 Total project

To conclude the economical evaluation of the project, the total budget is presented in Table 7.4. It is calculated as the sum of the partial costs that have been presented in the previous section.

Description	Cost (€)
Personnel costs	9100
Software budget	430
Hardware costs	736
Total budget	**10266€**

Table 7.4: Breakdown of the execution budget of the project.

Finally, it is obtained the total amount that the development of this Master's Thesis would entail. It is calculated by adding the general expenses and the industrial profit to the material execution costs recently calculated. Specifically, the industrial profit corresponds to the 6% of the general expenses and a percentage of 13% for the general expenses. In addition, the corresponding taxes for Spain are also incorporated with *Impuesto sobre el Valor Añadido (IVA)* corresponds to the 21% of the total cost. The total amount is shown in Table 1.7.

Description	Cost (€)
Execution budget	10266
General expenses	1334.58
Industrial profit	615.96
SUM	**12216.54€**
IVA (21%)	2565.47€
TOTAL BUDGET	**14782.01€**

Table 7.5: Breakdown of the total budget of the project.

Therefore, the total projected budget is **FOURTEEN THOUSAND SEVEN HUNDRED AND EIGHTY-TWO EUROS AND ONE CENT**.