



Instituto Superior de Engenharia

Politécnico de Coimbra

DEPARTMENT OF SYSTEMS AND COMPUTER
ENGINEERING

Diagnose of the Inflammatory Bowel Disease using Deep Learning Models with Interpretability

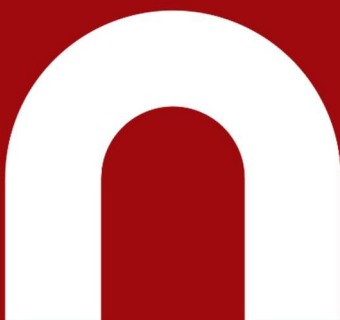
Dissertation to fulfill the Master's degree in Informatics Engineering
Specialization in Intelligent Data Analysis

Author

José Carlos Nunes Maurício

Supervisor

Inês Campos Monteiro Sabino Domingues



INSTITUTO POLITÉCNICO
DE COIMBRA

INSTITUTO SUPERIOR
DE ENGENHARIA
DE COIMBRA

Coimbra, Julho 2023

RESUMO

Doença de Crohn e Colite Ulcerosa são duas doenças crônicas que provocam uma inflamação nos tecidos de todo trato gastrointestinal, e são caracterizadas pelo termo: Doença Inflamatória do Intestino. Para os pacientes com esta patologia, um diagnóstico preciso e atempado é um ponto chave para a qualidade de vida das pessoas com o objetivo de ser prescrito um tratamento que minimize os sintomas provocados pela inflamação do intestino. Contudo, os médicos gastroenterologistas sentem alguma dificuldade em avaliar as imagens endoscópicas de modo a reconhecer as características das duas doenças crônicas. Portanto, através deste trabalho procura-se encontrar uma forma de automatizar o diagnóstico destes pacientes durante o exame endoscópico reduzindo a subjetividade que está sujeito pela observação de um gastroenterologista, utilizando seis CNNs: AlexNet, ResNet50, VGG16, InceptionV3, ResNet50-MobileNetV2 e um modelo híbrido. Foram ainda utilizadas neste estudo cinco ViTs: ViT-B/32, ViT-S/32, ViT-B/16, ViT-S/16 and R26+S/32. *Knowledge distillation* foi aplicada de forma a construir modelos mais simples, com menos parâmetros, com base na aprendizagem das arquiteturas pré-treinadas em grandes volumes de dados. Conclui-se que no âmbito das ViTs consegue-se reduzir 25x o número de parâmetros mantendo uma boa performance e reduzir o tempo de inferência em 5,32 segundos. Para as CNNs os resultados mostram que é possível reduzir em 107x o número de parâmetros, reduzindo por consequência o tempo de inferência em 3,84 segundos. Aliado a isto, ainda foi realizada a interpretabilidade dos modelos que permitiu concluir que ambas as arquiteturas (*teacher and student models*) conseguem identificar úlceras, situações de sangramento e lesões provocadas pela inflamação da doença.

Palavras-chave: Colite Ulcerosa, Doença de Crohn, *Deep Learning*, Destilação do Conhecimento, DeiT

ABSTRACT

Crohn's Disease and Ulcerative Colitis are two chronic diseases that cause inflammation in the tissues of the entire gastrointestinal tract, and are characterised by the term: Inflammatory Bowel Disease. For patients with this condition, an accurate and timely diagnosis is a key point in people's lives in order to prescribe a treatment that minimises the symptoms caused by inflammation of the intestine. However, gastroenterologists find it difficult to evaluate endoscopic images in order to recognise the characteristics of the two chronic diseases. Therefore, through this work, we seek to find a way of automating the diagnosis of these patients during the endoscopic examination, reducing the subjectivity that is subject to the observation of a gastroenterologist, using six CNNs: AlexNet, ResNet50, VGG16, InceptionV3, ResNet50+MobileNetV2 and Hybrid model. Also, five ViTs were used in this study: ViT-B/32, ViT-S/32, ViT-B/16, ViT-S/16 and R26+S/32. Knowledge distillation was applied to build simpler models, with fewer parameters, based on the learning of the pre-trained architectures on large volumes of data. It is concluded that in the ViTs framework, it is possible to reduce 25x the number of parameters while maintaining good performance and reducing the inference time by 5.32 seconds. For CNNs, the results show that it is possible to reduce 107x the number of parameters, reducing consequently the inference time in 3.84 seconds. Allied with this, the interpretability of the models was carried out and it was concluded that both architectures (teacher and student models) can identify ulcers, bleeding situations and lesions caused by the inflammation of the disease.

Keywords: Ulcerative Colitis, Crohn's Disease, Deep Learning, Knowledge Distillation, DeiT

EPÍGRAFE

Let us strive for the impossible. The great achievements throughout history have been
the conquest of what seemed the impossible.

Charlie Chaplin

DEDICATÓRIA

Dedico esta dissertação de mestrado aos meus pais que me têm apoiado incondicionalmente ao longo da minha vida, e mais ainda durante a realização desta dissertação. Serão sempre uma inspiração para mim, seja pela sua resiliência, seja pela sua capacidade de trabalho, seja pela sua força que têm na luta por aquilo que acreditam.

AGRADECIMENTOS

A realização desta dissertação de mestrado foi um longo caminho a percorrer, que só foi possível fazê-lo através do apoio e contributo de várias pessoas. A quem eu estou muito grato por me terem apoiado fortemente nesta conquista e deixo os meus sinceros agradecimentos.

À minha orientadora, professora Doutora Inês Domingues, que foi exímia no processo de orientação. Demonstrou desde o início muito interesse pelo tema e preocupação em encontrar uma solução para a resolução do problema. Agradeço-lhe por todo o seu empenho, todas as suas contribuições, por ter sempre acreditado em mim e no projeto, por me ter motivado e incentivado nesta longa caminhada. Tendo sido esta forma de estar da professora determinantes para o sucesso deste projeto.

Ao meu amigo e colega, Francisco Mesquita, que ajudou a tornar todo este processo mais facilmente ultrapassável graças ao seu companheirismo e às trocas de ideias que fomos fazendo durante a elaboração das nossas dissertações.

Aos meus pais, que tiveram um papel importante nesta jornada. A quem dedico esta dissertação e a quem devo um agradecimento especial. Foi graças ao seu apoio incondicional e ao seu incentivo que eu consegui percorrer todo este caminho. Apesar de por vezes ter sacrificado alguns momentos em família para me poder dedicar a este projeto, eles sabiam o quanto era importante para mim. Agradeço-lhes a sua compreensão e por me terem apoiado sempre nos bons e maus momentos.

E, a todas as outras pessoas que não mencionei, mas que contribuíram de alguma forma para a minha motivação na realização deste projeto.

INDEX

Resumo	i
Abstract	ii
Epígrafe	iii
Dedicatória	iv
Agradecimentos	v
Index	1
Index of tables	3
Index of figures	4
List of abbreviations	6
List of acronyms	7
1 Introduction	8
2 Background Knowledge	10
2.1 Inflammatory Bowel Disease and Diagnose	10
2.2 Deep Learning models	14
2.3 Knowledge Distillation	16
2.4 Interpretability	20
3 Literature review	22
3.1 Vision Transformers vs Convolutional Neural Networks	22
3.1.1 Research Methodology	22
3.1.2 Findings	25
3.2 Deep Learning to Detect Inflammatory Bowel Disease	30
3.2.1 Research Methodology	30
3.2.2 Findings	31

4	Methodology and Data	39
4.1	Description of dataset	39
4.2	Experimental setup	44
4.3	Pre-processing	45
4.4	Data augmentation	47
4.5	Convolutional Neural Networks	47
4.6	Vision Transformers	47
4.7	Knowledge Distillation	49
4.8	Interpretability	50
4.9	Evaluation	50
5	Results	53
5.1	Convolutional Neural Networks	53
5.2	Vision Transformers	56
6	Interpretability	60
7	Discussion	70
8	Conclusion	72
8.1	Strengths	72
8.2	Limitations	73
8.3	Future work	73
	References	74
	Appendices	85
	Appendix A - Articles	86
	Appendix B - Supplementary material	136

INDEX OF TABLES

3.1	Data sources and number of results obtained for the literature review of ViTs vs CNNs.	23
3.2	Data sources and search string used for the literature review of ViTs vs CNNs. .	23
3.3	List of selected studies for the literature review of ViTs vs CNNs.	24
3.4	Data sources and the number of results obtained for the literature review of IBD.	30
3.5	Data sources and search string or title used for the literature review of IBD. . .	31
3.6	List of selected studies for the literature review of IBD	32
3.7	Overview of studies selected for the literature review of ViTs vs CNNs.	37
3.8	Overview of studies selected for the literature review of IBD.	38
4.1	Description of the type of lesions presented in the CrohnIPI dataset	43
4.2	Description of the created datasets.	45
4.3	Configurations of the nine performed experiments	46
5.1	Results of the CNNs in experiment 1.	53
5.2	Results of the CNNs in experiment 2.	53
5.3	Results of the CNNs in experiment 3.	54
5.4	Results of the CNNs in experiment 4.	54
5.5	Results of the CNNs in experiment 5.	54
5.6	Results of the CNNs in experiment 6.	55
5.7	Results of the CNNs in experiment 7.	55
5.8	Results of the CNNs in experiment 8.	55
5.9	Results of the CNNs in experiment 9.	55
5.10	Results of the ViTs in experiment 1.	56
5.11	Results of the ViTs in experiment 2.	56
5.12	Results of the ViTs in experiment 3.	57
5.13	Results of the ViTs in experiment 4.	57
5.14	Results of the ViTs in experiment 5.	57
5.15	Results of the ViTs in experiment 6.	58
5.16	Results of the ViTs in experiment 7.	58
5.17	Results of the ViTs in experiment 8.	58
5.18	Results of the ViTs in experiment 9.	59

INDEX OF FIGURES

2.1	Illustration of imaging exams for the IBD diagnosis	12
2.2	Examples of the different Mayo score degrees	14
2.3	Example of the architecture CNN	15
2.4	Example of the architecture ViT	16
2.5	Structure of the knowledge distillation	17
2.6	Example of the Response-base knowledge	18
2.7	Example of the architecture of DeiT	19
3.1	Yearly distribution of selected studies comparing ViTs and CNNs	24
3.2	Distribution of the selected studies by application area	25
3.3	Yearly distribution of selected studies on IBD	32
4.1	Experimental setup	40
4.2	Images regarding the 8 classes present on the Kvasir dataset	41
4.3	Class distribution for lower gastrointestinal tract pathologies	42
4.4	Ulcerative colitis image from the HyperKvasir dataset	44
4.5	Class distribution of the three datasets	45
4.6	Pre-processing examples	46
4.7	Architecture of the ResNet50+MobileNetV2 model	48
4.8	Architecture of the hybrid model	48
6.1	Interpretability of the CNNs teacher models in experiment 1	60
6.2	Interpretability of the CNNs teacher models in experiment 2	61
6.3	Interpretability of the CNNs teacher models in experiment 3	61
6.4	Interpretability of the CNNs teacher models in experiment 4	62
6.5	Interpretability of the CNNs teacher models in experiment 5	62
6.6	Interpretability of the CNNs teacher models in experiment 6	63
6.7	Interpretability of the CNNs teacher models in experiment 7	63
6.8	Interpretability of the CNNs teacher models in experiment 8	64
6.9	Interpretability of the CNNs teacher models in experiment 9	64
6.10	Interpretability of the ViTs teacher models in experiment 1	65
6.11	Interpretability of the ViTs teacher models in experiment 2	65
6.12	Interpretability of the ViTs teacher models in experiment 3	66
6.13	Interpretability of the ViTs teacher models in experiment 4	66

6.14	Interpretability of the ViTs teacher models in experiment 5	67
6.15	Interpretability of the ViTs teacher models in experiment 6	67
6.16	Interpretability of the ViTs teacher models in experiment 7	68
6.17	Interpretability of the ViTs teacher models in experiment 8	68
6.18	Interpretability of the ViTs teacher models in experiment 9	69
8.1	Interpretability of the CNNs student models in experiment 1	136
8.2	Interpretability of the CNNs student models in experiment 2	137
8.3	Interpretability of the CNNs student models in experiment 3	137
8.4	Interpretability of the CNNs student models in experiment 4	138
8.5	Interpretability of the CNNs student models in experiment 5	138
8.6	Interpretability of the CNNs student models in experiment 6	139
8.7	Interpretability of the CNNs student models in experiment 7	139
8.8	Interpretability of the CNNs student models in experiment 8	140
8.9	Interpretability of the CNNs student models in experiment 9	140
8.10	Interpretability of the ViTs student models in experiment 1	141
8.11	Interpretability of the ViTs student models in experiment 2	141
8.12	Interpretability of the ViTs student models in experiment 3	142
8.13	Interpretability of the ViTs student models in experiment 4	142
8.14	Interpretability of the ViTs student models in experiment 5	143
8.15	Interpretability of the ViTs student models in experiment 6	143
8.16	Interpretability of the ViTs student models in experiment 7	144
8.17	Interpretability of the ViTs student models in experiment 8	144
8.18	Interpretability of the ViTs student models in experiment 9	145
8.19	Interpretability of the DeiT student models in experiment 1	145
8.20	Interpretability of the DeiT student models in experiment 2	146
8.21	Interpretability of the DeiT student models in experiment 3	146
8.22	Interpretability of the DeiT student models in experiment 4	147
8.23	Interpretability of the DeiT student models in experiment 5	147
8.24	Interpretability of the DeiT student models in experiment 6	148
8.25	Interpretability of the DeiT student models in experiment 7	148
8.26	Interpretability of the DeiT student models in experiment 8	149
8.27	Interpretability of the DeiT student models in experiment 9	149

LIST OF ABBREVIATIONS

Acc Accuracy

Pre Precision

ViT Vision Transformers

LIST OF ACRONYMS

AUC	Area Under Curve
AUCROC	Area Under the Curve of the Receiver Operating Characteristic
CD	Crohn's disease
CNN	Convolutional Neural Networks
DNN	Deep Neural Networks
FN	False Negative
FP	False Positive
Grad-CAM	Gradient-weighted Class Activation Mapping
IBD	Inflammatory Bowel Disease
KD	Knowledge Distillation
K-NN	K-Nearest Neighbors
LIME	Local Interpretable Model-agnostic Explanations
MCC	Mathews Correlation Coefficient
MES	Mayo Endoscopic Score
NPV	Negative Predictive Value
PPV	Positive Predictive Value
RISE	Randomized Image Sampling for Explanations
Shap	SHapley Additive exPlanations
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TNR	True Negative Rate
TPR	True Positive Rate
UC	Ulcerative colitis

1 INTRODUCTION

Inflammatory Bowel Disease is the term used to classify two chronic diseases: Ulcerative Colitis and Crohn's Disease, which cause digestive derangements and inflammation in the entire gastrointestinal tract, from the mouth to the anus. So far, its cause is scientifically unknown. However, some environmental factors that interact with genetics cause a reaction in the immune system [1, 2]. Crohn's disease is a chronic inflammatory disease that causes inflammation in the mucosa of the intestine, most commonly in the lower part of the small intestine, but can cause inflammation anywhere in the gastrointestinal tract, from the mouth to the anus. It occurs in patients aged 15-35 years and causes pain, diarrhea, fever, and other symptoms. On the other hand, ulcerative colitis is associated with blood in the stool, intense pain, and diarrhea. Unlike CD, UC affects the mucosal layer of the colon, causing lesions in the large intestine and rectum [3].

The number of patients with this condition (IBD) is increasing worldwide and can affect people of all ages, including children and the elderly. Unlike other inflammatory diseases, there is no medical treatment that can cure inflammation on the intestine. However, an accurate and early diagnosis allows the gastroenterologist to prescribe a treatment that minimises the symptoms of the disease and offers a quality of life to patients. In addition, if the inflammation is not treated in time, it can lead to colon cancer and damage to the walls of the intestine [4, 5].

The use of deep learning models to help the gastroenterologist in the diagnosis of patients with inflammatory bowel disease is important and, with time, more necessary. Not only in Portugal but worldwide the number of people who have this chronic disease is increasing and its diagnosis depends on the evaluation that the gastroenterologist makes of the endoscopic images, which makes this diagnosis subject to a great subjectivity [3]. Therefore, it is crucial to find tools that automate the diagnosis of the disease to prescribe the most suitable treatment for the patient.

Therefore, this work aims to find out through several experiments if, using deep learning algorithms, it is possible to automate the process of IBD diagnosis, and make it more accurate. It also seeks, through the implementation of Knowledge distillation, to compress these models trained on large volumes of data for lighter architectures and easy implementation in endoscopic systems, in order to allow a fast diagnosis during the endoscopic examination. Ensuring the same accuracy in predicting inflammatory bowel disease. In conjunction, it also aims to make the models more informative for the

gastroenterologist with their interpretability and explainability. With the aim of the medical specialist understanding which characteristics are identified by the algorithm during the prediction of Crohn's disease or Ulcerative colitis.

In this sense, this work contributes to a comparative study, where six CNNs and five ViTs are used to process images collected using colonoscopy and video capsule endoscopic to recognise which type of inflammatory bowel disease is present: Crohn's disease or Ulcerative Colitis. Together with knowledge distillation to simpler models that guarantee similar performance [6] as pre-trained models on large volumes of data. In addition, this work offers a comparative study of five interpretability models: Grad-CAM, LIME, SHAP values, RISE, and Occlusion sensitivity. This allows the gastroenterologist to know which characteristics of the intestinal mucosa the deep learning model has identified are linked to the predicted pathology. This work has made still other contributions:

- Publication of a paper in Applied Sciences: "Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review" [7];
- Publication of a paper in IbPRIA 2023: "Deep Neural Networks to distinguish between Crohn's disease and Ulcerative colitis" [8];
- Elaboration of a paper to be submitted in RECPAD 2023: "Interpretability of Deep Neural Networks to diagnose Inflammatory Bowel Disease" [9];
- Elaboration of a paper to be submitted in CIARP 2023: "Knowledge Distillation of Vision Transformers and Convolutional Networks to Predict Inflammatory Bowel Disease" [10];
- Development of code to implement the deep models ¹.

The organization of this document is divided into seven chapters: chapter 2 presents a theoretical background on the disease and the deep learning approach used; chapter 3 describes the methodology and the findings used in the literature review; chapter 4 describes the created datasets for this study and the implemented experimental methodology; chapter 5 presents the quantitative results obtained by the teachers and student's models; chapter 6 presents the information obtained after the interpretability of the models; chapter 7 compares the results obtained in this study with the results of the other authors and critically assesses the findings made; chapter 8 summarises the findings, suggests some future research directions, and presents the strengths and limitations of the work.

¹Available at: https://github.com/MauricioAI/tese_mestrado_ISEC

2 BACKGROUND KNOWLEDGE

This chapter shows a theoretical background about inflammatory bowel disease and its variants, as well as the medical imaging tests that exist to diagnose patients, the deep learning algorithms used in this work, knowledge distillation for ViTs architectures and CNNs, and the interpretability algorithms selected for this work.

2.1 Inflammatory Bowel Disease and Diagnose

Inflammatory bowel disease is characterized by chronic inflammation throughout the intestinal tract, which joins two different pathologies, such as Crohn's disease and ulcerative colitis. It is estimated that in Portugal it already affects 7,000 to 15,000 people, with 2.9 cases per 100,000 inhabitants per year with ulcerative colitis and 2.4 cases per 100,000 inhabitants per year with Crohn's disease [11].

In Europe, the prevalence of the disease is more than 1 in 198 people for ulcerative colitis and 1 in 310 people for Crohn's disease [12]. And if the comparison is made worldwide, the prevalence of ulcerative colitis is 7.6 in 246 cases per 100,000 people per year, preceded by Crohn's disease with 3.6 in 214 cases per 100,000 people per year. This disease has a higher incidence in North America and Europe. However, in recent decades, it has seen an increase in countries such as China and India [13]. Besides the fact that the number of patients with inflammatory bowel disease has been increasing, there are also limitations in their diagnosis of the disease, which makes it difficult to prescribe treatment on time.

Inflammatory bowel disease (IBD) is characterized as a disease of unknown cause that results from the interaction between genetic and environmental factors, triggering an immune response that causes digestive disorders and inflammation in the gastrointestinal tract. These types of diseases are divided into ulcerative colitis and Crohn's disease [4]. However, some studies have shown that some of the diagnoses of inflammatory bowel disease refer to a different category of disease, called indeterminate colitis. This pathology refers to colitis with an extreme degree of severity that has features of both diseases (ulcerative colitis and Crohn's disease), making it difficult to distinguish between them [3].

IBD can provoke damage in the intestine wall when the inflammation is not controlled over time as [11]. These damages include:

- **Abscesses:** pockets of infection that may rupture through the intestinal wall;

- **Stenosis:** parts of tightening in the bowel;
- **Fistulas:** abnormal connections between two organs or structures, which may be caused by inflammation and pressure inside the intestine that breaks down tissue;
- **Colon cancer:** persistent inflammation in the colon increases the risk of colon cancer.

Crohn's disease affects any part of the gastrointestinal tract, from the mouth to the anus. The severity of the disease and its location determine the signs and symptoms, which have a wide spectrum of clinical presentations. The most common symptoms of Crohn's disease are abdominal pain, watery diarrhea, and weight loss. Abdominal pain is usually caused by colic and can last for years before being diagnosed. It occurs in the right lower quadrant due to the involvement of the terminal ileum and is sometimes confused with appendicitis [14]. Diarrhea in most cases is watery, but it may contain blood in 1% to 2% of patients, depending on the severity of the disease. Weight loss is mostly due to diarrhea, malnutrition, which is caused by the bad absorption of nutrients from the small intestine [4], and anorexia, which results from a fear of eating.

Ulcerative colitis is restricted to the large intestine and involves a superficial inflammatory process. This disease only affects the colon and rectum. Whereas Crohn's disease affects any part of the gastrointestinal tract. Normally, the inflammation starts in the rectum and can be continuously extended to the terminal ileum, where it is terminated by backwash ileitis. Symptoms vary from person to person, but the disease typically manifests as mild abdominal discomfort to acute pain from bowel movements and abdominal cramps [15].

There may also be symptoms such as diarrhea transit intestinal caused by colonic inflammation, in which the degree of severity may be related to the extent of the inflammation. That is, in the case of rectal inflammation, it is associated with the passage of mucus in the stool. On the other hand, in the case of inflammation occurring throughout the colon, it may sometimes be associated with the presence of blood (10% of patients). Approximately 1%–3% of patients have experienced at least one episode of haemorrhage and need surgical intervention. Fatigue, fever, and weight loss may also be present. As the disease progresses, patients may develop strictures in the colon that can lead to obstruction and pain. The Montreal classification system aids in the clinical management of the patient by describing the extent of inflammation of the colon and, most often, the degree of severity of the disease [14].

The diagnosis of inflammatory bowel disease is made through a combination of tests such as endoscopy or colonoscopy and imaging studies (e.g., contrast radiography, magnetic resonance imaging, and computed tomography). In addition to these tests,

stool samples and blood tests may also be requested [16]. Endoscopy consists of using a flexible tube with a micro-camera at the tip (an endoscope) that penetrates through the mouth and passes through the esophagus, the stomach, and the duodenum. The camera at the tip of the tube allows to film the condition of the mucous in the upper gastrointestinal tract. During the examination, it is possible to take samples of the mucous, cells, or fluids from the upper gastrointestinal tract, stop internal bleeding, and perform procedures such as opening strictures [17]. Figure 2.1 shows some examples of imaging exams to diagnose inflammatory bowel disease.

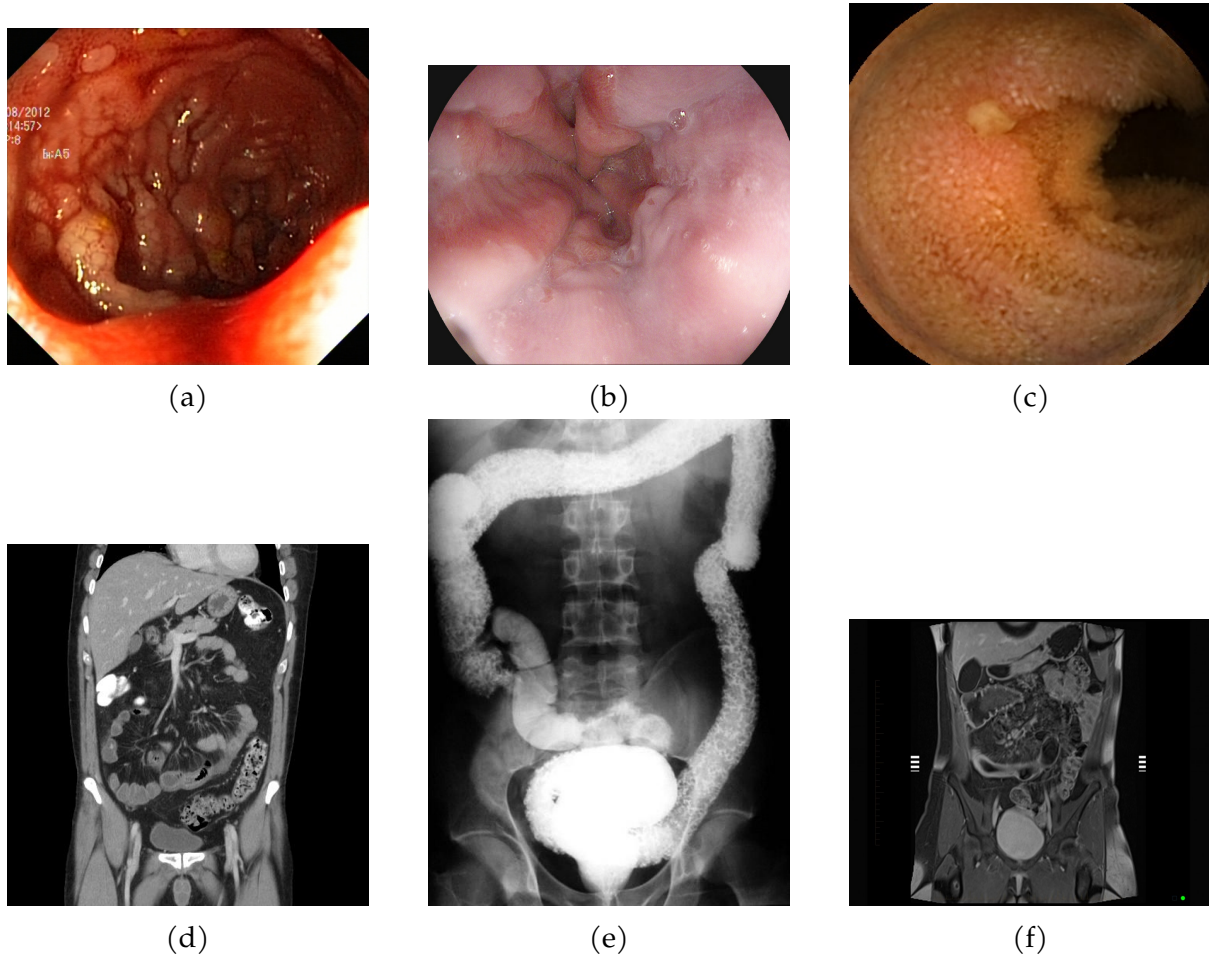


Figure 2.1: Illustration of imaging exams for the IBD diagnosis: (a) Colonoscopy; (b) Endoscopy; (c) Video Capsule Endoscopy; (d) Computed tomography [18]; (e) Contrast radiography; (f) Magnetic resonance imaging [19].

The colonoscopy is performed using a long, flexible tube with a micro-camera at the tip, which is introduced into the rectum and allows the gastroenterologist to look inside the large intestine (colon) for changes, as irritated and swollen tissues, polyps, or cancer. During the colonoscopy, it is still possible to remove abnormal tissue or polyps. Moreover, mucous samples (biopsies) are taken during the examination to study the disease [20].

Contrast radiography is an imaging exam that allows you to visualize organs using X-

rays and the administration of a contrast medium. Organs and tissues such as blood vessels, the stomach, and the colon, unlike bones, cannot easily block conventional X-rays. Therefore, the intake of the contrast serves to highlight these specific areas in the body and helps the radiologist visualize in greater detail the organs in the X-rays image [21].

Computed tomography is a diagnostic imaging procedure, that can be performed with or without contrast. This procedure combines X-rays with computer technology to collect images of the inside of the body. It displays detailed images of any part of the body, including bones, muscles, fat, organs, and blood vessels. On this exam, the X-ray beam that moves in a circle around the body allows you to see the organs or the structures in different ways, providing much more detail, which can help in diagnosing tumours, finding internal bleeding, or analyzing for other internal injuries [22].

Magnetic resonance imaging (MRI) is a non-invasive medical imaging technique that uses a magnetic field and computer-generated radio waves to generate detailed images of the organs and tissues of the body. This imaging technique produces 3D images that can be seen from different angles and can also produce high-resolution images. It seeks to aid in the diagnosis of aneurysms of brain vessels, disorders of the eye and inner ear, multiple sclerosis, spinal cord disorders, tumours, and brain injuries from trauma [20].

Video Capsule Endoscopy consists of an endoscopic capsule that measures 24 to 32 mm in length and 11 to 13 mm in diameter and contains a wireless camera capable of capturing 3 to 40 frames per second (fps). The capsule is ingested by the patient like a pill. After swallowing, it progresses along the gastrointestinal tract until it is excreted by the patient, allowing the mucosa surface to be visualized. A lead sensor array is added around the patient's abdomen, which is connected to the data recorder [23].

This exam plays an important role in the evaluation of lesions and bleeding in the small bowel, but has also been included in the diagnosis of inflammatory bowel disease and screening for colorectal neoplasia. Using software associated with the endoscopic equipment, the gastroenterologist can visualize the images collected from the gastrointestinal tract by the capsule endoscope. Before this exam is performed, patients are advised to fast or consume a liquid diet during 10 to 12 hours before the exam [23].

In the context of IBD, endoscopic monitoring of disease activity plays an important role in the assessment of lesions on the intestinal mucosa provoked by healing inflammation and is necessary to orient patient treatment. Therefore, scoring systems, that allow the estimate of the degree of disease severity, were developed. For ulcerative colitis disease, there are two systems: Mayo score and Ulcerative Colitis Endoscopic Index of Severity (UCEIS). For monitoring Crohn's disease there are Crohn's Disease Endoscopic Index of Severity (CDEIS) and the Simple Endoscopic Score for Crohn's Disease (SES-CD) [24].

Mayo score was developed by [25] and consists of a 0-3 scale measuring the degree of endoscopic appearance, providing information on the severity of UC. The levels are: 0 - Normal or inactive disease (normal mucosa); 1 - Mild disease (erythema, decrease vascular pattern, mild friability); 2 - Moderate disease (marked erythema, absent vascular pattern, friability, erosions); 3 - Severe disease (spontaneous bleeding, ulceration). Figure 2.2 presents an example of each Mayo score. The UCEIS score is an index that evaluates disease activity based on endoscopic findings: i) Vascular pattern; ii) Bleeding; iii) Lesions and ulcers [26]. The authors demonstrated that the UCEIS score is more accurate than the Mayo score in recognising disease remission.

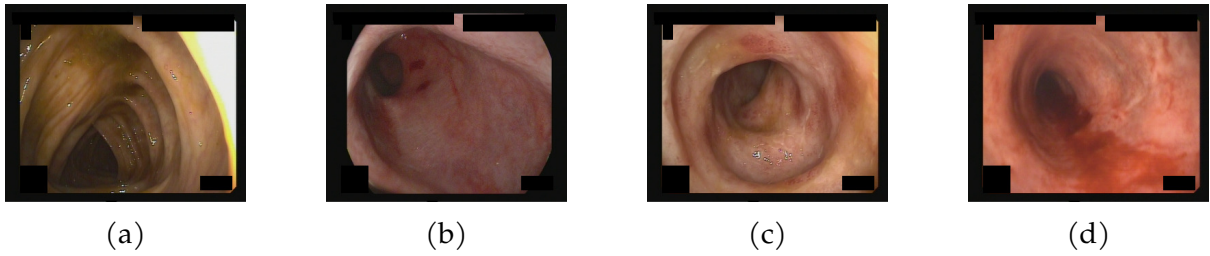


Figure 2.2: Examples of the different Mayo score degrees gathered through the LIMUC dataset [27]: (a) Mayo 0; (b) Mayo 1; (c) Mayo 2; (d) Mayo 3.

CDEIS score checks the presence of deep and superficial ulcers, ulcerated and non-ulcerated stenoses, as well as the extent of disease and ulcers. This index is used to guide the gastroenterologist on the patient's treatment strategies and to predict remission without corticosteroids. It is recognised as an accurate tool for assessing the degree of severity of Crohn's disease [28]. However, due to its complexity, its use in clinical practice is not recommended. In this sense, the SES-CD was developed and is limited to the amount of surface area affected by ulcers, the size of ulcers and the presence of strictures [29]. Its result can be converted to CDEIS through the formula (2.1), due to the demonstrated correlation between the two scores [30].

$$CDEIS = 0.76 \times SES-CD + 0.29 \quad (2.1)$$

The collected images for the elaboration of this study were gathered using two types of mechanisms for the diagnosis of the disease. To collect images concerning ulcerative colitis was used the colonoscopy exam, and the video capsule endoscopy exam was used to diagnose patients with Crohn's disease.

2.2 Deep Learning models

In the scope of pattern recognition in images, CNN networks are the most used architectures. They have been considered a powerful tool for dealing with large amounts of data [31]. This type of network is a specific type of Artificial Neural Network (ANN)

that consists of neurons which self-optimize through learning. These neurons have the function of receiving an input, performing an operation and returning an output [32].

Figure 2.3 presents an illustrative example of a CNN architecture. The typical CNN architecture starts with convolutional layers that pass through the kernels or filters, from left to right of the image, extracting computationally interpretable features. The first layer extracts low-level features (e.g., colours, gradient orientation, edges, etc.), and subsequent layers extract high-level features. Next, the pooling layers reduce the information extracted by the convolutional layers, preserving the most important features. Pooling layers have no trainable parameters [32]. Finally, the fully-connected layers are fed with the flattened output of the convolutional and pooling layers and perform the classification.

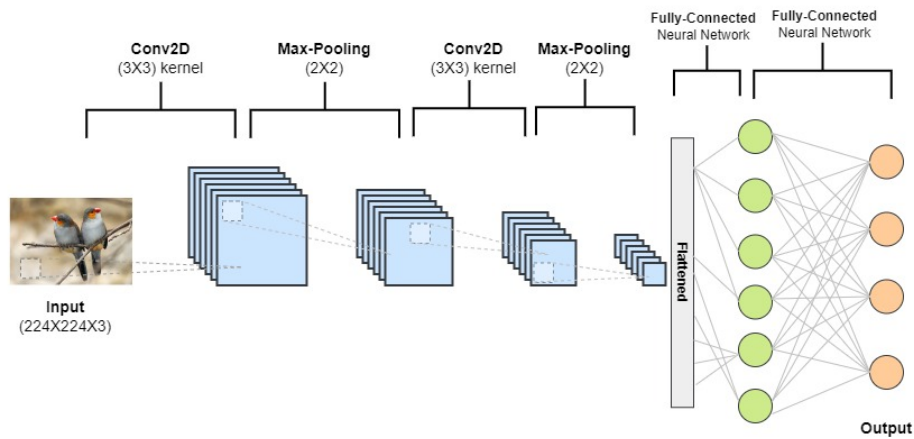


Figure 2.3: Example of the architecture CNN, based on [33].

These networks contain still non-linear layers, that are represented by activation functions like softmax and ReLU, which aim to adjust the output of the previous layer to the next one.

Recently, the authors [34], have proposed a new architecture for images, the ViTs, inspired by the success of Transformers for text. ViTs receive a 2D image as input which is divided into patches of equal size, as the standard transformer is configured to receive as input a sequence of 1D token embeddings. These patches are linearly projected creating a sequence of token embeddings, to which will be added a positional embedding that represents the information of the position of a patch in the original image. A class token is also added, similar to the BERT architecture [35], whose goal is to create an embedding associated with each sequence of embedded patches that allows the output of the Transformer encoder to make a representation of the image and be classified by the MLP structure with a hidden layer. The transformer encoder is the compilation of multiheaded self-attention layers and MLP blocks, which contain two

non-linear GELU layers. LayerNormalization layers are added to these layers to help normalize the learned information, as well as skip connections [34].

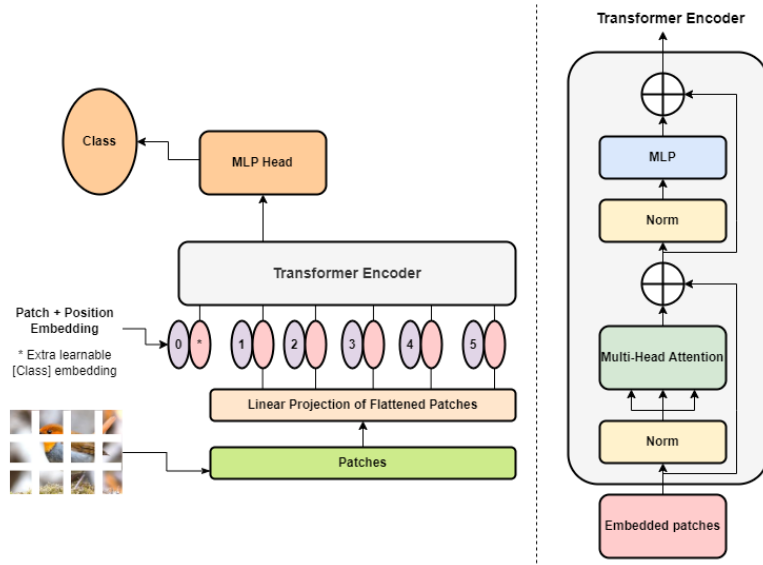


Figure 2.4: Example of the architecture ViT, based on [34].

The authors [34] relied on the configurations used by the BERT architecture to pre-train the ViT architectures. In this sense, some variants were created: Base, Large, Huge and Small with different dimensions. For instance, a variant with the nomenclature ViT-B/32 means that we are in the presence of a Base variant of ViT architectures, whose patch size is 32x32. In addition to these variants, hybrid models were also introduced with the aim of replacing the sequence of patches extracted from the original image with feature maps of CNNs. This means that a CNN is used to extract the feature maps from the original image, and they are linearly projected to patches embedded with the Transformer encoder dimensions.

2.3 Knowledge Distillation

Knowledge Distillation was proposed by [36] and consists in distilling the knowledge learned by a large network, designated as a teacher, to a smaller network, considered a student. It aims to facilitate the deployment of deep models in systems where computational capacity and equipment memory are limited, creating a student model lighter than the teacher model, but with the same performance.

In this scope, it is necessary to take into consideration some concepts: knowledge, distillation and teacher-student architectures [37]. Although the authors propose a type of knowledge and a way of distilling that knowledge, over time other types of knowledge and other ways of distilling it have been proposed. Therefore, the concept of knowledge is intended for the type of knowledge that can be transferred between the teacher

and the student, which consists of three types: Response-based knowledge, Feature-based knowledge and Relation-based knowledge. On the other hand, the concept of distillation translates how the student model will receive the knowledge transferred by the teacher, which can be divided into three categories: Offline distillation, Online distillation and Self-distillation [37], as shown in Figure 2.5.

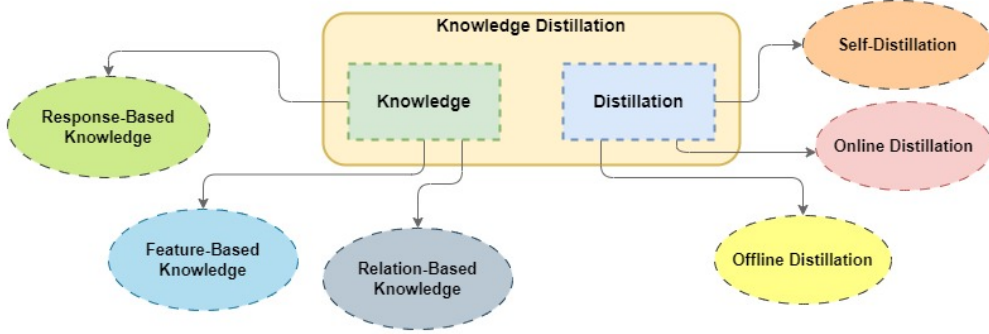


Figure 2.5: Structure of the knowledge distillation, based on [37].

Response-Based knowledge is demonstrated in Figure 2.6 and is based on the distillation of the outputs of the last layer of the teacher model, designated by logits that are computed through a softmax function for a probability associated with each class. This probability can be computed using the formula 2.2, where Z_i represents the logit for the i -th class and the temperature T is used to control the importance of each soft target [37].

$$p(Z_i, T) = \frac{\exp(Z_i/T)}{\sum_j \exp(Z_j/T)} \quad (2.2)$$

The temperature is normally equal to 1, but during distillation training, can be moulded with the aim that the higher it is, the more soft are the produced targets. However, after training, the temperature returns to 1 [36].

Feature-Based knowledge is proposed by [38] and aims to use the outputs of the middle layers (i.e., feature maps) as distilled knowledge for the student model. This methodology adaptively seeks, during distillation, to match the feature maps from the intermediate layers of the teacher model to the intermediate layers of the student model, based on the loss function [37].

Relation-Based knowledge [39], unlike the other ways of transferring knowledge, seeks to explore the relationships between different layers or between data samples. Initially, when it was proposed, it consisted in computing the product between features of two layers and distilling the knowledge based on the correlations between feature maps. However, these correlations may contain not only information about features but also about the relationships between data samples [37].

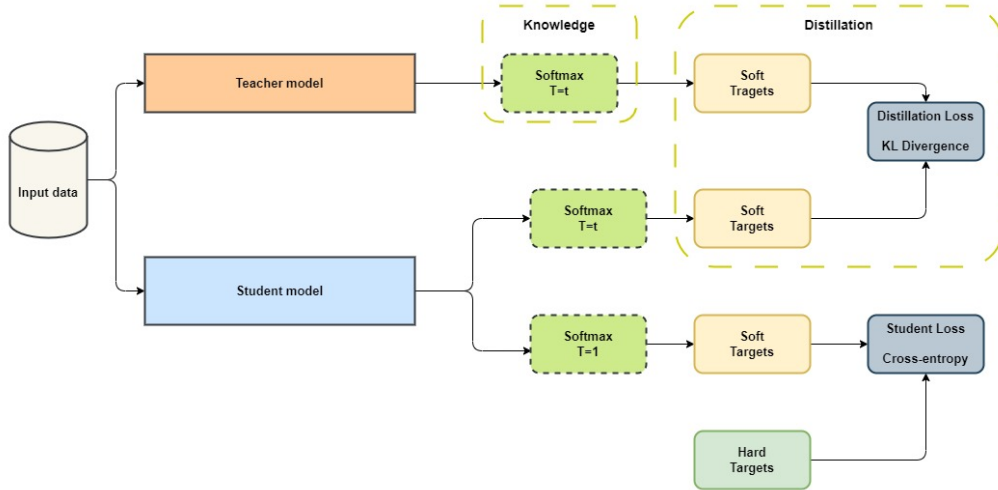


Figure 2.6: Example of the Response-base knowledge, based on [37].

Offline distillation is the most common way to transfer knowledge from a pre-trained network (teacher model) to a smaller network (student model). This technique was the one used by the authors [36]. This process is based on two stages: i) in the first stage, before distillation takes place, a large network is trained with the training dataset that will serve as the teacher model; ii) in the second stage, the same training dataset is used to train the student model and guide the distillation process, in which logits or intermediate features are transferred from the teacher model.

In the Online Distillation process [40, 41], a teacher model or a student model is not defined, which means that the networks may play the role of teacher or student. Therefore, it is described as an enterprise process that allows training the two architectures at the same time with the same dataset, making them learn from each other and update themselves during the process [37]. On the other hand, Self-distillation proposed by [42] consists in using the same network as teacher and student models. In this process, it can be used a pre-trained network that will use the knowledge of the deeper layers, such as features maps to distil to the lower layers such as Fully-connected.

Based on this context, for this work, we intend to distil the selected architectures ViTs and CNNs. Therefore, it was chosen the use of the Response-Based knowledge method to transfer the knowledge learned by the teacher model to the student model and the Offline distillation method to conduct the process of knowledge distillation between the two architectures. However, during the writing of this subchapter, it was discovered another methodology proposed by [43] to distil the knowledge from the teacher model to the student model regarding the ViTs architectures.

This methodology consists in distilling the knowledge through attention. Figure 2.7 exemplifies the authors' proposed methodology [43], wherein a distillation token is added before the first layer to incorporate other embeddings such as class and patches tokens. This distillation token consists of the label predicted by the teacher model, the

result returned by the last layer of the model, which will be assumed as a true label and will be computed along with the class and patches tokens in the self-attention layer. The self-attention layer is a mechanism that computes between a query vector and a vector of pairs (key, value) for output, and its result is scaled and normalized with the softmax function to obtain k weights. The output of this mechanism is a weighted sum of a set of k -value vectors [44].

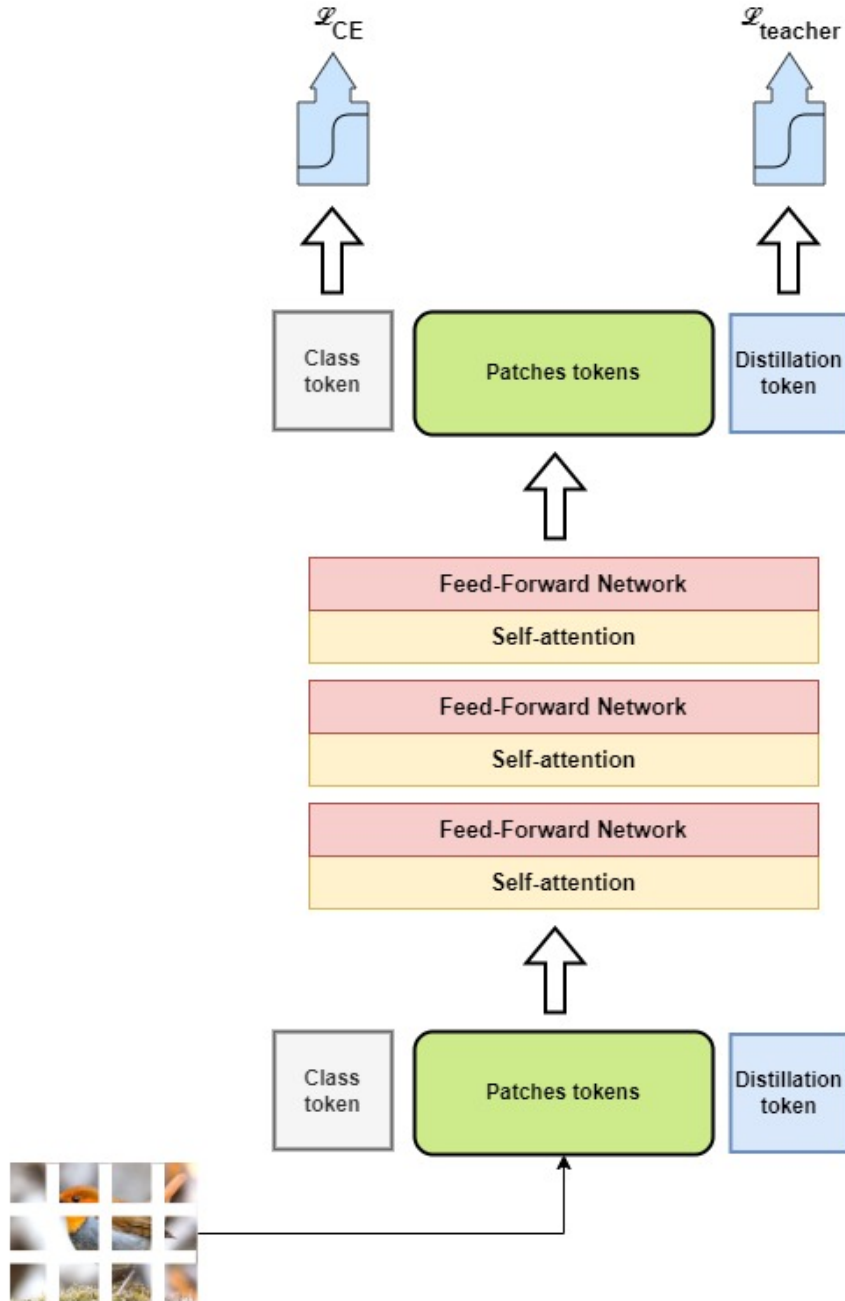


Figure 2.7: Example of the architecture of DeiT, based on [43].

In the next layer is added a Feed-Forward Network which consists of two linear layers, which are separated by a GeLU activation function. Its function is based on the first

layer making a linear projection of the patch dimensions from D to $4D$, and the second layer will make projections of the patch dimensions from $4D$ to D [43]. Considering D as the dimension of the tokens within the patches. The distillation proposed by [36] seeks to minimize the loss function KL Divergence between the teacher’s softmax and the student’s softmax and uses cross-entropy to minimize the loss between the hard labels and the student’s softmax. The authors [36] resort to cross-entropy to minimize the loss between the student and the hard labels, as well as between the softmax of the teacher and the student.

In the present study, the distillation of knowledge within ViTs is done using the two approaches and seeking to make a comparison between the two ways: (i) distillation through attention; (ii) distillation of the weights using the temperature to scale them. The goal is to find out which is the most efficient for the current application.

2.4 Interpretability

To do interpretability of the deep learning models, five model-agnostic techniques were experimented with: Grad-CAM, Shap values, LIME, RISE, and Occlusion sensitivity. Since they have been used for interpreting deep learning models by the other authors [45]. This sub-chapter presents a summary description of these techniques.

Grad-CAM was proposed by [46] and consists in using the Class Activation Mapping (CAM) technique to identify the most important part of the image that the classifier considered to predict the class of the image received as input. Therefore, this interpretability algorithm receives as input the feature map of the last convolutional layer and calculates its gradients that are grouped to obtain weightings. Higher weightings represent the greater importance of the feature maps for predicting an image in a particular class.

Shap values is a visualization tool that allows making deep learning and machine learning models more explainable by calculating the importance of each feature for the prediction made by the model [47]. This tool for deep learning model interpretability consists in previously calculating the shap values associated with each pixel in the image, providing the human being with information on how much each pixel contributed to classifying that image within a particular class [48].

Local Interpretable Model-Agnostic Explanations (LIME) is an algorithm that can explain the predictions of any classifier or regressor in a local way. In other words, what this model agnostic technique tries to do is to use the values of the input features and with some perturbations observe what the resulting impact is on the output, thus visually returning the contributions of each feature to the prediction of a sample [49]. In the case of the problem of this work, the LIME algorithm receives as input an image and based on the prediction made by the classifier returns the received image, as well

as a mask associated with the features that contributed to the prediction made. The features in the image that contribute positively to the prediction are marked green, and the features that contribute negatively to the prediction are marked red.

Vitali Petsiuk et. al [50] proposed a model designed as Randomized Input Sampling for Explanations (RISE) that consists in estimating a pixel importance map using gradients or other internal network state demonstrating how important each pixel is for the model's prediction. Generally, the importance of each pixel can be measured through a perturbation, such as reducing the pixel intensity to zero, blurring the region or adding noise to the image, and evaluating how this affects the classifier's decision. In the case of the RISE model, the importance of the pixels is estimated by darkening them in arbitrary combinations, decreasing their intensity to zero.

Occlusion sensitivity was proposed by [51] and aims to cause perturbations in the image by occluding pixels and observing the model output to understand the most important regions of the image, where the network was based to make the prediction. The perturbations are caused by replacing features of the image with an occlusion mask, which consists of a grey square.

3 LITERATURE REVIEW

For this work, two literature reviews were performed: the first consisted in understanding the differences between ViTs and CNNs, as well as in terms of performance, which architecture has been better in image processing; the second was based on reviewing the literature and understanding what has been done so far by other authors in the scope of diagnosis assisted by deep learning for inflammatory bowel disease. Therefore, this chapter will present the methodology used for the literature review and a summary of the information collected from the selected papers.

3.1 Vision Transformers vs Convolutional Neural Networks

Transformers have been the models of preference for Natural Language Processing (NLP) tasks. They offer scalability and computational efficiency, allowing to train models with more than 100 billion parameters, without causing saturation of the model performance. Therefore, inspired by the success of the transformers applied to NLP, it was proposed to use the same architecture with few modifications to do image classification [34]. The author's proposal is an architecture, called Vision Transformers (ViT), which consists of breaking the image into 2D patches and providing this linear sequence of patches as input to the model.

In contrast to this deep learning architecture, there is another very popular tool for processing large volumes of data called Convolutional Neural Networks (CNN). CNN is an architecture that consists of multiple layers, it typically contains convolutional layers, pooling layers, and fully connected layers [31]. CNN's have demonstrated good performance in big image classification datasets, computer vision tasks and NLP problems.

This literature review was published in [7] and the full paper can be found in Appendix A - Articles.

3.1.1 Research Methodology

Data Sources

ACM Digital Library, Google Scholar, Science Direct, Web of Science, and Scopus were chosen as the data sources to extract the Primary Studies. The number of results found after searching papers in each of the data sources is shown in Table 3.1.

Table 3.1: Data sources and number of results obtained for the literature review of ViTs vs CNNs.

Data source	Number of results	Number of selected papers
ACM Digital Library	19,179	1
Google Scholar	10,700	10
Science Direct	1,437	3
Scopus	55	2
Web of Science	90	1

Search String

In this work, ViTs and CNNs will be used. It is thus important to review their characteristics. To facilitate this, search strings tailored to identify relevant works that establish a comparison between these two architectures have been constructed. Table 3.2 provides a list of the search strings used in each electronic database.

Table 3.2: Data sources and search string used for the literature review of ViTs vs CNNs.

Data source	Search string
ACM Digital Library	((Vision Transformers) AND (convolutional neural networks) AND (images classification) AND (comparing))
Google Scholar	((ViT) AND (CNN) AND (Images Classification) OR (Comparing) OR (Vision Transformers) OR (convolutional neural networks) OR (differences))
Science Direct	((Vision Transformers) AND (convolutional neural networks) AND (images classification) AND (comparing))
Scopus	((ViT) AND (CNN) AND (comparing))
Web of Science	((ViT) AND (CNN) AND (comparing))

Inclusion Criteria

The inclusion criteria set to select the papers were that the studies were recent, had been written in English, and were published between January 2021 and December 2022. This choice of publication dates is based on the fact that ViTs were not proposed until the end of 2020 [34]. In addition, the studies had to demonstrate a comparison between CNNs and ViTs for image classification and could use any pre-trained model of the two architectures. Studies that presented a proposal for a hybrid architecture, where they combined the two architectures into one, were also considered. The dataset used during the studies did not have to be a specific one, but it had to be a dataset of images that allowed classification using both deep learning architectures.

Exclusion Criteria

Studies that oriented their research on using only one of the two deep learning architectures (i.e., Vision Transforms, or Convolutional Neural Networks) were excluded. Additionally, papers that were discovered to be redundant when searches were conducted throughout the chosen databases were eliminated. It was also defined that one of the exclusion criteria will be that the papers would have more than seven citations.

In summary, with the application of these criteria, 10,690 papers were excluded from the Google Scholar database, 89 papers from Web of Science, 53 papers from Scopus, 19,158 papers from ACM Digital Library, and 1434 papers from Science Direct.

Results

After applying the inclusion and exclusion criteria to the papers obtained in each of the electronic databases, 17 papers were selected for the literature review. Figure 3.1 shows a distribution of the selected papers by year of publication. Table 3.3 lists all the papers that were selected, as well as their year of publication and what type of publication they were.

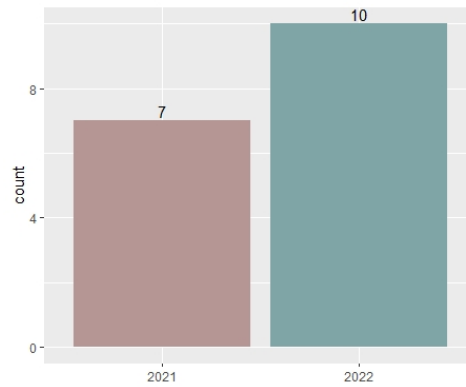


Figure 3.1: Distribution of the selected studies by years for the literature review of ViTs vs CNNs.

Table 3.3: List of selected studies for the literature review of ViTs vs CNNs.

Reference	Title	Year	Type
[52]	Adversarial Robustness Comparison of Vision Transformer and MLP-Mixer to CNNs	2021	Conference
[53]	Are Transformers More Robust Than CNNs?	2021	Conference
[54]	Detecting Pneumonia using Vision Transformer and comparing with other techniques	2021	Conference
[55]	Do Vision Transformers See Like Convolutional Neural Networks?	2021	Conference
[56]	Vision Transformer for Classification of Breast Ultrasound Images	2021	Conference
[57]	ConvNets vs. Transformers: Whose Visual Representations are More Transferable?	2021	Conference
[58]	A vision transformer for emphysema classification using CT images	2021	Journal
[59]	Comparing Vision Transformers and Convolutional Nets for Safety Critical Systems	2022	Conference
[60]	Convolutional Nets Versus Vision Transformers for Diabetic Foot Ulcer Classification	2022	Conference
[61]	Convolutional Neural Network (CNN) vs Vision Transformer (ViT) for Digital Holography	2022	Conference
[62]	Cross-Forgery Analysis of Vision Transformers and CNNs for Deepfake Image Detection	2022	Conference
[63]	Traffic Sign Recognition with Vision Transformers	2022	Conference
[64]	An improved transformer network for skin cancer classification	2022	Journal
[65]	CNN and transformer framework for insect pest classification	2022	Journal
[66]	Single-layer Vision Transformers for more accurate early exits with less overhead	2022	Journal
[67]	Vision transformer-based autonomous crack detection on asphalt and concrete surfaces	2022	Journal
[68]	Vision Transformers for Weeds and Crops Classification of High-Resolution UAV Images	2022	Journal

Figure 3.2 shows the distribution of the selected studies by application area. In the figure, most of the papers are generic in their application area. In these papers without a specific application area, the authors try to better understand the characteristics of the two architectures. For example, between CNNs and ViTs, the authors have tried to understand which of the architectures is more transferable. If architectures based on

transformers are more robust than CNNs. And if the ViT will be able to see the same information as CNN with a different architecture. Within the health domain, some studies have been developed in different sub-areas, such as breast cancer, to show that ViT can be better than CNNs. The figure also shows that some work has been done, albeit to a lesser extent, in other application areas. Agriculture stands out with two papers comparing ViTs with CNNs.

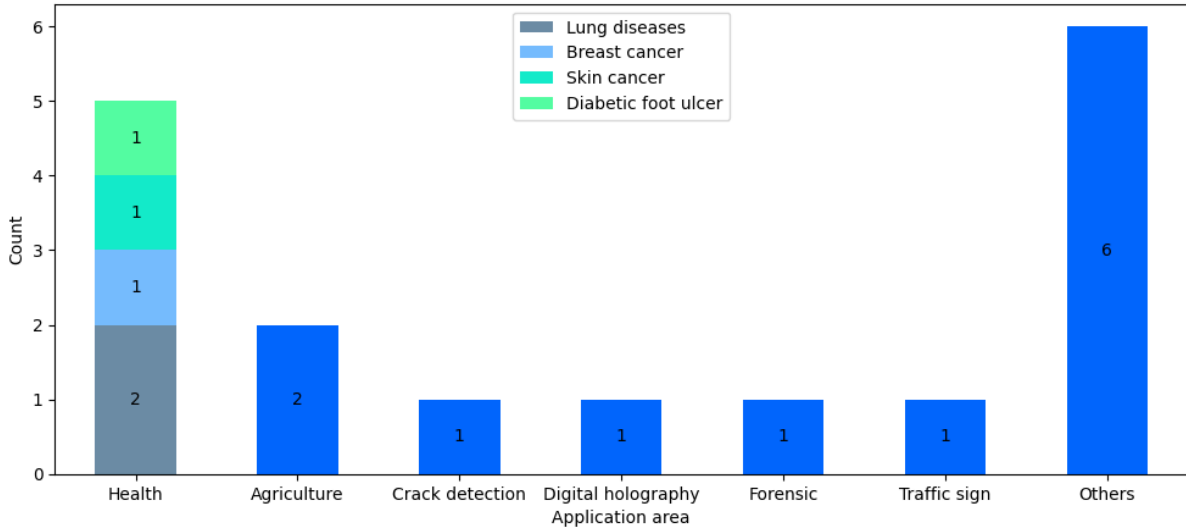


Figure 3.2: Distribution of the selected studies by application area.

3.1.2 Findings

An overview of the studies selected through the research methodology is shown in Table 3.7. This information summarizes the authors' approach, the findings, and other architectures that were used to build a comparative study. As a result, this chapter will provide a summary of the information included in the gathered papers.

In the study developed in [59], the authors aimed to compare the two architectures (i.e., ViT and CNN), as well as the creation of a hybrid model that corresponded to the combination of the two. The experiment was conducted using the ImageNet dataset and perturbations were applied to the dataset images. It was concluded that ViT can perform better and be more resilient on images with natural or adverse disturbances than CNN. It was also found in this work that the combination of the two architectures results in a 10% improvement in accuracy (Acc).

The work done in [61] aimed to compare Vision Transformers (ViT) with Convolutional Neural Networks (CNN) for digital holography, where the goal was to reconstruct amplitude and phase by extracting the distance of the object from the hologram. In this work, DenseNet201, DenseNet169, EfficientNetB4, EfficientNetB7, ViT-B/16, ViT-B32 and ViT-L/16 architectures were compared with a total of 3400 images.

They were divided into four datasets, original images with or without filters, and negative images with or without filters. The authors concluded that ViT despite having an accuracy like CNN, was more robust because, due to the self-attention mechanism, it can learn the entire hologram rather than a specific area.

The authors in [54] studied the performance of ViT in comparison with other architectures to detect pneumonia, through chest X-ray images. Therefore, a ViT model, a CNN network developed by the authors and the VGG-16 network were used for the study which focussed on a dataset with 5856 images. After the experiments performed, the authors concluded that ViT was better than CNN with 96.45% accuracy, 86.38% validation accuracy, 10.8% loss and 18.25% validation loss. In this work, it was highlighted that ViT has a self-attention mechanism that allows splitting the image into small patches that are trainable, and each part of the image can be given an importance. However, the attention mechanism as opposed to the convolutional layers makes ViT's performance saturate fast when the goal is scalability.

In the study [68] the goal was to compare ViT with state-of-art CNN networks to classify UAV images to monitor crops and weeds. The authors compared the influence of the size of the training dataset on the performance of the architectures and found that ViT performed better with fewer images than CNN networks in terms of F1-Score. They concluded that ViT-B/16 was the best model to do crop and weed monitoring. In comparison with CNN networks, ViT could better learn the patterns of images in small datasets due to the self-attention mechanism. In the scope of lung diseases, the authors in [11] investigated the performance of ViT models to automatically classify emphysema subtypes through Computed Tomography (CT) images in comparison with CNN networks. In this study, they performed a comparative study between the two architectures using a dataset collected by the authors (3192 patches) and a public dataset of 168 patches taken from 115 HRCT slides. In addition to this, they also verified the importance of pre-trained models. They concluded that ViT failed to generalize when trained with fewer images, because when comparing the pre-training accuracy with 91.27% on the training and 70.59% on the test.

In the work in [56], a comparison between state-of-the-art CNNs and ViT models for Breast ultrasound image classification was developed. The study was performed with two different datasets: the first containing 780 images and the second containing 163 images. The following architectures were selected for the study: ResNet50, VGG 16, Inception, NASNET, ViT-S/32, ViT-B/32, ViT-Ti/16, R + ViT-Ti/16 and R26 + ViT-S/16. ViT models were found to perform better than CNN networks for image classification. The authors also highlighted that ViT models could perform better when they were trained with a small dataset, because via the attention mechanism, it was possible to collect more information from different patches, instead of collecting information from the image.

Benz et al., in [52], compared ViT models, with the MLP-Mixer architecture and with CNNs. The goal was to evaluate which architecture was more robust in image classification. The study consisted of generating perturbations and adverse examples in the images and understanding which of the architectures was most robust. However, this study did not aim to analyse the causes. Therefore, the authors concluded that ViT were more robust than CNNs to adversarial attacks and from a features perspective CNN networks were more sensitive to high-frequency features. It was also described that the shift-variance property of convolutional layers may be at the origin of the lack of robustness of the network in the classification of images that have been transformed.

The authors in [62] performed an analysis between ViT and CNN models aimed at detecting deep fake images. The experiment consisted in using the ForgeryNet dataset with 2.9 million images and 220 thousand video clips, together with three different image manipulation techniques, where they tried to train the models with real and manipulated images. By training the ViT-B model and the EfficientNetV2 network the authors demonstrated that the CNN network could generalize better and obtain higher training accuracy. However, ViT could have better generalization, reducing the bias in the identification of anomalies introduced by one or more different techniques to introduce anomalies.

Chao Xin et al. [64] aimed to compare their ViT model with CNN networks and with another ViT model to perform image classification to detect skin cancer. The experiment conducted by the authors used a public HAM10000 dataset with dermatoscopic skin cancer images and a clinical dataset collected through dermoscopy. In this study, a multi-scale image and the overlapping sliding window were used to serialize the images. They also used contrastive learning to improve the similarity of different labels and minimize the similarity in the same label. Thus, the ViT model developed was better for skin cancer classification using these techniques. However, the authors also demonstrated the effectiveness of balancing the dataset on the model performance, but they did not present the F1-Score values before the dataset is balanced to verify the improvement.

The authors in [66] aimed to study if ViT models could be an alternative to CNNs in time-critical applications. That is, for edge computing instances and IoT networks, applications using deep learning models consume multiple computational resources. The experiment used pre-trained networks such as ResNet152, DenseNet201, InceptionV3, and SL-ViT with three different datasets in the scope of images, audio, and video. They concluded that the ViT model introduced less overhead and performed better than the architectures used. It was also shown that increasing the kernel size of convolutional layers and using dilated convolutional ones caused a reduction in the accuracy of a CNN network.

In a study carried out in [67], the authors tried to find in ViTs an alternative solution to

CNN networks for asphalt and concrete crack detection. The authors concluded that ViTs, due to the self-attention mechanism, had better performance in crack detection images with intense noise. CNN networks in the same images suffered from a high number of false negative rates, as well as the presence of biases in image classification.

Haolan Wang in [63] aimed to analyse eight different Vision Transformers and compare them with the performance of a pre-trained CNN network and without the pre-trained parameters to perform traffic signal recognition in autonomous driving systems. In this study, three different datasets with images of real-world traffic signals were used. This allowed the authors to conclude that the pre-trained DenseNet161 network had a higher accuracy than the ViT models to do traffic sign recognition. However, it was found in this work that ViT models performed better than the DenseNet161 network without the pre-trained parameters. From this work, it was also possible to conclude that the ViT models with a total number of parameters equal to or greater than the CNN networks, used during the experiment, had a shorter training time.

The work done in [60] compared CNN networks with Vision Transformers models for the classification of Diabetic Foot Ulcer images. For the study, the authors decided to use the following architectures: Big Image Transfer (BiT), EfficientNet, ViT-base and Data-efficient Image Transformers (DeiT) upon a dataset composed of 15,683 images. A further aim of this study is to compare the performance of deep learning models using Stochastic Gradient Descent (SGD) [69] with Sharpness-Aware Optimization (SAM) [70, 71]. These two tools are optimizers that seek to minimize the value of the loss function, improving the generalization ability of the model. However, SAM minimizes the value of the loss function and the sharpness loss, looking for parameters in the neighbourhood with a low loss. Therefore, this work concluded that the SAM optimizer originated an improvement in the values of F1-Score, AUC, Recall and Precision in all the architectures used. However, the authors did not present the training and test values that allow for evaluating the improvement in the generalization of the models. Therefore, the BiT-ResNetX50 model with the SAM optimizer obtained the best performance for the classification of Diabetic Foot Ulcer images with F1-Score = 57.71%, AUC = 87.68%, Recall = 61.88%, and Precision = 57.74%.

The authors in [65] performed a comparative study between ViT models and CNN networks used in the state of the art with a model developed by them, where they combined CNN and transformers to perform insect pest recognition to protect agriculture worldwide. This study involved three public datasets: the IP102 dataset, the D0 dataset and Li's dataset. The algorithm created by the authors consisted of using the sequence of inputs formed by the CNN feature maps to make the model more efficient, and a flexible attention-based classification head was implemented to use the spatial information. Comparing the results obtained, the proposed model obtained a better performance in insect pest recognition with an accuracy of 74.897%. This work

demonstrated that fine-tuning worked better on Vision Transformers than CNN, but on the other hand, this caused the number of parameters, the size, and the inference time of the model to increase significantly with respect to CNN networks. Through their experiments, the authors also demonstrated the advantage of using decoder layers in the proposed model to perform image classification. The greater the number of decoder layers, the greater the accuracy value of the model. However, this increase in the number of decoder layers increased the number of parameters, the size, and the inference time of the model. In other words, the architecture to process the images consumes far greater computational resources, which may not compensate for the increase in accuracy value with few layers. In the case of this study, the increase from one layer to three decoder layers represented only an increase of 0.478% in the accuracy value.

Several authors in [53, 55, 57] went deeper into the investigation and aimed to understand how the learning process of Vision Transformers works if ViT could be more transferable and better understand if the transform-based architecture were more robust than CNNs.

In this sense, the authors in [55] intended to analyse the internal representations of ViT and CNN structures in image classification benchmarks and found differences between them. One of the differences was that ViT has greater similarity between high and low layers, while CNN architecture needs more low layers to compute similar representations in smaller datasets. This is due to the self-attention layers implemented in ViT, which allows it to aggregate information from other spatial locations, vastly different from the fixed field sizes in CNNs. They also observed that ViTs in the lower, self-attention layers can access information from local heads (small distances) and global heads (large distances). Whereas CNNs have access to information locally in the lower layers.

On the other hand, the authors in [57] systematically analysed the transfer learning capacity in the two architectures. The study was conducted by comparing the performance of the two architectures on single-task and multi-task learning problems, using the ImageNet dataset. Through this study, the authors concluded that the transform-based architecture contained more transferable representations compared to CNNs for fine-tuning, presenting better performance and robustness in multi-task learning problems.

In another study carried out in [53], the goal was to prove if ViT were more robust than CNN as the most recent studies have shown. The authors developed their work comparing the robustness of the two architectures using two different types of perturbations: adversarial samples, which consist in evaluating the robustness of deep learning architectures in images with human-caused perturbations (i.e., data augmentation) and out-of-distribution samples, which consists in evaluating the robustness of the architectures in benchmarks of classification images. Through this experiment, it

was demonstrated that by replacing the activation function ReLU with the activation function of transformer-based architecture (i.e., GELU) the CNN network was more robust than ViT in adversarial samples. In this study, it was also demonstrated that CNN networks were more robust than ViT in patch-based attacks. However, the authors concluded that the self-attention mechanism was the key to the robustness of the transformer-based architecture in most of the experiments performed.

In conclusion, the ViT architecture is more robust than CNN networks for images that have noise or are augmented. It manages to perform better compared to CNN due to the self-attention mechanism because it makes the overall image information accessible from the highest to the lowest layers [59]. On the other hand, CNNs can generalize better with smaller datasets and get better accuracy than ViTs but, in contrast, ViTs have the advantage of better learning information with fewer images. This is because the images are divided into small patches, so there is a greater diversity of relationships between them.

3.2 Deep Learning to Detect Inflammatory Bowel Disease

In this chapter, a summary of the findings made by some related work within the scope of this study is presented. Over the years, the use of computer vision tools in the diagnosis of inflammatory bowel disease has become essential. Since the evaluation of images obtained through imaging examinations is done by professional gastroenterologists, the criteria used to determine the patient's diagnosis are subjective and may differ among doctors [72]. This process becomes a consequence for the patient as it results in a delay in prescribing an effective treatment to combat the disease.

3.2.1 Research Methodology

Data Sources

PubMed and Google Scholar were chosen as the data sources to extract the Primary Studies. The number of results found after searching papers in each of the data sources is shown in Table 3.4.

Table 3.4: Data sources and the number of results obtained for the literature review of IBD.

Date query	Data source	Number of results	Number of selected papers
12/09/2022	Google Scholar	40 010	19
11/10/2022	PubMed	3	0
16/10/2022	Google Scholar	72	22
16/10/2022	PubMed	18 000	7

Search String

Based on the aim of this study, some search strings, tuned for each selected data source, were developed. The search was also made by title. Table 3.5 provides a list of the search strings and titles used in each electronic database.

Table 3.5: Data sources and search string or title used for the literature review of IBD.

Date query	Data source	Search string/Title
12/09/2022	Google Scholar	Deep Learning to diagnosis Ulcerative Colitis and Crohn disease using endoscopy images
11/10/2022	PubMed	Distinguish between Crohn Disease and Ulcerative colitis with deep Learning
16/10/2022	Google Scholar	((Inflammatory Bowel Disease) AND (Deep Learning) OR (endoscopy) OR (colonoscopy))
16/10/2022	PubMed	((Inflammatory Bowel Disease) AND (Deep Learning))

Inclusion Criteria

The inclusion criteria set to select the articles was that studies had to be published between January 2016 and December 2022. In addition to this, the studies had to address the use of deep learning to distinguish inflammatory bowel diseases. Whenever this was not possible, at least they had to apply deep learning to classify endoscopic or colonoscopy images of the disease.

Exclusion Criteria

Studies that directed their research toward using Deep Learning to classify other bowel diseases were excluded as polyps, haemorrhoids, etc. Also, were excluded articles that were duplicated with those that had already been collected. Papers that oriented the investigation to processing videos collected during colonoscopy or endoscopy were also discarded.

Results

Accounting for all the papers selected on each of the searches performed, forty-eight articles were collected. In the first approach, seven papers were discarded because they were repeated, and seven more papers were discarded because they had few citations. Then, by reading the abstract, seven more papers were discarded because they were not within the scope of the study of this work. In total, thirteen papers were counted for the literature review. Figure 3.3 shows the distribution of papers by year of publication. Table 3.6 lists all the papers that were selected for this literature review.

3.2.2 Findings

Table 3.8 presents an overview of the results obtained and the architectures used by the literature in the field of inflammatory bowel disease diagnosis. Throughout this chapter, a summary of the findings made and the experiments developed by the authors of the papers selected for this literature review will be exposed.

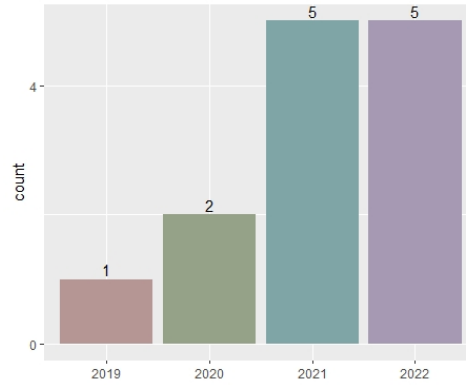


Figure 3.3: Distribution of the selected papers for the literature review of IBD by years.

Table 3.6: List of selected studies for the literature review of IBD

Reference	Title	Year	Type
[72]	Performance of a Deep Learning Model vs Human Reviewers in Grading Endoscopic Disease Severity of Patients With Ulcerative Colitis	2019	Journal
[73]	Deep learning algorithms for automated detection of Crohn's disease ulcers by video capsule endoscopy	2020	Journal
[74]	Development and Validation of a Deep Neural Network for Accurate Evaluation of Endoscopic Images From Patients With Ulcerative Colitis	2020	Journal
[75]	A deep learning framework for autonomous detection and classification of Crohn's disease lesions in the small bowel and colon with capsule endoscopy	2021	Journal
[76]	Accurate diagnosis of endoscopic mucosal healing in ulcerative colitis using deep learning and machine learning	2021	Journal
[77]	Automated Detection of Crohn's Disease Intestinal Strictures on Capsule Endoscopy Images Using Deep Neural Networks	2021	Journal
[78]	Deep Learning Algorithm for the Confirmation of Mucosal Healing in Crohn's Disease, Based on Confocal Laser Endomicroscopy Images	2021	Journal
[79]	Deep learning and capsule endoscopy: automatic identification and differentiation of small bowel lesions with distinct haemorrhagic potential using a convolutional neural network	2021	Journal
[80]	Automatically detecting Crohn's disease and Ulcerative Colitis from endoscopic imaging	2022	Conference
[81]	Artificial intelligence enabled automated diagnosis and grading of ulcerative colitis endoscopy images	2022	Journal
[82]	Development and Validation of a Deep Neural Network for Accurate Identification of Endoscopic Images From Patients With Ulcerative Colitis and Crohn's Disease	2022	Journal
[83]	Development of a Convolutional Neural Network-Based Colonoscopy Image Assessment Model for Differentiating Crohn's Disease and Ulcerative Colitis	2022	Journal
[84]	Diagnosis of ulcerative colitis from endoscopic images based on deep learning	2022	Journal

The authors in [78] conducted their study to use deep learning algorithms to distinguish between active and non-active inflammation of Crohn's disease in patients. For the study, the authors collected 6,205 images from 54 patients, which are divided into 32 patients with active Crohn's disease and 22 patients who form a control group. However, the authors in the control group enclose images of patients who do not have any inflammation and patients who have had inflammation. In this study, a CNN+LSTM was combined and a CNN was used to make the automatic diagnosis. The combination of a CNN+LSTM is the best model to do this with 95.36% of Acc, 94.6% of sensitivity and 92.78% of specificity. However, it is not shown which CNN was used.

In the study developed in [84], the authors sought to classify whether ulcerative colitis disease was in remission or not. They also wanted to automatize the classification of the severity grade of UC based on Mayo scores 1-3. Mayo 0 images were considered healthy patients, which is not true because patients with Mayo 0 have already been

diagnosed with ulcerative colitis disease, but at that moment the disease is in remission/inactive. The authors in this study used colonoscopy images collected from the First Affiliated Hospital of Kunming Medical University. They propose an architecture that combines a CNN and an RNN, called UC-DenseNet, and compare the results with other architectures: VGG-19, ResNet152, InceptionV3 and DenseNet201. The proposed architecture stands out obtaining 97.6% of Acc, 97.5% of Pre, 98.0% of recall, 97.6% of F1-Score and 97.5% of AUC for the classification of the existence of remission or not. An Acc=90.6%, Pre=87.4%, recall=86.3%, and F1-Score=86.8% were obtained for the grade of disease severity.

The literature in [83] collected 57,597 colonoscopy images about Crohn's disease, ulcerative colitis, and from healthy patients. However, this dataset is not publicly available. The authors aimed to use 13,872 images to train the CNN ResNetXt-101 network and hold 1,452 images for testing. In which, they aimed to compare with the analysis done by six experienced clinicians. The comparison consisted of analysis per image of the classes, but also per patient to evaluate the diagnosis made. Based on the results obtained, the CNN network for the classification of CD images obtained 92.39% of Acc, 87.53% of sensitivity, 94.78% of specificity, 89.19% of PPV and 88% of F1-Score. For UC it obtained: 93.35% of Acc, 90.49% of sensitivity, 98.14% of specificity, 89.94% of PPV, 95.11% of NPV and 90% of F1-Score. They concluded that the CNN network was better than clinicians at classifying images. On the other hand, clinicians were able to be better most of the time in per-patient analysis.

The study in [75] aimed to detect Crohn's disease lesions caused by the different degrees of severity located in the small intestine and colon. The authors collected 7,744 images from 38 patients, from three centres in southeastern Denmark. They used the ResNet50 network in five different image pre-processing (e.g., Original, Contrast increase, Histogram equalization, GradientX and Dephezing), this pre-processing consisted of modifying the texture of the images to improve the performance of the network. To evaluate the performance of the network, the authors divided the dataset by patient and random which consisted of 70% for training, 10% for validation, and 20% for testing. They concluded that overall for the per-patient split, the CNN obtained 98.30% of accuracy, 95.72% of sensitivity, and 99.77% of specificity. In the random split, it obtained 98.58% of Acc, 96.21% of sensitivity and 100% of specificity.

Ryan W. Stidham et. al [72] aimed to compare the performance of a deep learning model with 10 human reviewers for the classification of the severity grade of ulcerative colitis disease. For the study, 16,514 images were used as the Mayo score 0-3. The dataset was divided into 90% for creating the model and 10% for testing, of the 90%, 80% was used for training and 10% for fine-tuning. The authors demonstrated that the InceptionV3 network performed very similarly to the reviewers when comparing the values of the Kappa statistics (k) metric. Where the network obtained $k=0.84$ and the

experienced reviewers obtained $k=0.86$. Also, they concluded that the model demonstrated a good ability to distinguish endoscopic remission (Mayo 0-1) from moderate-to-severe (Mayo 2-3) with an AUROC of 0.966, a PPV of 0.87, a sensitivity of 83.0%, a specificity of 96.0% and an NPV of 0.94.

The authors in [77] aimed to automate the detection of intestinal strictures in Crohn's disease. In this study, the authors also used images of normal mucosal and mucosal ulcers. In total, the dataset contained 27,892 images gathered by capsule endoscopy. To evaluate the performance of the network the authors split the dataset using the 10-fold cross-validation method with seven patients for training and three for testing. To do the classification and imaging the authors selected the EfficientNet-B5 network. It is shown in this study that the network can obtain an average accuracy of 93.5% for the classification of images of strictures and non-strictures. The authors further concluded that the network has an excellent differentiation between strictures and normal mucosa (AUC=0.989), strictures and all ulcers (AUC=0.942), and between strictures and different grades of ulcers (mild, moderate, and severe ulcers) with an AUC of 0.992, 0.975 and 0.829, respectively.

The study in [73] sought to gather images by a video capsule endoscopy to detect ulcers in Crohn's disease. Therefore, the authors collected a dataset with 17,640 images from 49 patients and performed two experiments: i) they divided the dataset into five subsets of equal size where 80% of the patients were for training and 20% for testing; ii) they divided the dataset into N-1 patients for training and 1 unseen patient for testing, this process was randomized during 10 times. To classify the images they used a CNN Xception network. In summary, the CNN network in the first experiment obtained on average AUC=0.991, Acc=0.9612, sensitivity=0.9498, specificity=0.9698, PPV=0.958 and NPV=0.9642. In the second experiment, it averaged AUC=0.974, Acc=0.9066, sensitivity=0.8882, specificity=0.9077, PPV=0.9151 and NPV=0.9003.

Kento Takenak et. al [74] aimed to develop and evaluate a DNN to determine endoscopic remission and predict histologic remission in ulcerative colitis using colonoscopy images. For the study, the authors used 40,758 colonoscopy images and 6,885 biopsy results. Although the authors mention in the article that they have developed a DNN they do not describe what the structure of the same. However, the authors present that the developed network can determine endoscopic remission with 90.1% of Acc, 93.3% of sensitivity, 87.8% of specificity, 84.3% of PPV, 94.9% of NPV and 79.8% of Kappa coefficient. But it was better at determining histologic remission with 92.9% of Acc, 92.4% of sensitivity, 93.5% of specificity, 93.7% of PPV, 92.2% of NPV and 85.9% of Kappa coefficient.

The authors in [76] sought to diagnose endoscopic mucosal healing in ulcerative colitis using Deep Learning and machine learning algorithms. In this study, 856 endoscopic images with MES 0-3 were used and two experiments were performed: i) classify MES

0-1 vs MES 2-3 images to detect the presence or absence of mucosal healing; ii) classify MES 0 vs MES 1 images to determine complete or incomplete healing. The workflow designed by the authors was to do feature extraction with the InceptionV3 network and then classify with a DNN (unrevealed), SVM and K-NN classifier. The final results were obtained by combining the results of the three classifiers using ensemble learning (voting). However, the concrete way in which the voting is performed is not specified.

The study developed in [79] consisted in using a CNN to identify and differentiate multiple lesions in the small intestine with distinct bleeding potentials. The authors used a dataset of 53,555 images, which they divided into 80% for training and 20% for validation to train the Xception network. Based on Saurin's classification the images were divided into three categories: no bleeding potential - P0; uncertain/intermediate bleeding potential - P1; high bleeding potential -P2. In the end, the authors concluded that the CNN was able to differentiate multiple small bowel abnormalities and their respective bleeding potential classification with 99% of Acc, 88% of sensitivity, 99% of specificity, 87% of PPV and 99% of NPV.

The study in [82] sought to collect colonoscopy images from five hospitals in China to identify ulcerative colitis, Crohn's disease or healthy bowel. The authors were able to gather 49,154 colonoscopy images and also during the condition compared the performance of the model with 10 endoscopists. These images were only gathered for this study and are not publicly available. The team of human reviewers consisted of 5 reviewers with little experience and 5 very experienced reviewers. For the classification of the images, the authors used the ResNet50 network. When comparing the results of the deep learning model with the 10 reviewers, the authors showed that the model was on average faster at reading at 6.00 seconds and the reviewers with 2,425.00 seconds. On a per-patient analysis, the model was better than the reviewers with 0.991 of Acc versus 0.922 for the highly experienced reviewers and 0.780 for the less experienced reviewers. And in a per-lesion analysis, it was also superior with 0.904 of Acc against 0.597 for the less experienced reviewers and 0.699 for the very experienced reviewers.

The literature in [81] used the HyperKvasir dataset to obtain 840 images of ulcerative colitis disease, to diagnose the existence of the disease by classifying it with the others in the gastrointestinal tract. And to identify the degree of severity of ulcerative colitis. In this study, the authors used four convolutional networks: ResNet50, VGG-19, InceptionV3, and DenseNet121. Based on the experiments performed the authors concluded that the four networks showed better performance in diagnosing ulcerative colitis and non-ulcerative colitis, where the VGG-19 network obtained 98.49% of Acc, 98.61% of sensitivity, 98.24% of specificity, 97.66% of F1-Score and 99.88% of AUC. On the other hand, the networks failed to achieve a good performance in grade severity classification, where DenseNet121 obtained 87.50% of Acc, 79.00% of sensitivity, 91.00% of specificity, 91.29% of F1-Score and 90% of AUC.

The authors in the study [80] set a goal for their study to detect Crohn's disease from ulcerative colitis on endoscopic images. Throughout the study, three classification tasks were performed: i) distinguish the existence and non-existence of IBD; ii) distinguish Crohn's disease from ulcerative colitis; iii) distinguish ulcerative colitis from no ulcerative colitis. The authors used the SI-cura dataset, where they collected 4,388 images of ulcerative colitis, 5,948 images of Crohn's disease, 1,067 images of IBD, and 2,822 of no pathology. The SI-cura project is an Italian initiative that seeks to develop solutions based on artificial intelligence to distinguish pathologies with different natures. This dataset was used for this project and was not made available. The networks ResNet18, ResNet34, ResNet50, ResNet101 and ResNet152 were chosen to classify the images. They also created an ensemble method that combines ResNet34, with ResNet50 and ResNet101. They concluded that the ensemble method was the best in diagnosing the existence or non-existence of IBD pathology with 0.940 of MCC, 1.00 of TNR, 0.968 of TPR, 0.912 of NPV and 1.00 of PPV.

Based on the findings made by other authors, we seek to stand out by proposing to pool the images by the type of inflammatory bowel disease to distinguish between Crohn's disease and Ulcerative colitis, comparing ViTs with CNNs for processing the collected endoscopic images. A methodology that includes image pre-processing techniques for removing misleading features from endoscopic exam images is also suggested. Moreover, the present study also includes interpretability techniques to understand what regions of the mucosal the classifier identifies as the presence of the disease.

Combined with this, we seek through knowledge distillation to build more lightweight architectures, possible to deploy in endoscopic systems.

Table 3.7: Overview of studies selected for the literature review of ViTs vs CNNs.

Ref.	Dataset	Images Size	Number of Classes	Evaluated Architectures	Best Architecture	Best Results
[52]	ImageNet-1K (more than 1.431 M images) for training and ImageNet-C for validation	224x224px	2	ViT-B/16, ViT-L/16, Mixer-B/16, Mixer-B/16, RN18 (SWSL), RN50 (SWSL), RN18 and RN50	ViT-L/16	82.89% of Acc
[53]	ImageNet-A, ImageNet-C and Stylized ImageNet	224x224px	N/A	ResNet50 and DeiT-S	N/A	N/A
[54]	5856 images collected by X-ray	250x250px to ViT and 224x224px to CNN	2	ViT, CNN and VGG16	ViT	96.45% Acc, 86.38% val. Acc, 10.92% loss and 18.25% val. Loss
[55]	ImageNetILSVRC2012 (1.78 M images)	224x224px	1,000	ViT-B/32, ViT-L/16, ViT-H/14, ResNet50 and ResNet152	N/A	N/A
[56]	Public dataset 1 [85] with 780 images; Public dataset 2 [86] with 163 images	224x224px	3	ViT-S/32, ViT-B/32, ViT-Ti/16, R26 + S/16, R + Ti/16, VGG, Inception and NASNET	ViT-B/32	86.7% Acc and 95% AUC
[57]	Flower 102 (4,080 to 11,016 images); CUB 200 (11,788 images); Indoor 67 (15,620 images); NY Depth V2 (1,449 images); WikiArt; COVID-19 Image Data Collection (700 images); Caltech101 (9,146 images); FG-NET (1,002 images)	384 × 384px; 224 × 224px; 300 × 300px	40 to 102	R-101 × 3, R-152 and Swim-B × 4, ViT-B/16, ViT-L/16 and Swim-B	N/A	N/A
[58]	3192 images collected by CT and 160 images of a public dataset with CT biomarkers	61x61px	4	AlexNet, VGG-16, InceptionV3, MobileNetV2, ResNet34, ResNet50 and ViT	ViT	95.95% of Acc
[59]	ImageNet-C benchmark	224x224px	2	ViT-L/16, CNN, hybrid model (BiT-M + ResNet152) × 4	Hybrid model	99.20% of Acc
[60]	Dataset provided in DFUC 2021 challenge (15,683 images)	224x224px	4	EfficientNetB3, BiT-ResNeXt50, ViT-B/16 and DeiT-S/16	BiT-ResNeXt50	88.49% AUC, 61.53% F1-Score, 65.59% recall and 60.53% precision
[61]	3400 images collected by holographic camera	512x512px	10	EfficientnetB7, Densenet169 and ViT-B/16	ViT-B/16	99% of Acc
[62]	ForgeryNet with 2.9 M images	N/A	2	EfficientNetV2 and ViT-B	N/A	N/A
[63]	German dataset (51,830 images); Indian dataset (1976 images); Chinese dataset (18,168 images)	128x128px	15, 43 and 103	DenseNet161, ViT, DeepViT, MLP-Mixer, CvT, PiT, CarT, CCT, CrossViT and Twins-SVT	CCT	99.04% of Acc
[64]	HAM10000 dataset (10,015 images); 1016 images collected by dermatology	224x224px	3	MobileNetV2, ResNet50, InceptionV2, ViT and Proposed ViT model	Proposed ViT model	94.10% Acc, 94.10% precision and 94.10% F1-Score
[65]	IP102 dataset (75,222 images); D0 dataset (4508 images); Li's dataset (5629 images)	224x224px; 480x480	10 to 102	ResNet, EfficientNetB0, EfficientB1, RepVGG, VGG-16, ViT-L/16 and proposed hybrid model	Proposed hybrid model	99.47% Acc on the D0 dataset and 97.94% Acc on the Li's dataset
[66]	CIFAR-10 and CIFAR-100 (6000 images); Speech commands (100,503 1-second audio clips); GTZAM (1,00,030-second audio clips); DISCO (1935 images)	224x224px; 1024x576px. Spectrograms: 229x229 samples; 512x256 samples	10 to 100	SL-ViT, ResNet152, DenseNet201 and InceptionV3	SL-ViT	71.89% of Acc
[67]	CrackTree260 (260 images); Ozegetel (458 images); Lab's on dataset (80,000 images)	256x256px; 448x448px	2	TransUNet, U-Net, DeepLabv3+ and CNN + ViT	CNN + ViT	99.55% of Acc and 99.57% of precision
[68]	10,265 images collected by Pilgrim technologies UAV with Sony ILCE-7R-36 megapixels	64x64px	5	ViT-B/16, ViT-B/32, EfficientNetB, EfficientNetB1 and ResNet50	ViT-B/16	99.8% of Acc

Table 3.8: Overview of studies selected for the literature review of IBD.

Ref.	Objective	Images Size	Number of Images	Evaluated Architectures	Best Architecture	Best Results
[72]	Analyze if deep learning models are capable of grading the endoscopic severity of UC at a level comparable to that of experienced human reviewers.	320x256px	16,514 images	InceptionV3	InceptionV3	0.966 of AUROC, 0.87 of PPV, 83.0% of sensitivity and 96.0% of specificity
[73]	Create and test a deep learning method for automatically detecting small intestinal ulcers in Crohn's patients.	299x299px	17,640 images	Xception	Xception	0.991 of AUC, 0.9612 of Acc, 0.9498 of sensitivity, 0.9698 of specificity, 0.958 of PPV and 0.9642 of NPV
[74]	Create a system using deep neural networks to analyze endoscopic pictures from ulcerative colitis patients in a reliable, objective, and in real-time.	N/A	40,758 images	DNN	DNN	92.9% of Acc, 92.4% of sensitivity, 93.5% of specificity, 93.7% of PPV, 92.2% of NPV, 85.9% of Kappa coefficient
[75]	Detect Crohn's disease lesions caused by the different degrees of severity located in the small intestine and colon.	N/A	7,744 images	ResNet50	ResNet50	98.58% of Acc, 96.21% of sensitivity, 100% of specificity
[76]	Building a computer-aided diagnostic system that reliably detects mucosal healing in UC patients utilizing deep learning and machine learning.	299x299px	856 images	CNN, SVM, K-NN and Ensemble learning (voting)	Ensemble learning (voting)	0.9453 of Acc, 0.9278 of AUC, 0.8923 of sensitivity, 0.9633 of specificity, 0.8923 of Pre and 0.8923 of F1-Score
[77]	Automated the detection of intestinal stenosis in Crohn's disease.	456x456px	27,892 images	EfficientNet-B5	EfficientNet-B5	90.1% of accuracy and 0.965 of AUC
[78]	Use deep learning algorithms to distinguish between active and non-active inflammation in Crohn's patients.	300x300px	6,205 images	CNN and CNN+LSTM	CNN+LSTM	95.3% of Acc, 94.6% of sensitivity, 92.78% of specificity and 93% of precision-recall
[79]	Create a CNN-based model utilizing pictures from a capsule endoscopy to identify and distinguish between several small intestinal lesions with varied haemorrhagic potential.	N/A	53,555 images	Xception	Xception	99% of Acc, 88% of sensitivity, 99% of specificity, 87% of PPV and 99% of NPV
[80]	Development of artificial intelligence programs to distinguish between various diseases, such as inflammatory bowel disease, namely ulcerative colitis and Crohn's disease, based on endoscopic imaging of patients and healthy controls.	700x700px	14,193 images	ResNet18, ResNet34, ResNet50, ResNet101, ResNet152 and Ensemble method	Ensemble method	0.940 of MCC, 1.000 of TNR, 0.968 of TPR, 0.912 of NPV and 1.000 of PPV
[81]	Endoscopic evaluation to reliably grade disease activity, and detect ulcerative colitis and non-ulcerative colitis in patients.	299x299px	2,642 images and 840 images	VGG19, DenseNet121, ResNet50 and InceptionV3	VGG19	98.49% of Acc, 98.61% of sensitivity, 98.24% of specificity, 97.66% of F1-Score and 0.9988 of AUC
[82]	Developing and validating a deep learning diagnostic system to identify between UC and CD using a large number of colonoscopy images.	N/A	49,154 images	ResNet50	ResNet50	99.1% of Acc
[83]	Classify images about Crohn's disease, ulcerative colitis, and from healthy patients to compare with the analysis done by six experienced clinicians.	256x256px	57,597 images	ResNetXt-101	ResNetXt-101	98.35% of Acc, 98.14% of sensitivity, 98.46% of specificity, 96.93% of PPV, 99.07% of NPV and 0.98 of F1-Score
[84]	Classify whether ulcerative colitis disease was in remission or not. And make the classification of the severity grade of UC based on Mayo score 1-3.	256x256px	Dataset A: 9,928 images; Dataset B: 4,378 images	VGG19, ResNet152, InceptionV3, DenseNet201 and UC-DenseNet	UC-DenseNet	0.989 of Acc, 0.989 of Pre, 0.986 of Recall, 0.989 of F1-Score and 0.988 of AUC

4 METHODOLOGY AND DATA

In this chapter, the description of the created datasets will be presented, as well as a summary description of the datasets that were found referring to the two types of inflammatory bowel disease. This chapter will also describe the methodology implemented in this study and the algorithms used to classify the images, as well as the interpretability algorithms used to interpret the output of the deep learning models.

In order to improve the diagnosis of inflammatory bowel disease, a methodology based on eight phases is proposed (Figure 4.1). In the first phase, the authors collected images of the two inflammatory bowel diseases (i.e. Ulcerative colitis and Crohn's disease). In the second phase, frames were extracted from the videos collected during an endoscopic examination performed on the patients.

In the third phase, a pre-processing was applied to the images. In the fourth phase, data augmentation was performed on the training dataset. In the fifth phase, the convolutional networks and vision transformers were implemented and configured. In the sixth phase, the performance of the CNNs and ViTS was evaluated using the classification metrics: Accuracy, Loss, Precision, Recall, F1-Score, Area Under Curve, Mathew's Correlation Coefficient, as well as the Inference time. In the seventh phase, knowledge distillation of the deep learning models was attempted. Finally, the interpretability of the models, as well as their students resulting from knowledge distillation, was implemented. This methodology is illustrated in Figure 4.1 and will be further explained in the following chapters.

4.1 Description of dataset

This study aims to classify endoscopic images in order to distinguish between inflammatory bowel diseases. In this sense, public datasets with images of Ulcerative colitis and Crohn's disease were considered. Kvasir and HyperKvasir datasets contain images related to bowel diseases such as polyps, ulcerative colitis, haemorrhoids, etc. LIMUC dataset contains images related only to ulcerative colitis. CrohnIPI (a dataset available upon request) contains images related to Crohn's disease.

The Kvasir dataset was collected using endoscopic equipment at Vestre Viken Health Trust in Norway. Which consists of 4 hospitals and provides healthcare to 470,000 people [87]. This dataset consists of 4 000 images labelled and verified by medical experts from Vestre Viken Health Trust (VV) and the Cancer Registry of Norway (CRN). CRN

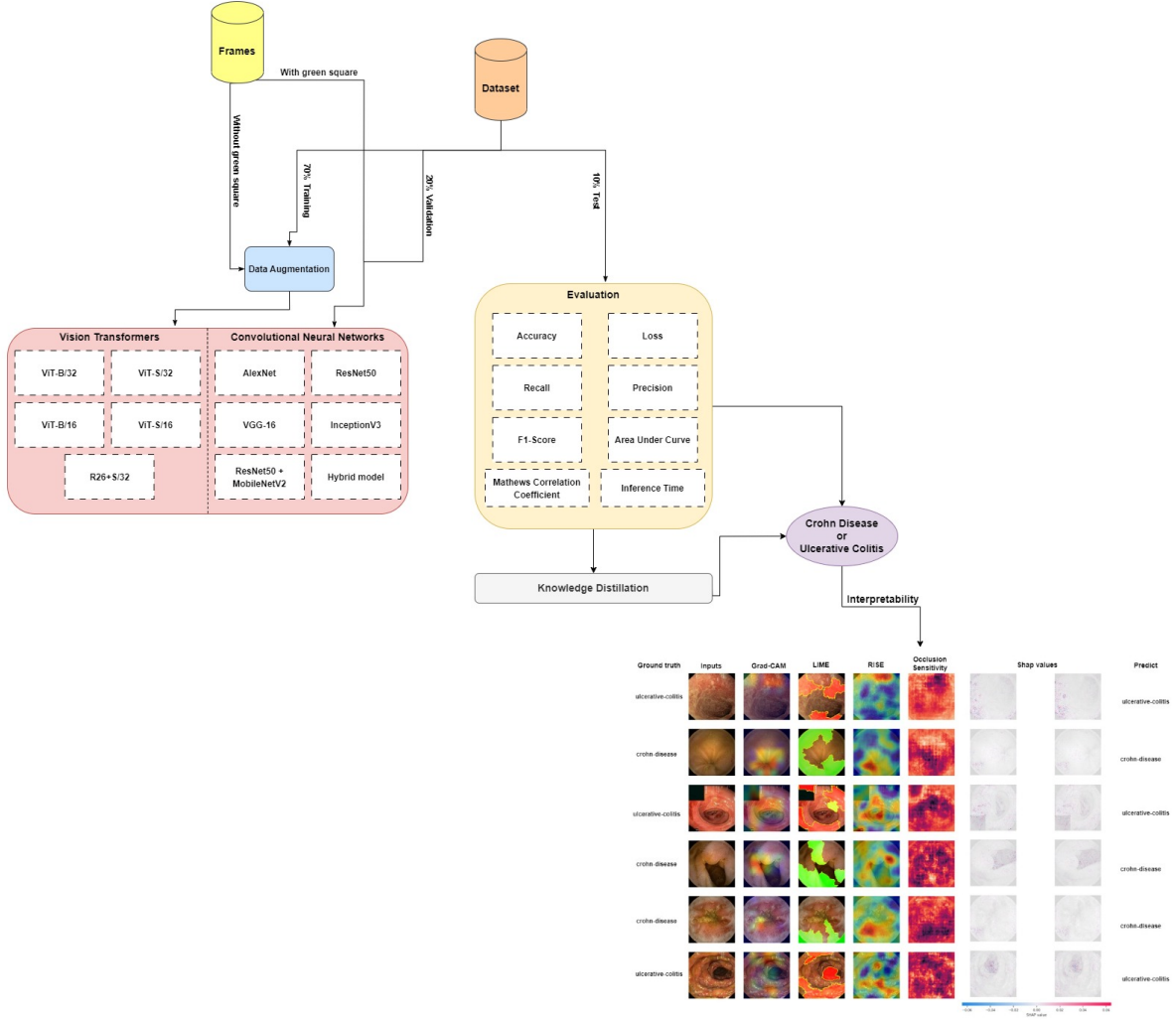


Figure 4.1: Experimental setup.

is recognised as an independent institution under the custody of Oslo University Hospital Trust and is part of the South East Norway Regional Health Authority. It is responsible for the national screening programmes to detect cancer or precancerous lesions early with the aim of preventing the number of cancer deaths [87].

The eight classes present in the dataset depict anatomical landmarks, pathological findings and endoscopic procedures in the gastrointestinal tract. The anatomical landmarks include Z-line, pylorus, and cecum; the pathological findings include esophagitis, polyps, and ulcerative colitis, it further includes images on the removal of dyed and lifted polyp lesions, and dyed resection margins. The images have different resolutions between 720x576 and 1920x1072 pixels. Some images have a green figure that represents the position and configuration of the endoscope inside the bowel [87]. In Figure 4.2 it is possible to observe some images referring to the classes present in the Kvasir dataset.

HyperKvasir is considered the largest multi-class imaging and video dataset of the gas-

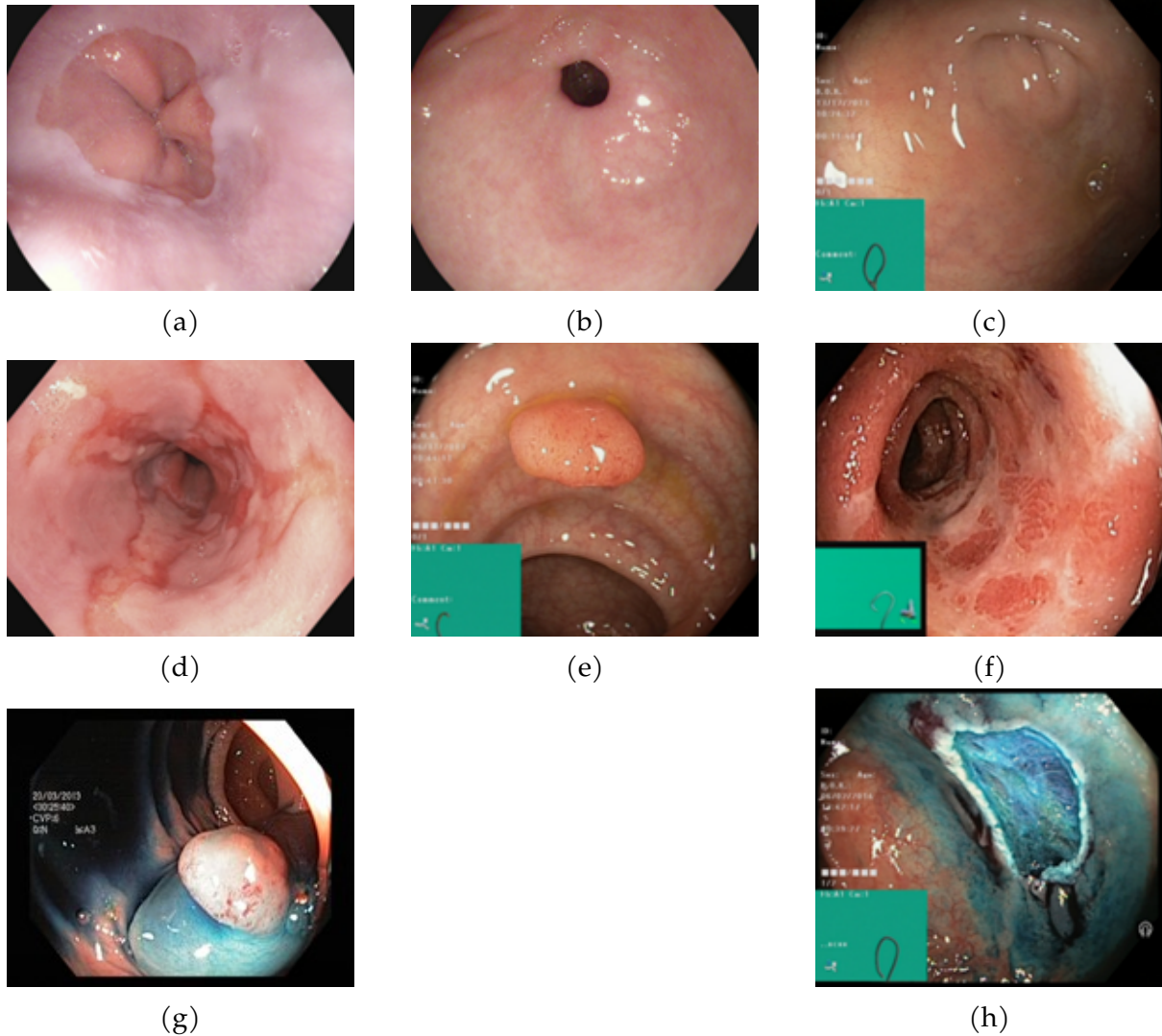


Figure 4.2: Images regarding the 8 classes present on the Kvasir dataset: (a) Z-line, (b) Pylorus, (c) Cecum, (d) Esophagitis, (e) Polyps, (f) Ulcerative Colitis, (g) Dyed and Lifted Polyps and (h) Dyed Resection Margins.

gastrointestinal tract today. The images were collected during real gastro and colonoscopy examinations, using standard endoscopic equipment from Olympus (Olympus Europe, Germany) and Pentax (Olympus Europe, Germany), at Baerum Hospital in Norway between 2008 and 2016. It was labelled by experienced gastrointestinal endoscopists. This dataset has 110 079 images and 374 videos, which represent anatomical landmarks, and pathological and normal findings [88, 89]. The dataset is divided into four different parts: labelled images, unlabelled images, segmented images and labelled videos. The labelled images part consists of a total of 10 662 images in JPEG format, which represent twenty-three classes. For this type of image, four main categories of findings were defined:

- **Anatomical landmarks:** they refer to features of the gastrointestinal tract that is used as guidance during endoscopic procedures;

- **Quality of mucosal views:** it consists in the complete visualisation of the entire mucosa to avoid overlooking uncovered pathology. For the colon there is a classification of mucosal visualisation, Boston Bowel Preparation Scale;
- **Pathological findings:** Abnormalities affecting all parts of the gastrointestinal tract due to the pathological disease discovered. Most pathologies can be recognised more or less easily on the mucous membrane of the intestinal walls;
- **Therapeutic Intervention:** therapeutic intervention after some injury or pathological finding has been detected. For example: lifting and resecting a polyp, dilation of a stenosis or injection of a bleeding ulcer;

Within each category are directories that represent each class, for example, images intended for polyps are in the polyps directory. The images of each class are not balanced. Figure 4.3 shows a histogram demonstrating the unbalance of the classes related to the pathologies of the lower gastrointestinal tract. The unlabelled images represent a total of 99 417 images. In addition to the unlabelled images, extracted features and cluster assignments are also offered in an Attribute-Relation File Format (ARFF). The segmented images consist of the original image, a segmentation mask and a bounding box for 1000 images of the polyps class. The bounding box consists of the outermost pixels of the polyp. While, in the mask, the pixels representing the polyp tissue are present in the foreground (area in white) and the background (area in black) contains no pixels of the polyp. Finally, the labelled videos represent a total of 373 videos that have different findings and landmarks. This number of videos represents 11.62 hours of videos and a further 1 million video frames that can be converted to images if required. In total, 171 videos were analysed by a gastroenterologist and labelled [89].

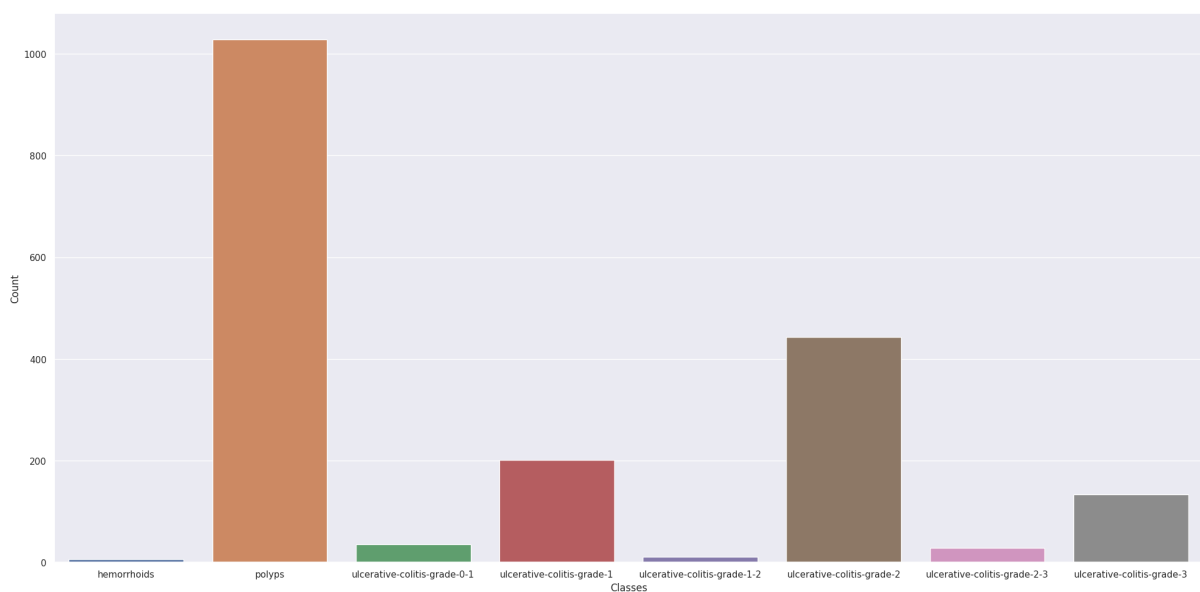


Figure 4.3: Distribution of the classes regarding to the pathologies of the lower gastrointestinal tract.

The CrohnIPI dataset [90] consists of a multicentre dataset approved by the French Data Protection Authority and the Ethics Community of Nantes, France. This dataset was collected between 2014 and 2018 using third-generation SBCE videos (Pillcam SB3 system, Medtronic, Minnesota, United States) of patients with Crohn’s disease or suspected to have the disease [91]. After collection, only videos where the presence of the lesions caused by Crohn’s disease was verified were selected and successive frames were extracted in JPEG format. Subsequently, these images were labelled by a gastroenterologist trainee and by three professionals with more than 10 years of experience in managing patients with inflammatory bowel disease. This dataset has been used to develop tools for illness diagnosis [92, 93].

The various types of lesions recognised in the videos captured by the endoscopic equipment are described in Table 4.1.

Table 4.1: Description of the type of lesions presented in the CrohnIPI dataset

Pathological lesion	Description
Erythema	Area of reddish villi
Edema	Enlarged or swollen or engorged
Aphthoid ulceration	Marked loss of the epithelial layer with a whitish centre and a red halo, surrounded by normal mucosa
Ulceration	Deep loss of tissue with a whitish bottom compared to the surrounding swollen or edematous mucosa
Stenosis	Tightening of the intestinal lumen delaying the passing of the capsule

The LIMUC dataset made available in [27] consists of 11276 images with a resolution of 352x288, which were collected through 1043 colonoscopy procedures. For the collection of these images, 564 patients who underwent a colonoscopy for the diagnosis of ulcerative colitis disease between December 2011 and July 2019 at the Department of Gastroenterology, Marmara University School of Medicine were selected. All images were first classified and reviewed by two experienced gastroenterologists according to the endoscopic Mayo score (EMS) and then by a third experienced gastroenterologist without knowing the label assigned previously by the other gastroenterologists. Differently labelled images were assigned a final Mayo score using the most voted one. Therefore, the dataset is divided into 6105 images labelled as Mayo 0, 3052 images labelled as Mayo 1, 1254 images labelled as Mayo 2 and 865 images labelled as Mayo 3.

As the aim of the present study is to automatically detect the type of inflammatory bowel disease (e.g. Crohn’s disease and Ulcerative Colitis), the images referring to the Ulcerative Colitis pathology of the HyperKvasir and LIMUC datasets will be used, as well as the images with the type of Crohn’s disease lesions present in the CrohnIPI dataset. Figure 4.4 shows an example of the images that exist in each collected dataset.

Therefore, three different datasets were created (Table 4.2):

- **Dataset 1:** this dataset consists of 850 instances of the HyperKvasir dataset referring to ulcerative colitis disease; and 1360 instances of the CrohnIPI dataset referring to Crohn’s disease. Also incorporate 64 video frames extracted from the labelled videos of the HyperKvasir dataset. Figure 4.5a shows the class distribution of this dataset;
- **Dataset 2:** this dataset consists of 446 instances of the LIMUC dataset, the instances are equally distributed by severity degree. That is, 25% of the total images referring to the severity degree were extracted from the original dataset; 64 video frames were extracted from the labelled videos from the HyperKvasir dataset; 850 instances from the HyperKvasir dataset; and 1360 instances from the CrohnIPI dataset. Figure 4.5b shows the class distribution of this dataset. The objective of the creation of this dataset was to mitigate the unbalance of the classes presented on Dataset 1 and to include more diversity of images with Ulcerative colitis;
- **Dataset 3:** this dataset consists of 1360 instances of the LIMUC dataset regarding the disease ulcerative colitis. The instances are equally distributed by severity grade. That is, 25% of the total images referring to severity grade were extracted from the original dataset; and 1360 instances of the dataset CrohnIPI referring to Crohn’s disease. Figure 4.5c shows the class distribution of this dataset. The existence of a dataset with ulcerative colitis images in large numbers allowed us to create this dataset with the classes balanced. Allowing, the realization of another experiment.

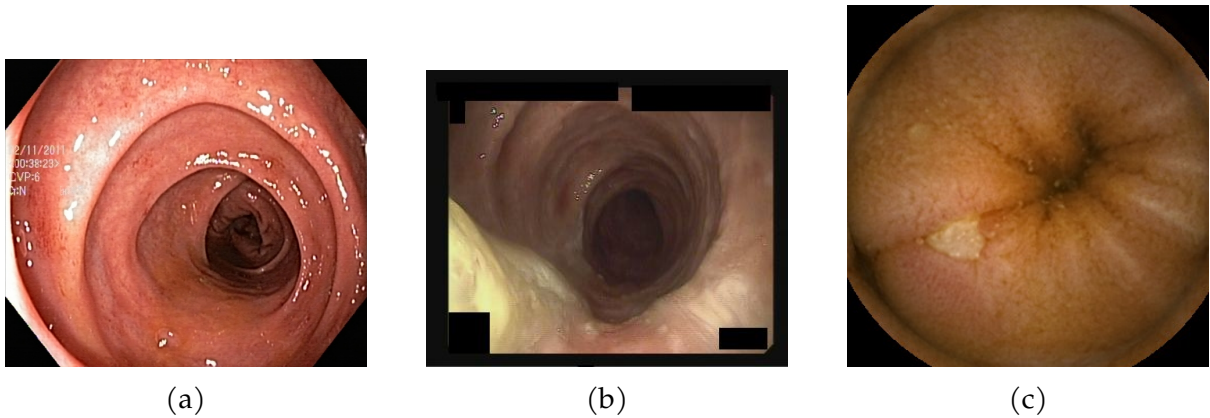


Figure 4.4: Ulcerative colitis image from the HyperKvasir dataset (a); Ulcerative colitis image from the LIMUC dataset (b); Crohn’s disease image from the CrohnIPI dataset (c).

4.2 Experimental setup

In conducting this study, Tensorflow, version 2.8.0, Tfimm, version 0.6.13, and Shap, version 0.40.0, were used. The programming environment for importing the libraries

was Google Colab with the NVIDIA A100 GPU ¹.

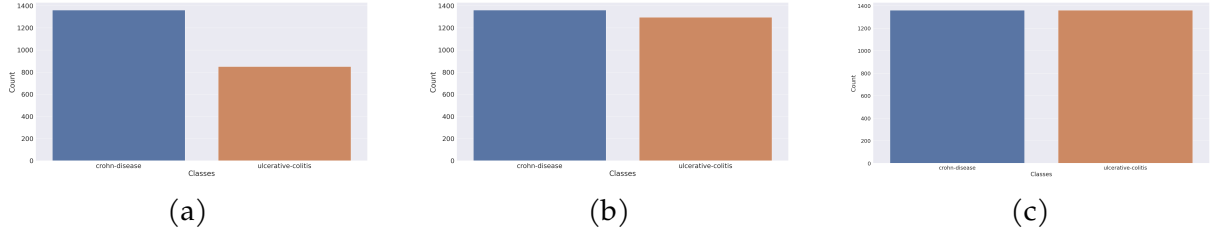


Figure 4.5: Class distribution: (a) Dataset 1; (b) Dataset 2; (c) Dataset 3.

Based on the datasets created, a total of nine experiments were realized, allowing different combinations. Therefore, it started with developing three experiments, where the first experiment consisted in using dataset 1. The second experiment had the aim of balancing the classes and diversifying the images regarding class ulcerative colitis, for which was used the dataset 2. Experiment three consisted of using dataset 3 and the partial use of the proposed methodology. That is, during this experiment, there are no frames extracted from videos and no image preprocessing techniques were applied.

Table 4.2: Description of the created datasets.

Dataset	Description
Dataset 1	From Ulcerative Colitis contains: i) 850 instances of the HyperKvasir dataset; ii) 64 video frames extracted from the labelled videos of the HyperKvasir dataset. From Crohn's disease contains 1,360 instances of the CrohnIPI dataset.
Dataset 2	From Ulcerative colitis contains: i) 850 instances of the HyperKvasir dataset; ii) 64 video frames extracted from the labelled videos of the HyperKvasir dataset; iii) 46 instances of the LIMUC dataset. From Crohn's disease contains 1,360 instances of the CrohnIPI dataset.
Dataset 3	From Ulcerative Colitis contains 1,360 instances of the LIMUC dataset. From Crohn's disease contains 1,360 instances of the CrohnIPI dataset.

The presence of the green square in the lower-left corner of the images obtained from the HyperKvasir dataset, rose the concern that could lead the models to identify the green square as a disease feature. Therefore, more six experiments were performed to test this hypothesis, by combining the three pre-processing techniques presented in Chapter 4.3 with Datasets 1 and 2 described in 4.1. We refer to Table 4.3 for the configurations of each experiment.

4.3 Pre-processing

In the images and frames extracted from a video of the HyperKvasir dataset, there was a green square in the lower left corner, which represents the position of the endoscope during the diagnostic examination. This may become a problem for classifying the

¹Code is available at: https://github.com/MauricioAI/tese_mestrado_ISEC

Table 4.3: Configurations of the nine performed experiments

Experiment	Pre-processing	Dataset
Experiment 1	N/A	Dataset 1
Experiment 2	N/A	Dataset 2
Experiment 3	N/A	Dataset 3
Experiment 4	Horizontal crop	Dataset 1
Experiment 5	Horizontal crop	Dataset 2
Experiment 6	Vertical crop	Dataset 1
Experiment 7	Vertical crop	Dataset 2
Experiment 8	Gaussian blur	Dataset 1
Experiment 9	Gaussian blur	Dataset 2

type of inflammatory bowel disease present in the images. Therefore, experiments were performed where some transformations were applied when this type of situation was verified.

One of the applied transformations was a horizontal crop across the entire width of the image to remove the green square in the lower left corner. Next, a vertical crop was applied over the entire height of the image to remove that part of the image. However, during these transformations, some information regarding the mucosa of the intestine is lost. To circumvent this loss, some more experiments were performed in which a Gaussian blur of 2/16 was applied only to the area where the green square was. The preprocessing techniques were proposed in [8]. For more details, refer to Appendix A - Articles, where the full paper can be found.

The goal was to preserve as much of the image information as possible but to disguise that feature in some images. An illustration of these pre-processing methods is available in Figure 4.6.

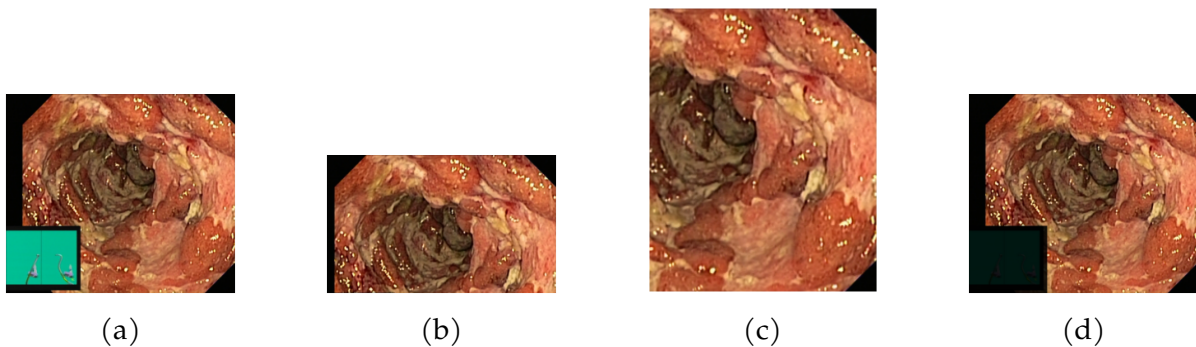


Figure 4.6: Pre-processing examples: (a) Original image; (b) Horizontal crop; (c) Vertical crop; (d) Gaussian blur.

4.4 Data augmentation

Before classifying the images, data augmentation was performed on the train set with horizontal and vertical Random Flip, Random Contrast with a factor of 0.15, Random Rotation with a factor of 0.2 and a Random Zoom with a portion of -0.2 for height and -0.3 for width.

4.5 Convolutional Neural Networks

The following architectures were used to classify images of inflammatory bowel diseases: ResNet50, VGG16, and InceptionV3. These architectures have been pre-trained with the ImageNet dataset, except the AlexNet network. Besides these architectures, another architecture was built that combines a ResNet50 with a MobileNetV2 network. Finally, a hybrid model was built where a CNN is combined with an LSTM. These architectures were chosen based on results presented in some studies of the literature review [75, 78, 80, 72, 81].

The architecture that combines the ResNet50 network and the MobileNetV2 network consists in an ensemble method, where the two architectures were combined. That is, the input images will be directed to the two architectures that will process these images without the last classification layer and the output of the two will be concatenated to be classified by a Dense layer with two neurons referring to class Crohn disease and Ulcerative Colitis. The construction of this model was inspired by [94, 95]. This architecture is illustrated in Figure 4.7.

The hybrid model built in this work was based on a similar architecture developed by other authors [96, 78, 97]. This architecture is based on: 8 Conv2D layers, 2 BatchNormalization layers, 4 MaxPooling2D layers, 1 Flatten layer, 7 Dense layers, 3 Dropout layers and 3 LSTM layers. The hybrid model is illustrated in Figure 4.8. The forget gate in LSTM networks determines which information should be remembered and which should be forgotten before sending it to the subsequent layers [78].

The architectures used to classify the images were trained for 200 epochs, using SparseCategoricalCrossentropy as the loss function, a batch size of 32, and Adam with a learning rate of $1.00e-05$ was set as the optimizer [98]. During training, an EarlyStopping callback was used with the patience of 5 to monitor the validation accuracy [99, 100]. In all experiments, the images were resized to 224x224 pixels.

4.6 Vision Transformers

In addition to convolutional neural networks, transformer-based deep learning architectures were also used in this work, so that we could compare the two types of deep

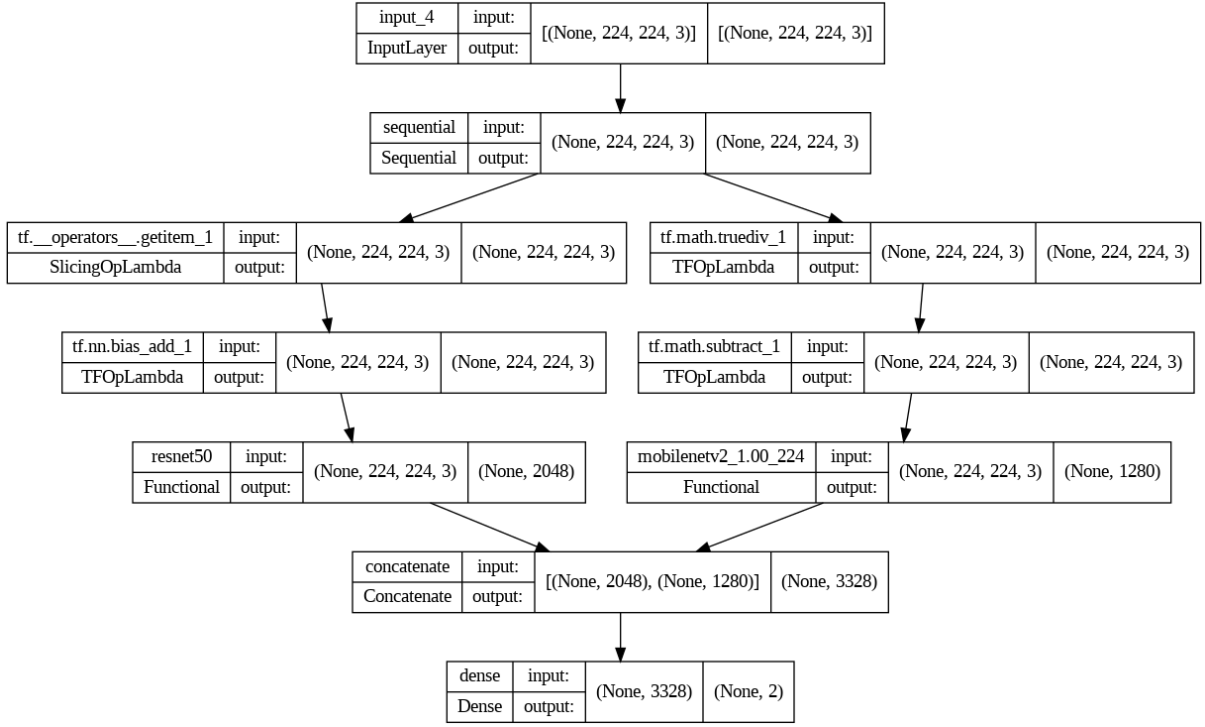


Figure 4.7: Architecture of the ResNet50+MobileNetV2 model.

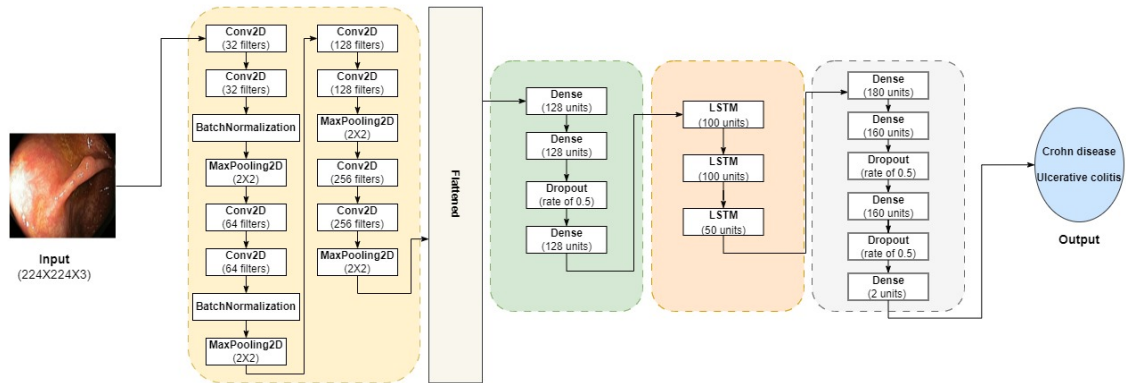


Figure 4.8: Architecture of the hybrid model.

learning architectures in classifying the type of inflammatory bowel disease. Therefore, in this chapter, the vision transformers used are presented.

Within the vision transformers, the models used were: ViT-B/32, ViT-S/32, ViT-B/16, ViT-S/16 and R26+S/32 [55, 56, 57]. These models were pre-trained with the Imagenet21k dataset. The R26+S/32 model was created by Google and made available, as well as other versions very similar to this one, which consists of combining a ResNet network with a vision transformer [101]. That is, the first part of the model name (R26) represents the number of convolutional layers that were used from the ResNet network and the second part of the name (S/32) means that those layers extracted from

the ResNet network were combined with a small vision transformer of 32 patches.

Using SparseCategoricalCrossentropy as the loss function, a batch size of 32, and Adam with a learning rate of $1.00e-05$ as the optimizer, the models used to classify the pictures were trained for 200 epochs [98]. An EarlyStopping callback with a patience of 5 was utilized during training to track the validity accuracy. The images were resized to 224x224 pixels in all experiments.

4.7 Knowledge Distillation

Deploying these bulky architectures into medical systems to assist gastroenterologists in diagnosing the disease can be very complex and even impractical due to the time it takes to process the images, as well as the computational resources required to store these models. Therefore, this sub-chapter will describe the process of distilling knowledge from core architectures to simpler and lighter architectures.

For the CNNs, a student with a total of 1,254,194 parameters was created, with the Adam optimizer using a learning rate of $1.00e-04$ and the SparseCategoricalCrossentropy loss function. During distillation, the Kullback-Leibler Divergence function represented by the formula (4.1) was set as the loss function, with an alpha of 0.8 and a temperature of 105 degrees.

$$D_{KL}(p \parallel q) = \sum_{i=1}^N p(x_i) \log\left(\frac{p(x_i)}{q(x_i)}\right) \quad (4.1)$$

For student training, 100 epochs and an EarlyStopping callback with a patience of 5 were set to monitor the validation accuracy.

In the case of ViTs, the distillation process was done in two different ways: i) using the temperature technique to scale the logits returned by the teacher's; ii) using the DeiT technique [43]. In the implementation of the first technique, an architecture with a total of 2,300,162 parameters was built to distil the 32-patched architectures and an architecture with a total of 3,442,274 parameters to distil the 16-patched architectures. The same configurations were used in the distillation of the CNNs.

In the second technique, an architecture with a total of 3,554,884 parameters was built to distil the architectures with 32 patches and an architecture with 3,112,516 parameters in total to do the distillation of the architectures with 16 patches. These students were configured with the AdamW optimizer with a weight decay of 0.0001 and a learning rate scaled of $6.25e-07$, obtained through the formula:

$$lr_{scaled} = \frac{lr}{512} \times batchsize \quad (4.2)$$

The formula consisted of the quotient between the learning rate (lr), which is equal to $1.00e-05$, and the base 512. The authors [43] describe the use of the 512 base as an improvement in performance. This result is multiplied by the 32 batch size.

The `SparseCategoricalCrossentropy` function was defined as the student's loss function and as the distillation function. The students were trained for 100 epochs, using an `EarlyStopping` callback with a patience of 5. In both distillation techniques of the ViTs, as well as in the knowledge distillation of CNNs, data augmentation was used [102, 103, 43].

4.8 Interpretability

It is important to understand in which areas of the images the models were based to predict the type of inflammatory bowel disease and in this sub-chapter, the interpretability algorithms used to interpret the output of the teachers and the students in the CNNs and ViTs will be described.

For the implementation of the CNNs, the following algorithms were selected: Grad-CAM, LIME, Shap values, RISE and Occlusion sensitivity to interpret the teachers' output. On the other hand, the Shap values and Grad-CAM algorithms were excluded from the interpretation of the students' CNNs. In the ViTs implementation, the Shap values, LIME, RISE and Occlusion sensitivity algorithms were selected to interpret the teachers' output. In the interpretation of the students, the Shap values algorithm was excluded.

Grad-CAM and Shap values algorithms were excluded from the interpretability of the students of the CNNs, and the Shap values algorithm was not used in the interpretability from the students of the ViTs because due to the compilation of the teacher-student models in the same architecture was not possible to get Gradients of the intermediate layers.

4.9 Evaluation

In all experiments, the dataset was split into 70% for training, 20% for validation, and 10% for testing [80].

In total, of the 64 frames extracted, 30 frames belonged to the same video in which the green square in the lower left corner was found. These frames were attached to the images in the validation dataset. The remaining frames that belonged to another video and did not have the green square were introduced in the training dataset.

Classification metrics were selected to evaluate the network's and the transformer's performance on the test dataset, as well as the student's performance on the test dataset.

These include Accuracy, Precision, Recall, F1-Score, Area Under Curve (AUC), Mathew's Correlation Coefficient (MCC), Inference time [104, 105] and Loss:

$$Loss = - \sum_{i=1}^n y_i \cdot \log \hat{y}_i \quad (4.3)$$

AUC which measures the algorithm's ability to distinguish between classes. The higher the AUC value, the better the algorithm can distinguish between classes [106]. Accuracy refers to the ratio of correct predictions made by the model. It is measured by the total number of correct predictions divided by the total number of predictions [107].

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (4.4)$$

True Positive (TP) is the number where the model correctly predicted the positive class, which in this work concerns Ulcerative colitis. True Negative (TN) is the number of predictions correctly made of the negative class, which means predicting Crohn's disease. On the other hand, False Positive (FP) is the number of predictions incorrectly made of the positive class. False Negative (FN) is the number of predictions incorrectly made from the negative class [108].

Precision calculates the ratio of true positives predicted. This is measured by the number of true positives divided by the sum between the number of true positives and the number of false positives [109].

$$Precision = \frac{TP}{TP + FP} \quad (4.5)$$

Recall measures the ratio of true positives and the amount of correctly predicted classes. It is obtained through the total number of true positives divided by the sum between the total number of true positives and the number of false negatives [109].

$$Recall = \frac{TP}{TP + FN} \quad (4.6)$$

F1-Score is the harmonic mean of Precision and Recall and provides an estimate of the misclassified cases, sometimes better than the classification accuracy metric [107]. This is due to the fact that accuracy can be misleading in imbalanced datasets, since it is based on the majority class.

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4.7)$$

MCC is a metric designed to evaluate the performance of algorithms in binary classification problems. It calculates the correlation coefficient between the true and predicted

classes. Unlike other classification metrics, MCC offers advantages because it remains unaffected by imbalanced datasets. The formula to calculate the correlation coefficient is shown in (4.8), and the resulting values range between -1 and 1. A value of 1 indicates a perfect classifier ($FP=FN=0$). Conversely, a value of -1 signifies confusion in classifying the classes ($TP=TN=0$) [104, 110].

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (4.8)$$

5 RESULTS

After the algorithms selected for this work had been executed following the proposed methodology, the results obtained by the chosen classification metrics were collected. In this sense, this chapter presents through the tables the quantitative results obtained by the CNNs and the ViTs, as well as the results obtained by the students of each architecture. The results from experiments 4 to 9 of the ViTs and CNNs were presented in [10]. The full paper can be found in Appendix A - Articles.

5.1 Convolutional Neural Networks

Tables 5.1 and 5.2 present the results of the classification metrics obtained by the CNNs during experiments 1 and 2, respectively. In these tables, it can be observed that all main architectures show a good performance in predicting the type of inflammatory bowel disease, as well as the respective students. An exception is observed in the student of the AlexNet network.

Table 5.1: Results of the CNNs in experiment 1.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	AlexNet	0.9953	0.0074	1.0000	0.9878	0.9939	0.9962	0.9901	1.02s	46,760,706
	ResNet50	1.0000	0.0452	1.0000	1.0000	1.0000	1.0000	1.0000	3.90s	23,591,810
	VGG16	1.0000	0.0006	1.0000	1.0000	1.0000	1.0000	1.0000	5.52s	134,268,738
	InceptionV3	1.0000	0.0410	1.0000	1.0000	1.0000	1.0000	1.0000	2.91s	21,806,882
	ResNet50+MobileNetV2	1.0000	0.0391	1.0000	1.0000	1.0000	1.0000	1.0000	3.61s	25,852,354
	Hybrid model	1.0000	0.3311	1.0000	1.0000	1.0000	1.0000	1.0000	2.42s	1,930,966
Knowledge Distillation	AlexNet student	0.4883	1.6641	1.0000	0.4263	0.5978	0.5871	0.2725	2.00s	1,254,194
	ResNet50 student	0.9718	0.1149	1.0000	0.9310	0.9643	0.9773	0.9427	1.81s	1,254,194
	VGG16 student	1.0000	0.0009	1.0000	1.0000	1.0000	1.0000	1.0000	1.67s	1,254,194
	InceptionV3 student	0.9953	0.0669	0.9877	1.0000	0.9938	0.9938	0.9901	1.40s	1,254,194
	ResNet50+MobileNetV2 student	0.9906	0.0621	1.0000	0.9759	0.9878	0.9924	0.9804	1.60s	1,254,194

Table 5.2: Results of the CNNs in experiment 2.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	AlexNet	0.9965	0.0207	0.9926	1.0000	0.9963	0.9963	0.9929	0.87s	46,760,706
	ResNet50	1.0000	0.0225	1.0000	1.0000	1.0000	1.0000	1.0000	3.66s	23,591,810
	VGG16	1.0000	0.0100	1.0000	1.0000	1.0000	1.0000	1.0000	5.61s	134,268,738
	InceptionV3	0.9894	0.0339	1.0000	0.9783	0.9890	0.9898	0.9789	3.78s	21,806,882
	ResNet50+MobileNetV2	1.0000	0.0329	1.0000	1.0000	1.0000	1.0000	1.0000	3.93s	25,852,354
	Hybrid model	0.9894	0.3453	1.0000	0.9783	0.9890	0.9898	0.9789	2.90s	1,930,966
Knowledge Distillation	AlexNet student	0.9894	0.0460	0.9852	0.9925	0.9888	0.9892	0.9787	2.49s	1,254,194
	ResNet50 student	0.9894	0.0539	1.0000	0.9783	0.9890	0.9898	0.9789	1.72s	1,254,194
	VGG16 student	0.9894	0.0803	0.9926	0.9853	0.9889	0.9895	0.9787	1.74s	1,254,194
	InceptionV3 student	0.9929	0.0384	1.0000	0.9854	0.9926	0.9932	0.9859	1.73s	1,254,194
	ResNet50+MobileNetV2 student	0.9894	0.0563	1.0000	0.9783	0.9890	0.9898	0.9789	1.88s	1,254,194

Table 5.3 presents the results obtained during experiment 3, where dataset 3 without the application of pre-processing techniques was used. During this experiment, the students of the ensemble method and the ResNet50 network were unable to correctly learn the information transmitted by the teacher model. It is also verified that, even

though the teacher models demonstrate good performance in the classification of the IBD images, when compared with the other experiments, there was a considerable reduction of the results during the distillation process.

Table 5.3: Results of the CNNs in experiment 3.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	AlexNet	1.0000	0.0022	1.0000	1.0000	1.0000	1.0000	1.0000	0.76s	46,760,706
	ResNet50	1.0000	0.0044	1.0000	1.0000	1.0000	1.0000	1.0000	2.44s	23,591,810
	VGG16	1.0000	0.0000	1.0000	1.0000	1.0000	1.0000	1.0000	3.43s	134,268,738
	InceptionV3	1.0000	0.0067	1.0000	1.0000	1.0000	1.0000	1.0000	2.94s	21,806,882
	ResNet50+MobileNetV2	1.0000	0.0121	1.0000	1.0000	1.0000	1.0000	1.0000	3.68s	25,852,354
	Hybrid model	0.9926	0.3328	1.0000	0.9852	0.9925	0.9927	0.9853	2.23s	1,930,966
Knowledge Distillation	AlexNet student	0.8963	0.2241	1.0000	0.8261	0.9048	0.8978	0.8107	1.52s	1,254,194
	ResNet50 student	0.5037	1.3675	1.0000	0.4981	0.6650	0.5109	0.1044	1.55s	1,254,194
	VGG16 student	1.0000	0.0004	1.0000	1.0000	1.0000	1.0000	1.0000	1.30s	1,254,194
	InceptionV3 student	0.9185	0.2287	1.0000	0.8581	0.9236	0.9197	0.8487	1.28s	1,254,194
	ResNet50+MobileNetV2 student	0.4926	1.2775	1.0000	0.4926	0.66000	0.5000	0.0000	1.39s	1,254,194

Table 5.4 shows the results obtained by the CNNs in experiment 4. It can be observed that both deep learning models obtained a good performance in image classification and through knowledge distillation it was possible to reduce the inference time in most models, except in the AlexNet network. However, in this experiment, it is verified that the ResNet50 and VGG16 students were not able to correctly learn the information distilled by the teacher model.

Table 5.4: Results of the CNNs in experiment 4.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	AlexNet	1.0000	0.0031	1.0000	1.0000	1.0000	1.0000	1.0000	0.75s	46,760,706
	ResNet50	1.0000	0.0297	1.0000	1.0000	1.0000	1.0000	1.0000	3.65s	23,591,810
	VGG16	1.0000	0.0016	1.0000	1.0000	1.0000	1.0000	1.0000	3.80s	134,268,738
	InceptionV3	1.0000	0.0596	1.0000	1.0000	1.0000	1.0000	1.0000	2.95s	21,806,882
	ResNet50+MobileNetV2	1.0000	0.0429	1.0000	1.0000	1.0000	1.0000	1.0000	4.20s	25,852,354
	Hybrid model	0.9953	0.4380	0.9877	1.0000	0.9938	0.9938	0.9901	2.31s	1,930,966
Knowledge Distillation	AlexNet student	1.0000	0.0026	1.0000	1.0000	1.0000	1.0000	1.0000	2.51s	1,254,194
	ResNet50 student	0.3991	1.3687	1.0000	0.3876	0.5586	0.5152	0.1084	2.06s	1,254,194
	VGG16 student	0.4272	2.6307	1.0000	0.3990	0.5704	0.5379	0.1739	1.61s	1,254,194
	InceptionV3 student	1.0000	0.0702	1.0000	1.0000	1.0000	1.0000	1.0000	1.45s	1,254,194
	ResNet50+MobileNetV2 student	0.9953	0.0476	1.0000	0.9878	0.9939	0.9962	0.9901	1.69s	1,254,194

Table 5.5 presents the results of experiment 5 regarding CNN networks. A comparison with experiment 4 reveals a slight decrease in performance for the teacher models, attributed to the inclusion of images from the LIMUC dataset associated with Ulcerative colitis disease. However, the students of the ResNet50 and VGG16 networks were able to learn better the information distilled from their teachers.

Table 5.5: Results of the CNNs in experiment 5.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	AlexNet	0.9858	0.0230	0.9852	0.9852	0.9852	0.9858	0.9716	1.08s	46,760,706
	ResNet50	0.9965	0.0339	1.0000	0.9926	0.9963	0.9966	0.9929	2.68s	23,591,810
	VGG16	0.9965	0.0038	0.9926	1.0000	0.9963	0.9963	0.9929	5.60s	134,268,738
	InceptionV3	0.9929	0.0408	1.0000	0.9854	0.9926	0.9932	0.9859	3.96s	21,806,882
	ResNet50+MobileNetV2	1.0000	0.0177	1.0000	1.0000	1.0000	1.0000	1.0000	4.18s	25,852,354
	Hybrid model	0.9929	0.3326	1.0000	0.9854	0.9926	0.9932	0.9859	3.31s	1,930,966
Knowledge Distillation	AlexNet student	0.9752	0.0998	0.9926	0.9571	0.9745	0.9759	0.9509	2.10s	1,254,194
	ResNet50 student	0.9610	0.1058	0.9185	1.0000	0.9575	0.9593	0.9244	1.73s	1,254,194
	VGG16 student	0.9929	0.1006	0.9926	0.9926	0.9926	0.9929	0.9858	1.76s	1,254,194
	InceptionV3 student	0.9823	0.1043	1.0000	0.9643	0.9818	0.9830	0.9651	3.21s	1,254,194
	ResNet50+MobileNetV2 student	0.9894	0.0738	0.9926	0.9853	0.9889	0.9895	0.9787	1.69s	1,254,194

Tables [5.6 - 5.9] show the results obtained by the CNN teacher and student models from experiments 6 to 9. In experiment 8, it is verified that the hybrid model of CNNs

Diagnose of the IBD using Deep Learning Models with Interpretability

did not obtain a good performance. In addition, in each experiment using the same distillation technique and when comparing inference time between ViTs and CNNs, it was possible to substantially reduce the inference time in the ViTs models. Although the student models present a higher number of parameters than the CNNs' student model.

Table 5.6: Results of the CNNs in experiment 6.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	AlexNet	1.0000	0.0024	1.0000	1.0000	1.0000	1.0000	1.0000	0.78s	46,760,706
	ResNet50	1.0000	0.0376	1.0000	1.0000	1.0000	1.0000	1.0000	2.27s	23,591,810
	VGG16	1.0000	0.0000	1.0000	1.0000	1.0000	1.0000	1.0000	3.56s	134,268,738
	InceptionV3	0.9906	0.0644	1.0000	0.9759	0.9878	0.9924	0.9804	2.94s	21,806,882
	ResNet50+MobileNetV2	1.0000	0.0338	1.0000	1.0000	1.0000	1.0000	1.0000	3.61s	25,852,354
	Hybrid model	1.0000	0.3582	1.0000	1.0000	1.0000	1.0000	1.0000	2.43s	1,930,966
Knowledge Distillation	AlexNet student	0.9906	0.0207	0.9753	1.0000	0.9875	0.9877	0.9802	2.15s	1,254,194
	ResNet50 student	0.9859	0.0625	0.9877	0.9756	0.9816	0.9863	0.9702	1.72s	1,254,194
	VGG16 student	1.0000	0.0006	1.0000	1.0000	1.0000	1.0000	1.0000	1.43s	1,254,194
	InceptionV3 student	0.9906	0.0711	1.0000	0.9759	0.9878	0.9924	0.9804	1.70s	1,254,194
	ResNet50+MobileNetV2 student	0.9906	0.0758	1.0000	0.9759	0.9878	0.9924	0.9804	1.90s	1,254,194

Table 5.7: Results of the CNNs in experiment 7.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	AlexNet	0.9858	0.0354	0.9926	0.9781	0.9853	0.9861	0.9717	1.07s	46,760,706
	ResNet50	1.0000	0.0262	1.0000	1.0000	1.0000	1.0000	1.0000	3.70s	23,591,810
	VGG16	1.0000	0.0007	1.0000	1.0000	1.0000	1.0000	1.0000	5.59s	134,268,738
	InceptionV3	0.9965	0.0369	1.0000	0.9926	0.9963	0.9966	0.9929	3.87s	21,806,882
	ResNet50+MobileNetV2	1.0000	0.0179	1.0000	1.0000	1.0000	1.0000	1.0000	4.96s	25,852,354
	Hybrid model	0.9823	0.3591	1.0000	0.9643	0.9818	0.9830	0.9651	2.67s	1,930,966
Knowledge Distillation	AlexNet student	0.9894	0.0467	0.9926	0.9853	0.9889	0.9895	0.9787	2.50s	1,254,194
	ResNet50 student	0.9929	0.0754	1.0000	0.9854	0.9926	0.9932	0.9859	1.72s	1,254,194
	VGG16 student	0.9929	0.0862	1.0000	0.9854	0.9926	0.9932	0.9859	1.88s	1,254,194
	InceptionV3 student	0.9929	0.0422	1.0000	0.9854	0.9926	0.9932	0.9859	1.66s	1,254,194
	ResNet50+MobileNetV2 student	0.9929	0.0592	1.0000	0.9854	0.9926	0.9932	0.9859	1.80s	1,254,194

Table 5.8: Results of the CNNs in experiment 8.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	AlexNet	1.0000	0.0007	1.0000	1.0000	1.0000	1.0000	1.0000	0.74s	46,760,706
	ResNet50	1.0000	0.0591	1.0000	1.0000	1.0000	1.0000	1.0000	2.58s	23,591,810
	VGG16	1.0000	0.0002	1.0000	1.0000	1.0000	1.0000	1.0000	3.65s	134,268,738
	InceptionV3	0.9953	0.0638	1.0000	0.9878	0.9939	0.9962	0.9901	3.48s	21,806,882
	ResNet50+MobileNetV2	1.0000	0.0356	1.0000	1.0000	1.0000	1.0000	1.0000	3.58s	25,852,354
	Hybrid model	0.4413	0.6797	1.0000	0.4050	0.5765	0.5492	0.1997	2.40s	1,930,966
Knowledge Distillation	AlexNet student	1.0000	0.0044	1.0000	1.0000	1.0000	1.0000	1.0000	2.15s	1,254,194
	ResNet50 student	1.0000	0.0448	1.0000	1.0000	1.0000	1.0000	1.0000	1.69s	1,254,194
	VGG16 student	0.9906	0.1358	1.0000	0.9759	0.9878	0.9924	0.9804	1.67s	1,254,194
	InceptionV3 student	0.9953	0.0457	0.9877	1.0000	0.9938	0.9938	0.9901	1.68s	1,254,194
	ResNet50+MobileNetV2 student	1.0000	0.0489	1.0000	1.0000	1.0000	1.0000	1.0000	2.22s	1,254,194

Table 5.9: Results of the CNNs in experiment 9.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	AlexNet	0.9858	0.0323	0.9778	0.9925	0.9851	0.9855	0.9717	1.04s	46,760,706
	ResNet50	1.0000	0.0104	1.0000	1.0000	1.0000	1.0000	1.0000	2.63s	23,591,810
	VGG16	1.0000	0.0016	1.0000	1.0000	1.0000	1.0000	1.0000	5.58s	134,268,738
	InceptionV3	1.0000	0.0271	1.0000	1.0000	1.0000	1.0000	1.0000	3.68s	21,806,882
	ResNet50+MobileNetV2	0.9965	0.0216	1.0000	0.9926	0.9963	0.9966	0.9929	4.27s	25,852,354
	Hybrid model	0.9787	0.3919	1.0000	0.9574	0.9783	0.9796	0.9583	2.72s	1,930,966
Knowledge Distillation	AlexNet student	0.9645	0.1238	0.9926	0.9371	0.9640	0.9657	0.9306	2.64s	1,254,194
	ResNet50 student	0.9787	0.0611	1.0000	0.9574	0.9783	0.9796	0.9583	1.74s	1,254,194
	VGG16 student	0.9858	0.3479	1.0000	0.9712	0.9854	0.9864	0.9720	3.20s	1,254,194
	InceptionV3 student	0.9752	0.0860	1.0000	0.9507	0.9747	0.9762	0.9515	1.76s	1,254,194
	ResNet50+MobileNetV2 student	0.9007	0.2484	1.0000	0.8282	0.9060	0.9048	0.8188	2.00s	1,254,194

5.2 Vision Transformers

Tables 5.10 and 5.11 present the results obtained by the ViTs during experiments 1 and 2, respectively. Table 5.12 allows visualizing the results obtained during experiment 3 by the ViTs. Table 5.13 shows the results obtained by the ViTs in experiment 4. Comparing the results of the four experiments, it is observed that in experiment 2 the fact of joining the three collected datasets caused an overall increase in the inference time of the architectures, as well as caused a drop in the performance of the teacher of the ViT-S/16 model.

On the other hand, comparing experiments 1 and 4, it is concluded that the application of a horizontal crop in the Ulcerative colitis images that were gathered from the HyperKvasir dataset caused a reduction of the inference time in the knowledge distillation through attention, as well as improved the students' performance. In experiment 3, it is visible that some architectures present an improvement in performance during both distillation processes. However, their inference time has increased.

Table 5.10: Results of the ViTs in experiment 1.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	ViT-B/32	1.0000	0.0013	1.0000	1.0000	1.0000	1.0000	1.0000	3.16s	87,417,602
	ViT-S/32	1.0000	0.0008	1.0000	1.0000	1.0000	1.0000	1.0000	4.01s	22,475,138
	ViT-B/16	1.0000	0.0003	1.0000	1.0000	1.0000	1.0000	1.0000	5.23s	86,417,594
	ViT-S/16	1.0000	0.0007	1.0000	1.0000	1.0000	1.0000	1.0000	4.54s	21,590,402
	R26+S/32	1.0000	0.0014	1.0000	1.0000	1.0000	1.0000	1.0000	6.15s	36,028,098
Knowledge Distillation	ViT-B/32 student	1.0000	0.0030	1.0000	1.0000	1.0000	1.0000	1.0000	1.52s	2,300,162
	ViT-S/32 student	0.9859	0.0305	0.9630	1.0000	0.9811	0.9815	0.9703	0.94s	2,300,162
	ViT-B/16 student	0.9437	0.1201	1.0000	0.8710	0.9310	0.9545	0.8898	1.96s	3,442,274
	ViT-S/16 student	0.9671	0.1367	0.9136	1.0000	0.9548	0.9568	0.9314	1.33s	3,442,274
	R26+S/32 student	0.9202	0.2733	0.7901	1.0000	0.8828	0.8951	0.8366	1.27s	2,300,162
DeiT	ViT-B/32 student	0.9343	0.1549	0.8395	0.9855	0.9067	0.9160	0.8630	1.39s	3,554,884
	ViT-S/32 student	0.9484	0.1330	0.8765	0.9861	0.9281	0.9345	0.8918	1.24s	3,554,884
	ViT-B/16 student	0.8920	0.2842	0.7160	1.0000	0.8345	0.8580	0.7809	1.39s	3,112,516
	ViT-S/16 student	0.6291	0.5473	0.0247	1.0000	0.0482	0.5123	0.1243	1.08s	3,112,516
	R26+S/32 student	0.9437	0.1396	0.8519	1.0000	0.9200	0.9259	0.8837	1.01s	3,554,884

Table 5.11: Results of the ViTs in experiment 2.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	ViT-B/32	0.9929	0.0147	0.9926	0.9926	0.9926	0.9929	0.9858	3.54s	87,417,602
	ViT-S/32	1.0000	0.0070	1.0000	1.0000	1.0000	1.0000	1.0000	4.15s	22,475,138
	ViT-B/16	1.0000	0.0001	1.0000	1.0000	1.0000	1.0000	1.0000	7.41s	86,417,594
	ViT-S/16	0.9965	0.0075	1.0000	0.9926	0.9963	0.9966	0.9929	4.31s	21,590,402
	R26+S/32	1.0000	0.0059	1.0000	1.0000	1.0000	1.0000	1.0000	8.16s	36,028,098
Knowledge Distillation	ViT-B/32 student	0.9787	0.0522	0.9852	0.9708	0.9779	0.9790	0.9575	1.53s	2,300,162
	ViT-S/32 student	0.9645	0.1216	0.9926	0.9371	0.9640	0.9657	0.9306	1.04s	2,300,162
	ViT-B/16 student	0.9397	0.4739	1.0000	0.8882	0.9408	0.9422	0.8863	2.10s	3,442,274
	ViT-S/16 student	0.9894	0.0406	0.9852	0.9925	0.9888	0.9892	0.9787	1.61s	3,442,274
	R26+S/32 student	0.9858	0.0508	0.9778	0.9925	0.9851	0.9855	0.9717	1.55s	2,300,162
DeiT	ViT-B/32 student	0.9255	0.1920	0.8519	0.9914	0.9163	0.9225	0.8579	1.60s	3,554,884
	ViT-S/32 student	0.9362	0.1899	0.9037	0.9606	0.9313	0.9348	0.8732	1.15s	3,554,884
	ViT-B/16 student	0.9149	0.2850	0.8370	0.9826	0.9040	0.9117	0.8370	1.32s	3,112,516
	ViT-S/16 student	0.9255	0.2518	0.8667	0.9750	0.9176	0.9231	0.8550	1.40s	3,112,516
	R26+S/32 student	0.9078	0.2604	0.8296	0.9739	0.8960	0.9046	0.8226	1.59s	3,554,884

Table 5.12: Results of the ViTs in experiment 3.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	ViT-B/32	1.0000	0.0040	1.0000	1.0000	1.0000	1.0000	1.0000	4.88s	87,417,602
	ViT-S/32	1.0000	0.0012	1.0000	1.0000	1.0000	1.0000	1.0000	2.85s	22,475,138
	ViT-B/16	1.0000	0.0001	1.0000	1.0000	1.0000	1.0000	1.0000	7.33s	86,417,594
	ViT-S/16	1.0000	0.0021	1.0000	1.0000	1.0000	1.0000	1.0000	3.64s	21,590,402
	R26+S/32	1.0000	0.0029	1.0000	1.0000	1.0000	1.0000	1.0000	6.76s	36,028,098
Knowledge Distillation	ViT-B/32 student	0.9926	0.0151	1.0000	0.9852	0.9925	0.9927	0.9853	1.77s	2,300,162
	ViT-S/32 student	0.9667	0.1389	1.0000	0.9366	0.9673	0.9672	0.9355	1.78s	2,300,162
	ViT-B/16 student	0.9889	0.0240	1.0000	0.9779	0.9888	0.9891	0.9780	1.04s	3,442,274
	ViT-S/16 student	1.0000	0.0046	1.0000	1.0000	1.0000	1.0000	1.0000	1.70s	3,442,274
	R26+S/32 student	0.9963	0.0064	1.0000	0.9925	0.9963	0.9964	0.9926	1.55s	2,300,162
DeiT	ViT-B/32 student	0.9704	0.0734	0.9774	0.9630	0.9701	0.9705	0.9408	1.06s	3,554,884
	ViT-S/32 student	0.9889	0.0287	0.9774	1.0000	0.9886	0.9887	0.9780	1.07s	3,554,884
	ViT-B/16 student	0.9704	0.0831	0.9774	0.9630	0.9701	0.9705	0.9408	1.55s	3,112,516
	ViT-S/16 student	0.9926	0.0216	0.9925	0.9925	0.9925	0.9926	0.9852	1.27s	3,112,516
	R26+S/32 student	0.9889	0.0376	0.9774	1.0000	0.9886	0.9887	0.9780	1.08s	3,554,884

Table 5.13: Results of the ViTs in experiment 4.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	ViT-B/32	1.0000	0.0011	1.0000	1.0000	1.0000	1.0000	1.0000	3.07s	87,417,602
	ViT-S/32	1.0000	0.0007	1.0000	1.0000	1.0000	1.0000	1.0000	2.63s	22,475,138
	ViT-B/16	1.0000	0.0029	1.0000	1.0000	1.0000	1.0000	1.0000	5.25s	86,417,594
	ViT-S/16	0.9953	0.0097	1.0000	0.9878	0.9939	0.9962	0.9901	3.26s	21,590,402
	R26+S/32	0.9953	0.0169	1.0000	0.9878	0.9939	0.9962	0.9901	5.88s	36,028,098
Knowledge Distillation	ViT-B/32 student	0.9624	0.1138	0.9012	1.0000	0.9481	0.9506	0.9218	1.81s	2,300,162
	ViT-S/32 student	0.9859	0.0251	1.0000	0.9643	0.9818	0.9886	0.9708	1.31s	2,300,162
	ViT-B/16 student	0.9953	0.0132	1.0000	0.9878	0.9939	0.9962	0.9901	1.74s	3,442,274
	ViT-S/16 student	0.9906	0.0147	0.9753	1.0000	0.9875	0.9877	0.9802	0.91s	3,442,274
	R26+S/32 student	0.9953	0.0097	1.0000	0.9878	0.9939	0.9962	0.9901	1.42s	2,300,162
DeiT	ViT-B/32 student	0.9671	0.0636	0.9506	0.9625	0.9565	0.9639	0.9302	0.93s	3,554,884
	ViT-S/32 student	0.9671	0.0983	0.9136	1.0000	0.9548	0.9568	0.9314	0.91s	3,554,884
	ViT-B/16 student	0.9577	0.1023	0.8889	1.0000	0.9412	0.9444	0.9122	1.03s	3,112,516
	ViT-S/16 student	0.9624	0.0849	0.9012	1.0000	0.9481	0.9506	0.9218	1.01s	3,112,516
	R26+S/32 student	0.9531	0.1419	0.8765	1.0000	0.9342	0.9383	0.9027	0.92s	3,554,884

Table 5.14 presents the results of experiment 5 by the ViT models. Based on the obtained results, it is perceived by the value of the AUC metric, that the architectures present a good ability to distinguish the images of the classes under study. Comparing the distillation results obtained by CNNs and ViTs, it is concluded that the students of CNNs can perform better in classifying images of inflammatory bowel disease. When comparing the results between the two distillation techniques of the ViT models, it can be seen that the results worsened in the distillation through attention, having for example the model ViT-S/16 reduced 4.96% in Accuracy, 5.19% in AUC and 9.67% in MCC metric. But inference time was shorter using the DeiT distillation technique.

Table 5.14: Results of the ViTs in experiment 5.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	ViT-B/32	0.9965	0.0116	0.9926	1.0000	0.9963	0.9963	0.9929	3.42s	87,417,602
	ViT-S/32	0.9965	0.0070	1.0000	0.9926	0.9963	0.9966	0.9929	2.87s	22,475,138
	ViT-B/16	1.0000	0.0004	1.0000	1.0000	1.0000	1.0000	1.0000	6.49s	86,417,594
	ViT-S/16	1.0000	0.0065	1.0000	1.0000	1.0000	1.0000	1.0000	3.67s	21,590,402
	R26+S/32	1.0000	0.0042	1.0000	1.0000	1.0000	1.0000	1.0000	6.45s	36,028,098
Knowledge Distillation	ViT-B/32 student	0.9539	0.1103	0.9704	0.9357	0.9527	0.9546	0.9083	1.76s	2,300,162
	ViT-S/32 student	0.9574	0.1615	0.9111	1.0000	0.9535	0.9556	0.9178	1.00s	2,300,162
	ViT-B/16 student	0.9894	0.0413	0.9778	1.0000	0.9888	0.9889	0.9789	1.22s	3,442,274
	ViT-S/16 student	0.9929	0.0376	1.0000	0.9854	0.9926	0.9932	0.9859	1.04s	3,442,274
	R26+S/32 student	0.9504	0.2228	0.9926	0.9116	0.9504	0.9521	0.9042	1.06s	2,300,162
DeiT	ViT-B/32 student	0.9362	0.1776	0.8963	0.9680	0.9308	0.9345	0.8739	1.08s	3,554,884
	ViT-S/32 student	0.9468	0.1330	0.9111	0.9762	0.9425	0.9454	0.8950	1.10s	3,554,884
	ViT-B/16 student	0.9397	0.2261	0.8815	0.9917	0.9333	0.9373	0.8837	1.16s	3,112,516
	ViT-S/16 student	0.9433	0.2013	0.8963	0.9837	0.9380	0.9413	0.8892	1.16s	3,112,516
	R26+S/32 student	0.9291	0.2303	0.8667	0.9832	0.9213	0.9265	0.8629	1.04s	3,554,884

Tables [5.15 - 5.18] present the results gathered from the same experiments by the ViT teacher and student models. The results represented in the tables express a behaviour similar to the one described during experiment 5, except that during experiment 6 the model R26+S/32 had a performance improvement with the distillation of knowledge through attention. The same was described by models ViT-B/32 and ViT-S/32 in experiment 8.

Table 5.15: Results of the ViTs in experiment 6.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	ViT-B/32	1.0000	0.0003	1.0000	1.0000	1.0000	1.0000	1.0000	3.21s	87,417,602
	ViT-S/32	1.0000	0.0040	1.0000	1.0000	1.0000	1.0000	1.0000	2.91s	22,475,138
	ViT-B/16	1.0000	0.0035	1.0000	1.0000	1.0000	1.0000	1.0000	5.62s	86,417,594
	ViT-S/16	1.0000	0.0008	1.0000	1.0000	1.0000	1.0000	1.0000	3.33s	21,590,402
	R26+S/32	1.0000	0.0033	1.0000	1.0000	1.0000	1.0000	1.0000	6.29s	36,028,098
Knowledge Distillation	ViT-B/32 student	0.9953	0.0038	0.9877	1.0000	0.9938	0.9938	0.9901	1.70s	2,300,162
	ViT-S/32 student	0.9906	0.0395	0.9753	1.0000	0.9875	0.9877	0.9802	1.58s	2,300,162
	ViT-B/16 student	0.9906	0.0509	0.9753	1.0000	0.9875	0.9877	0.9802	1.15s	3,442,274
	ViT-S/16 student	0.9906	0.0399	0.9753	1.0000	0.9875	0.9877	0.9802	1.65s	3,442,274
	R26+S/32 student	0.9061	0.3198	0.7531	1.0000	0.8592	0.8765	0.8087	0.97s	2,300,162
DeiT	ViT-B/32 student	0.9671	0.1054	0.9136	1.0000	0.9548	0.9568	0.9314	1.02s	3,554,884
	ViT-S/32 student	0.9718	0.0793	0.9383	0.9870	0.9620	0.9653	0.9404	1.28s	3,554,884
	ViT-B/16 student	0.9718	0.0738	0.9259	1.0000	0.9615	0.9630	0.9411	1.54s	3,112,516
	ViT-S/16 student	0.9718	0.0964	0.9259	1.0000	0.9615	0.9630	0.9411	1.10s	3,112,516
	R26+S/32 student	0.9671	0.0660	0.9383	0.9744	0.9560	0.9616	0.9302	1.03s	3,554,884

Table 5.16: Results of the ViTs in experiment 7.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	ViT-B/32	1.0000	0.0028	1.0000	1.0000	1.0000	1.0000	1.0000	4.52s	87,417,602
	ViT-S/32	0.9965	0.0051	1.0000	0.9926	0.9963	0.9966	0.9929	2.85s	22,475,138
	ViT-B/16	1.0000	0.0019	1.0000	1.0000	1.0000	1.0000	1.0000	7.72s	86,417,594
	ViT-S/16	1.0000	0.0018	1.0000	1.0000	1.0000	1.0000	1.0000	4.87s	21,590,402
	R26+S/32	0.9787	0.0480	1.0000	0.9574	0.9783	0.9796	0.9583	8.11s	36,028,098
Knowledge Distillation	ViT-B/32 student	0.9716	0.0697	0.9630	0.9774	0.9701	0.9713	0.9432	1.53s	2,300,162
	ViT-S/32 student	0.8794	0.6815	0.9926	0.8024	0.8874	0.8841	0.7808	1.35s	2,300,162
	ViT-B/16 student	0.9645	0.1487	0.9259	1.0000	0.9615	0.9630	0.9311	1.42	3,442,274
	ViT-S/16 student	0.9787	0.0924	0.9556	1.0000	0.9773	0.9778	0.9582	1.33s	3,442,274
	R26+S/32 student	0.9645	0.0733	0.9333	0.9921	0.9618	0.9633	0.9303	1.56s	2,300,162
DeiT	ViT-B/32 student	0.9574	0.1114	0.9333	0.9767	0.9545	0.9565	0.9154	1.14s	3,554,884
	ViT-S/32 student	0.9468	0.1741	0.9037	0.9839	0.9421	0.9450	0.8958	1.10s	3,554,884
	ViT-B/16 student	0.9433	0.2008	0.8963	0.9837	0.9380	0.9413	0.8892	1.37s	3,112,516
	ViT-S/16 student	0.9397	0.1884	0.8963	0.9758	0.9344	0.9379	0.8815	1.31s	3,112,516
	R26+S/32 student	0.9397	0.1539	0.9111	0.9609	0.9354	0.9385	0.8801	1.11s	3,554,884

Table 5.17: Results of the ViTs in experiment 8.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	ViT-B/32	1.0000	0.0024	1.0000	1.0000	1.0000	1.0000	1.0000	3.11s	87,417,602
	ViT-S/32	1.0000	0.0031	1.0000	1.0000	1.0000	1.0000	1.0000	2.76s	22,475,138
	ViT-B/16	1.0000	0.0012	1.0000	1.0000	1.0000	1.0000	1.0000	5.11s	86,417,594
	ViT-S/16	1.0000	0.0017	1.0000	1.0000	1.0000	1.0000	1.0000	3.26s	21,590,402
	R26+S/32	1.0000	0.0007	1.0000	1.0000	1.0000	1.0000	1.0000	5.86s	36,028,098
Knowledge Distillation	ViT-B/32 student	0.8779	0.4485	0.6790	1.0000	0.8088	0.8395	0.7532	2.66s	2,300,162
	ViT-S/32 student	0.9484	0.1526	0.8642	1.0000	0.9272	0.9321	0.8932	1.18s	2,300,162
	ViT-B/16 student	1.0000	0.0011	1.0000	1.0000	1.0000	1.0000	1.0000	1.76s	3,442,274
	ViT-S/16 student	1.0000	0.0035	1.0000	1.0000	1.0000	1.0000	1.0000	1.81s	3,442,274
	R26+S/32 student	0.8404	0.7652	0.5802	1.0000	0.7344	0.7901	0.6793	1.66s	2,300,162
DeiT	ViT-B/32 student	0.9484	0.1210	0.8765	0.9861	0.9281	0.9345	0.8918	0.97s	3,554,884
	ViT-S/32 student	0.9531	0.1131	0.8765	1.0000	0.9342	0.9383	0.9027	1.00s	3,554,884
	ViT-B/16 student	0.8779	0.3492	0.6790	1.0000	0.8088	0.8395	0.7532	1.53s	3,112,516
	ViT-S/16 student	0.9249	0.2402	0.8025	1.0000	0.8904	0.9012	0.8460	1.12s	3,112,516
	R26+S/32 student	0.9577	0.0981	0.9012	0.9865	0.9419	0.9468	0.9111	1.29s	3,554,884

Table 5.18: Results of the ViTs in experiment 9.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	ViT-B/32	1.0000	0.0025	1.0000	1.0000	1.0000	1.0000	1.0000	3.46s	87,417,602
	ViT-S/32	1.0000	0.0048	1.0000	1.0000	1.0000	1.0000	1.0000	4.00s	22,475,138
	ViT-B/16	1.0000	0.0006	1.0000	1.0000	1.0000	1.0000	1.0000	7.26s	86,417,594
	ViT-S/16	1.0000	0.0013	1.0000	1.0000	1.0000	1.0000	1.0000	4.80s	21,590,402
	R26+S/32	0.9965	0.0195	1.0000	0.9926	0.9963	0.9966	0.9929	6.74s	36,028,098
Knowledge Distillation	ViT-B/32 student	0.9681	0.1090	0.93333	1.0000	0.9655	0.9667	0.9378	1.45s	2,300,162
	ViT-S/32 student	0.9787	0.0897	0.9926	0.9640	0.9781	0.9793	0.9578	1.44s	2,300,162
	ViT-B/16 student	0.9716	0.0937	0.9778	0.9635	0.9706	0.9719	0.9433	1.94s	3,442,274
	ViT-S/16 student	0.9929	0.0215	0.9926	0.9926	0.9926	0.9929	0.9858	1.44s	3,442,274
	R26+S/32 student	0.9823	0.0506	0.9630	1.0000	0.9811	0.9815	0.9650	2.30s	2,300,162
DeiT	ViT-B/32 student	0.9681	0.1108	0.9481	0.9846	0.9660	0.9673	0.9366	1.46s	3,554,884
	ViT-S/32 student	0.9397	0.1967	0.8889	0.9836	0.9339	0.9376	0.8825	1.46s	3,554,884
	ViT-B/16 student	0.9255	0.2187	0.8741	0.9672	0.9183	0.9234	0.8539	1.68s	3,112,516
	ViT-S/16 student	0.9007	0.3263	0.8222	0.9652	0.8880	0.8975	0.8081	1.20s	3,112,516
	R26+S/32 student	0.9327	0.1409	0.9037	0.9683	0.9349	0.9382	0.8807	1.49s	3,554,884

6 INTERPRETABILITY

In addition to quantitatively evaluating the performance of the models used, it is important in this work to also qualitatively evaluate their performance through interpretability models. In order to understand which parts of the mucosa the algorithm used as a basis to make a particular decision. Therefore, this chapter presents the images resulting from the interpretability of the teacher models. The interpretability of the student models of ViTs and CNNs can be found in Appendix B - Supplementary material.

To make the interpretability of the models used in this work, six images were randomly selected during the creation of the datasets, referring to the classes Crohn's disease and Ulcerative Colitis.

Figures [6.1 - 6.9] present the resulting images of the interpretability of all teacher models, relative to the CNNs architectures in all experiments. In these images, it is also possible to observe which was the label of the original image and which was the predicted label by CNN. In this work, ViT models were also interpreted to understand which parts of the gut mucosa were more important for these architectures. On the other hand, it allows to compare with CNN architectures what information was learned by both architectures. Figures [6.10 - 6.18] present the interpretability done during all the experiments concerning the teacher models of the ViTs.

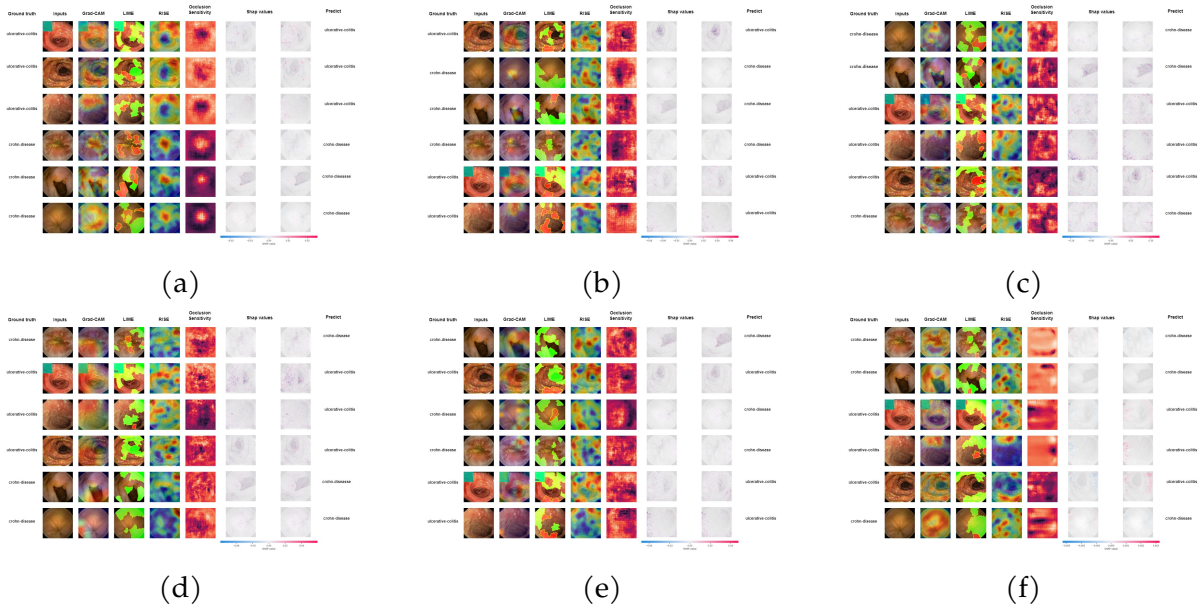


Figure 6.1: Interpretability of the CNNs teacher models in experiment 1: (a) AlexNet; (b) ResNet50; (c) VGG16; (d) InceptionV3; (e) Ensemble method; (f) Hybrid model.

Diagnose of the IBD using Deep Learning Models with Interpretability

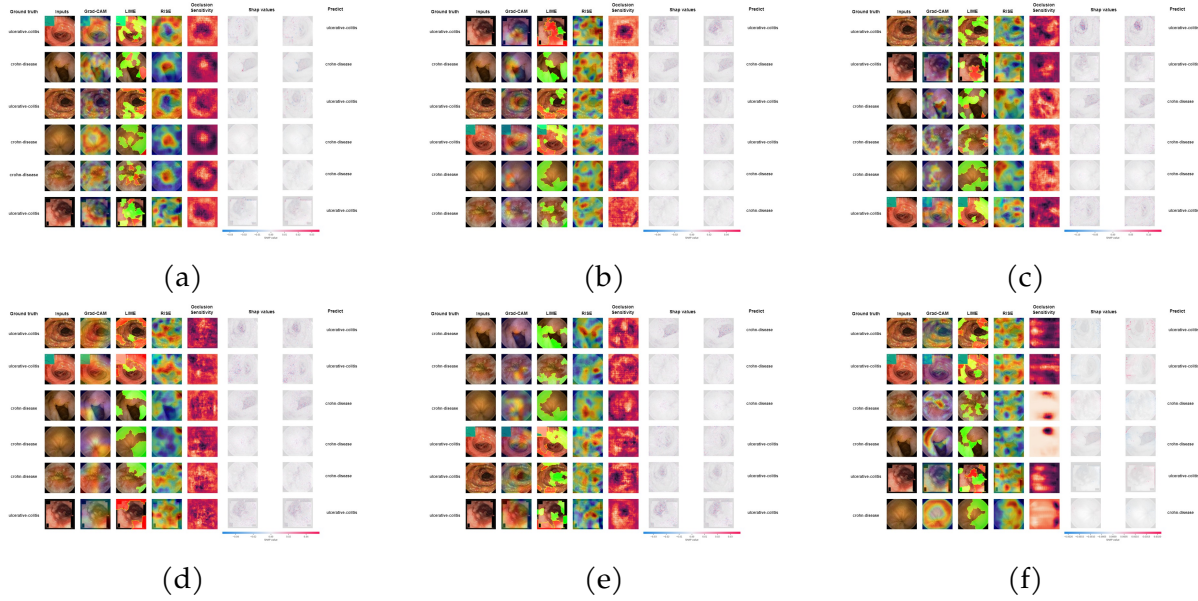


Figure 6.2: Interpretability of the CNNs teacher models in experiment 2: (a) AlexNet; (b) ResNet50; (c) VGG16; (d) InceptionV3; (e) Ensemble method; (f) Hybrid model.

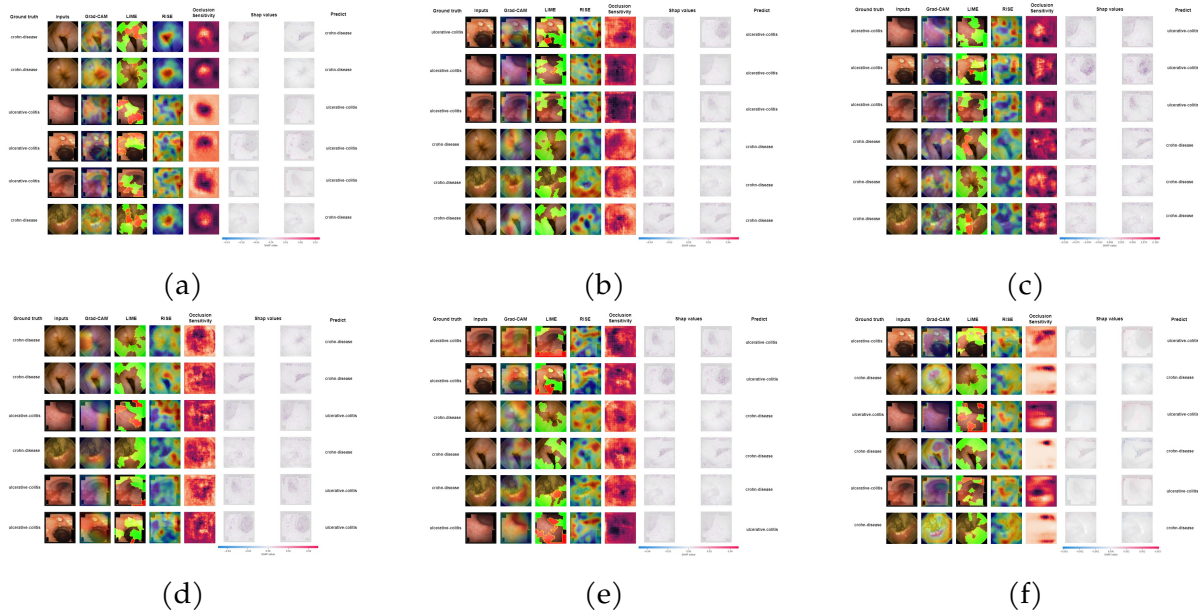


Figure 6.3: Interpretability of the CNNs teacher models in experiment 3: (a) AlexNet; (b) ResNet50; (c) VGG16; (d) InceptionV3; (e) Ensemble method; (f) Hybrid model.

In experiments 1 and 2 one can visualise through the interpretability images that the selected models have a tendency to rely on elements external to the intestinal mucosa to make the prediction of classes. This turns out to be wrong, considering that the green box does not indicate any useful information about ulcerative colitis disease. However, in the implementation of CNNs it is possible to realize through the algorithms Occlusion sensitivity and RISE identify that the networks ResNet50 and InceptionV3 have the ability to identify the lesions caused by Crohn's disease, and can identify eczema and ulcers in ulcerative colitis disease, but in the same experiments analyzing the in-

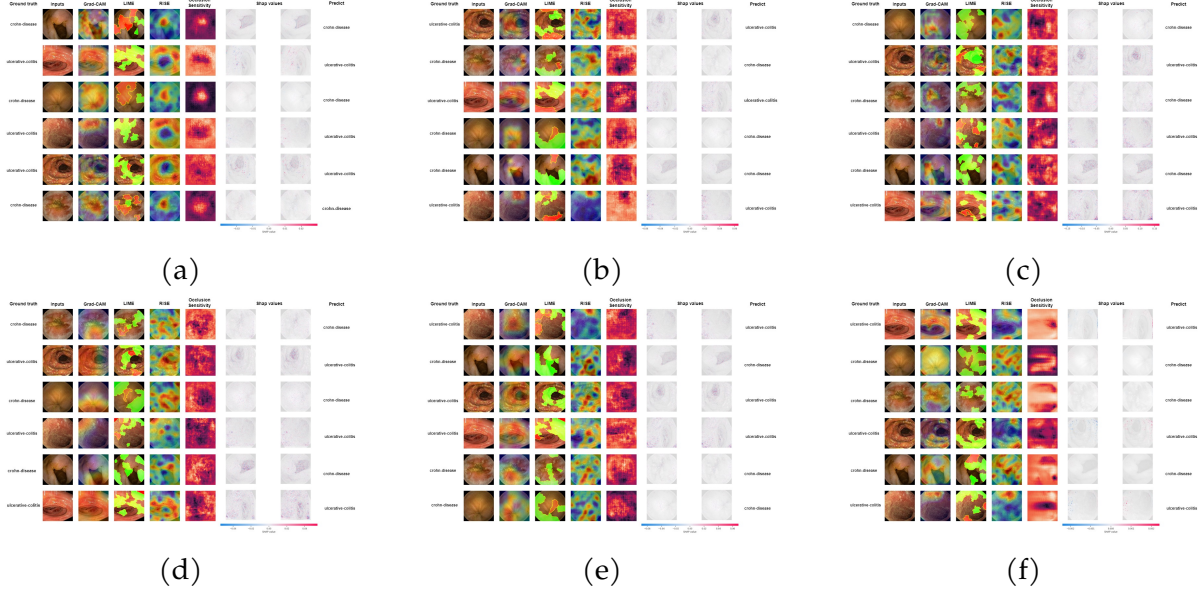


Figure 6.4: Interpretability of the CNNs teacher models in experiment 4: (a) AlexNet; (b) ResNet50; (c) VGG16; (d) InceptionV3; (e) Ensemble method; (f) Hybrid model.

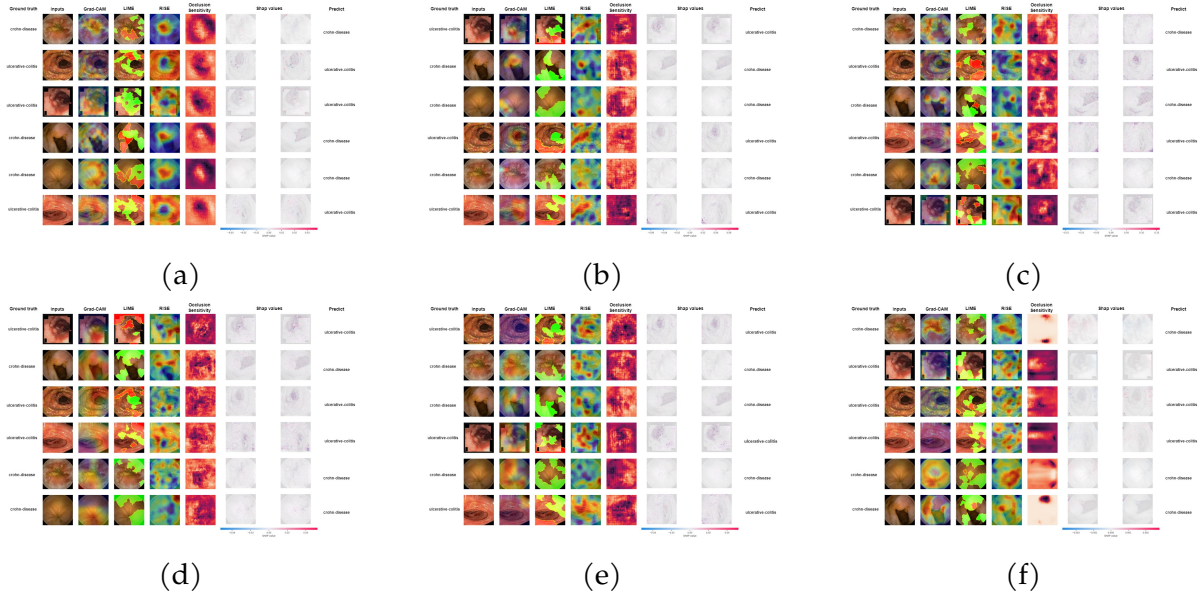


Figure 6.5: Interpretability of the CNNs teacher models in experiment 5: (a) AlexNet; (b) ResNet50; (c) VGG16; (d) InceptionV3; (e) Ensemble method; (f) Hybrid model.

formation presented by the same algorithms, it is identified that the ViTs are not able to identify the same features for the same diseases, despite having a good performance.

In contrast to these experiments, it is verified in Figure 6.12 that the ViTs have the ability to recognize bleeding situations in ulcerative colitis disease. While the implementation with CNNs was not possible to recognize these types of features linked to the disease. In all the experiments the Hybrid algorithm (CNN+LSTM) presents some difficulty in recognizing the abnormalities present in the intestinal mucosa linked to the pathologies.

Diagnose of the IBD using Deep Learning Models with Interpretability

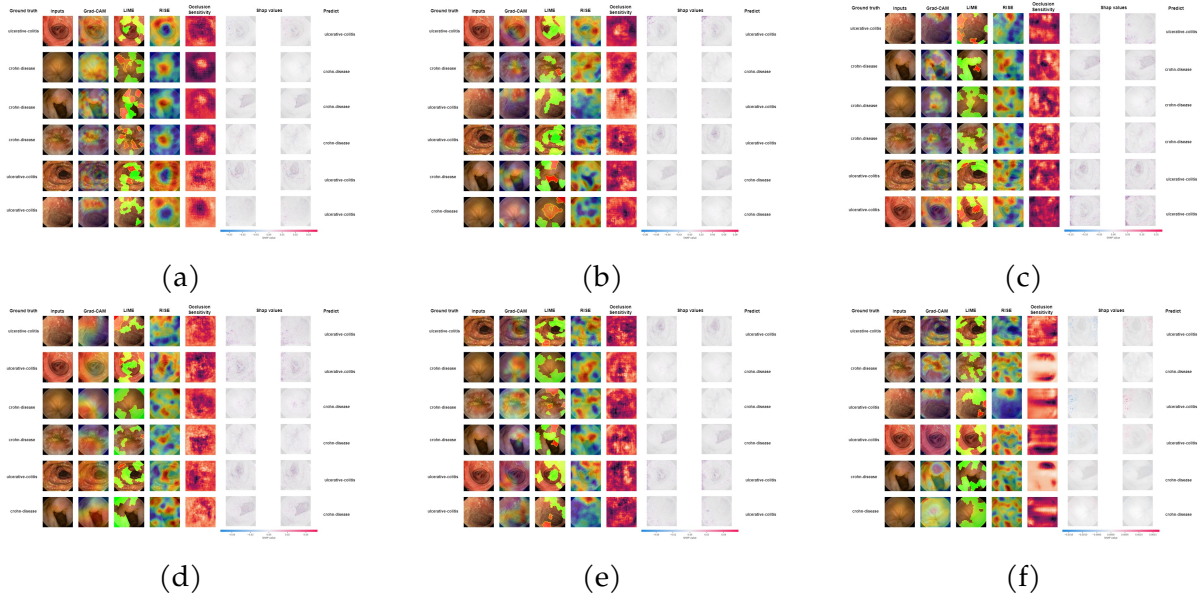


Figure 6.6: Interpretability of the CNNs teacher models in experiment 6: (a) AlexNet; (b) ResNet50; (c) VGG16; (d) InceptionV3; (e) Ensemble method; (f) Hybrid model.

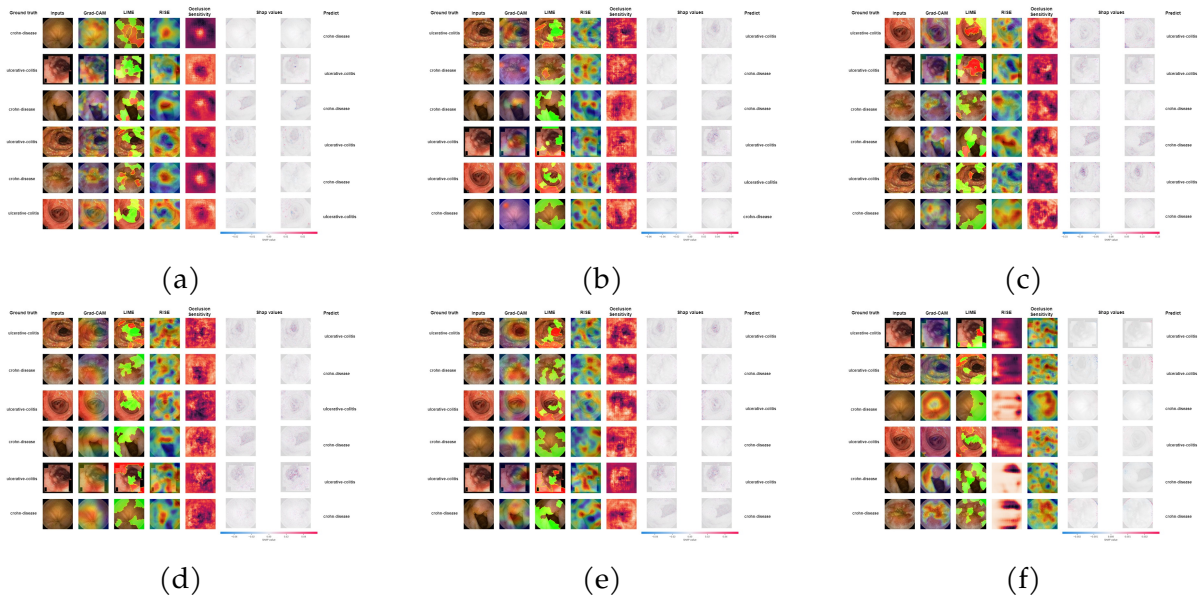


Figure 6.7: Interpretability of the CNNs teacher models in experiment 7: (a) AlexNet; (b) ResNet50; (c) VGG16; (d) InceptionV3; (e) Ensemble method; (f) Hybrid model.

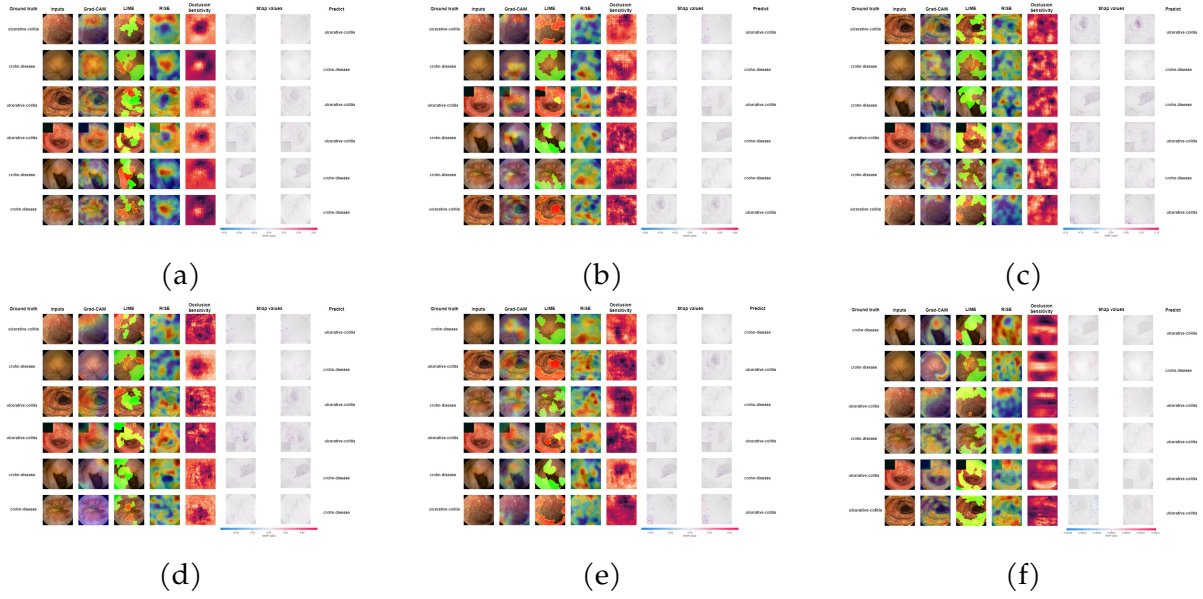


Figure 6.8: Interpretability of the CNNs teacher models in experiment 8: (a) AlexNet; (b) ResNet50; (c) VGG16; (d) InceptionV3; (e) Ensemble method; (f) Hybrid model.

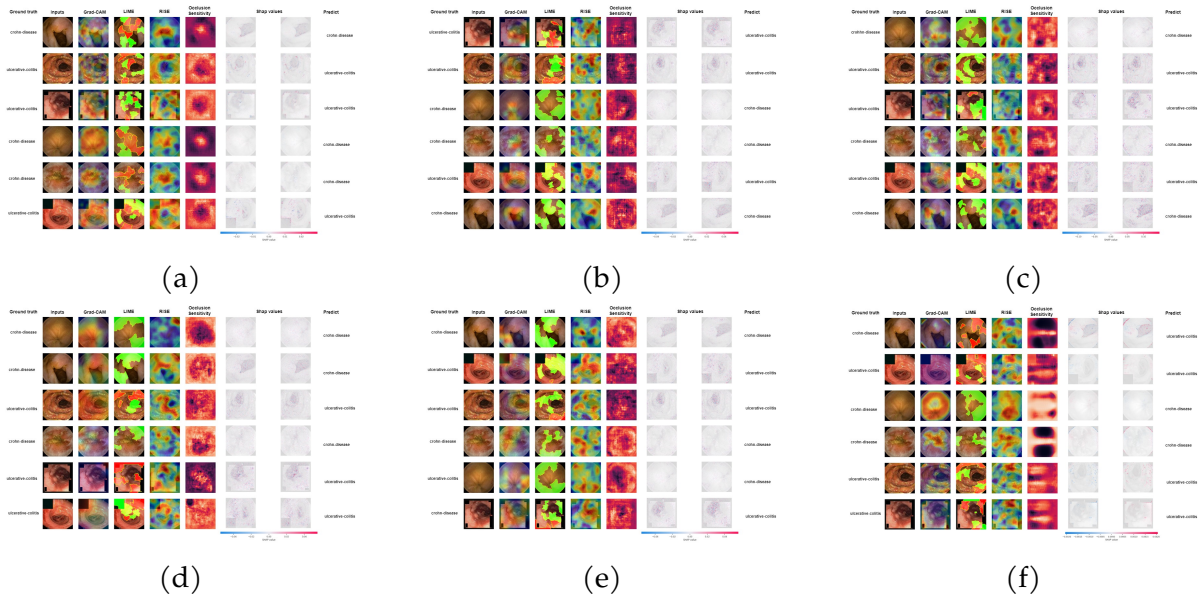


Figure 6.9: Interpretability of the CNNs teacher models in experiment 9: (a) AlexNet; (b) ResNet50; (c) VGG16; (d) InceptionV3; (e) Ensemble method; (f) Hybrid model.

Diagnose of the IBD using Deep Learning Models with Interpretability

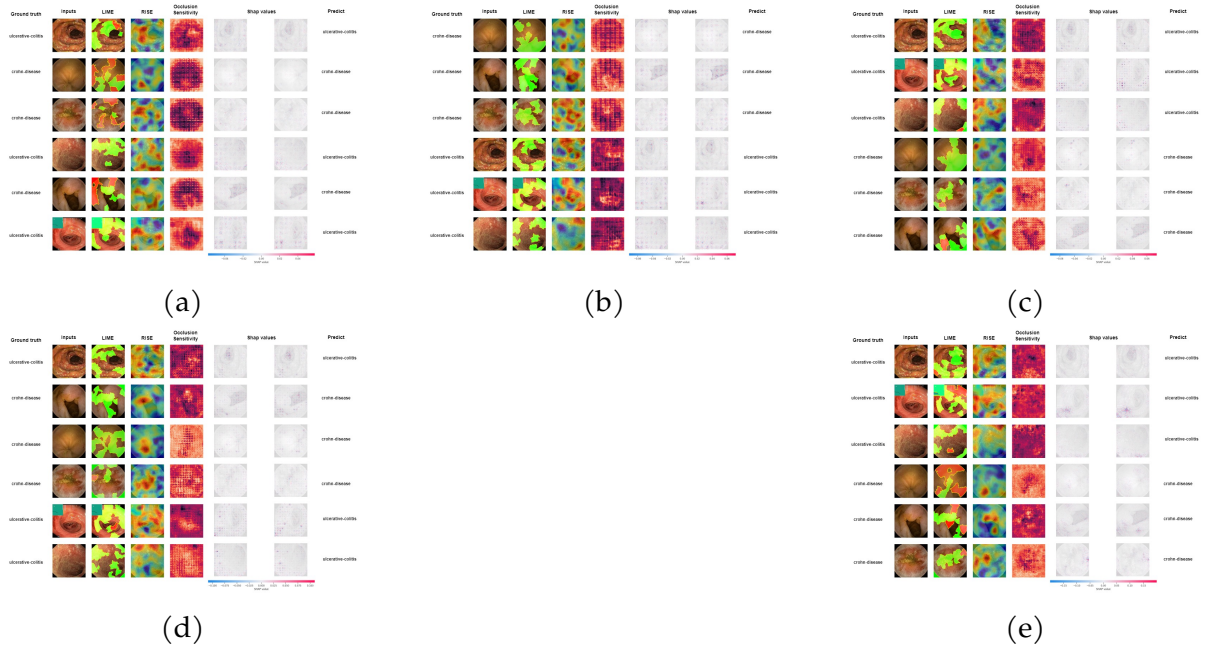


Figure 6.10: Interpretability of the ViTs teacher models in experiment 1: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.

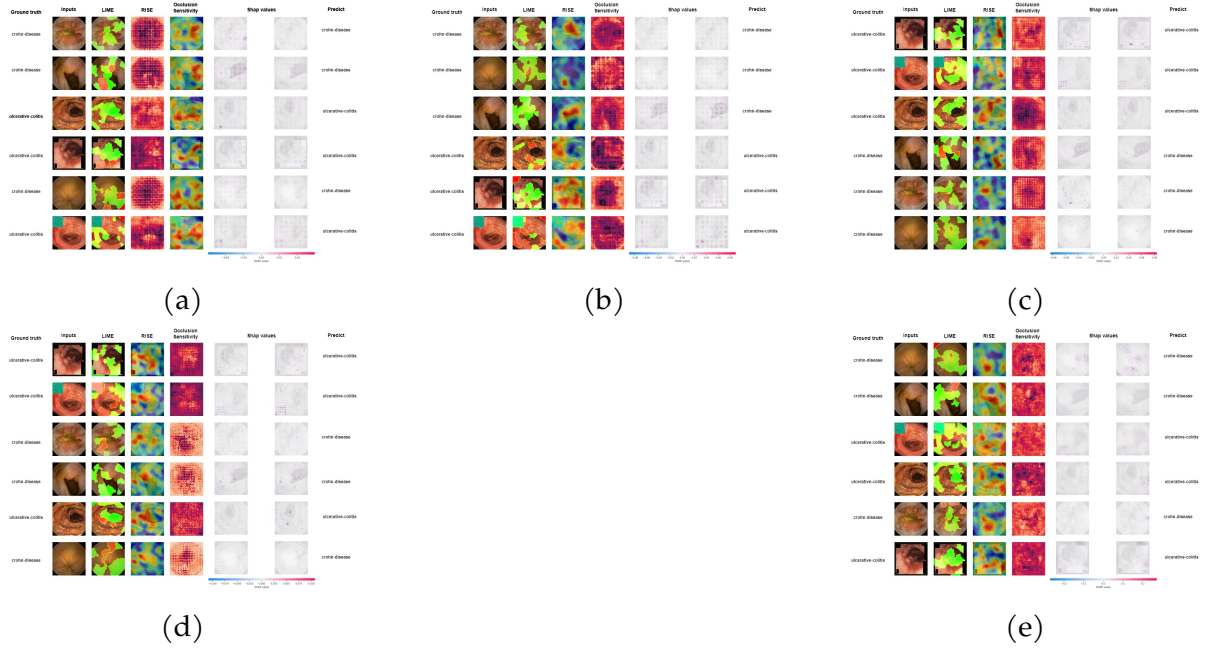


Figure 6.11: Interpretability of the ViTs teacher models in experiment 2: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.



Figure 6.12: Interpretability of the ViTs teacher models in experiment 3: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.



Figure 6.13: Interpretability of the ViTs teacher models in experiment 4: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.

Diagnose of the IBD using Deep Learning Models with Interpretability

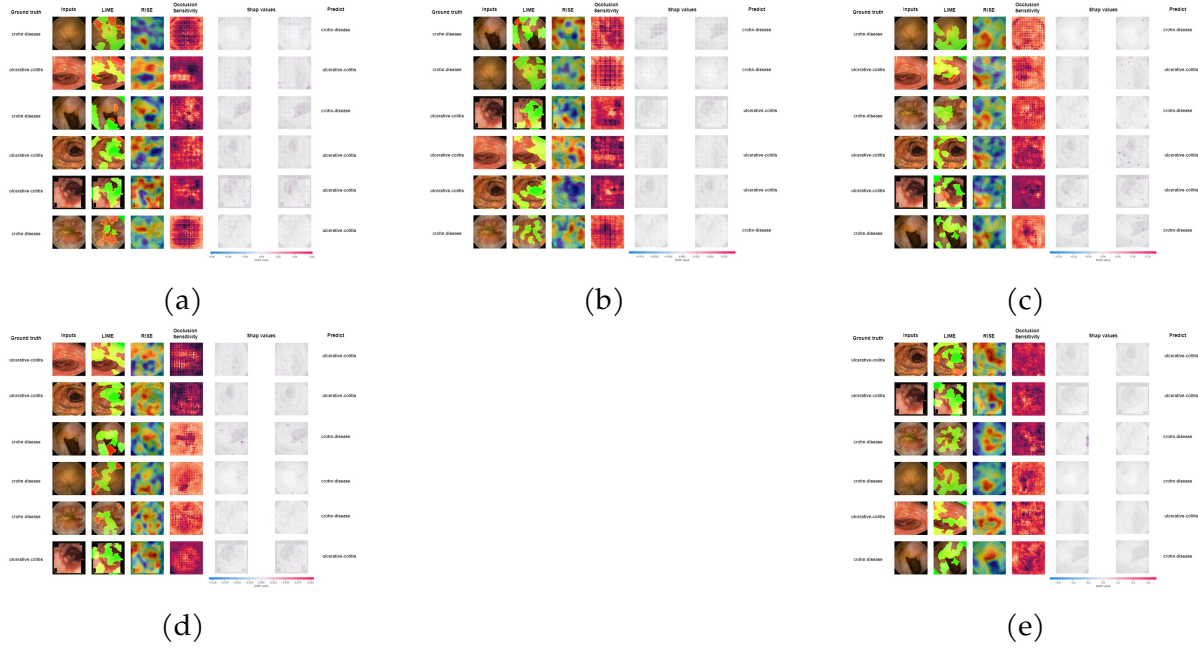


Figure 6.14: Interpretability of the ViTs teacher models in experiment 5: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.



Figure 6.15: Interpretability of the ViTs teacher models in experiment 6: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.



Figure 6.16: Interpretability of the ViTs teacher models in experiment 7: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.



Figure 6.17: Interpretability of the ViTs teacher models in experiment 8: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.

Diagnose of the IBD using Deep Learning Models with Interpretability

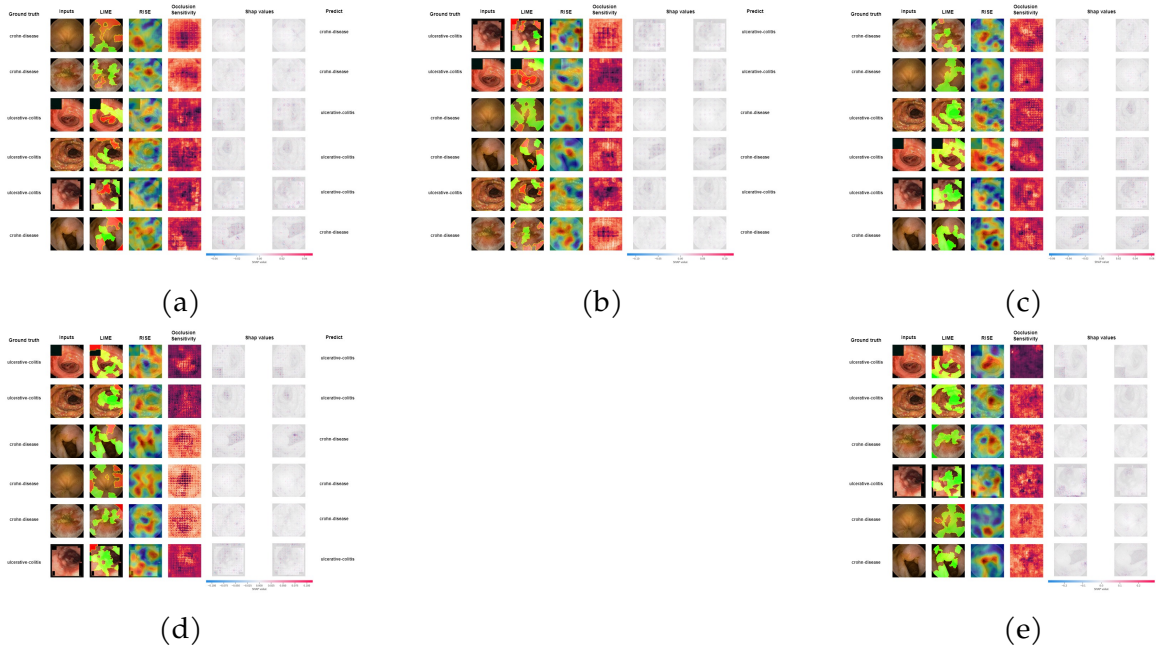


Figure 6.18: Interpretability of the ViTs teacher models in experiment 9: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.

7 DISCUSSION

This study serves as a reference for hospitals and gastroenterology doctors to automate the process of diagnosing patients with inflammatory bowel disease. The AlexNet network in experiment 1 was the algorithm that achieved the shortest inference time (0.75 seconds), with a good performance: 100% Accuracy, 100% Recall, 100% Precision, 100% F1-Score, 100% AUC, 100% F1-Score and 0.0031 Loss. However, due to the self-attention mechanism and the performance improvement that ViTs demonstrate over CNNs [7], ViT-S/16 with knowledge distillation could be a very promising alternative with: 99.06% accuracy, 97.53% recall, 100% precision, 98.75% f1-Score, 98.77% AUC, 98.02% MCC, 0.0147 Loss and took 0.91 seconds to infer the test results.

The models used in this study prove to be accurate in the diagnosis of inflammatory bowel disease, and through the analysis of the AUC and MCC metrics, the models show a good capacity to distinguish the classes, and the number of false positives and false negatives is very close to zero or even manages to be zero, respectively. The results obtained are superior to those presented by the authors [83] who used a ResNetXt-101 network. Furthermore, if we compare with the results presented in the same study corresponding to the accuracy of clinicians in diagnosing the disease, the accuracy of the models used manages to be much higher. However, the number of images used by the authors was higher than the one used in this study, as well as the images were collected using the same endoscopic tool.

Another study also developed an investigation where it compares the results of deep neural networks with clinicians in distinguishing Crohn's disease and Ulcerative colitis. The authors highlighted in their study the significant improvement in the accuracy of the developed neural network compared to the clinicians' accuracy in distinguishing IBD disease per patient and per lesion [82]. Analysing the data per lesion it is found that the results presented by the selected algorithms manage to be much higher than the clinicians' accuracy, as well as the CNNs used in this study demonstrated an improvement in performance in making the distinction between Crohn's disease and ulcerative colitis. On the other hand, the authors [80] in the study they developed seek to distinguish between ulcerative colitis and Crohn's disease, between ulcerative colitis and normal, and between the presence of IBD and normal. In this study, the authors developed an ensemble method that combines a ResNet50, a ResNet34 and a ResNet152, that when analysing the result obtained by the MCC metric in the distinction between Ulcerative colitis and Crohn's disease our ensemble method (ResNet50+MobileNetV2) describes an improvement in performance in making the same classification.

The authors [78] in their study proposed a hybrid model, where they combine a CNN with an LSTM to perform the confirmation of mucosa healing in Crohn's disease. The best result achieved by the authors was 95.3% accuracy, 92.78% sensitivity, 94.6% specificity and an AUCROC of 0.98. In this paper, a hybrid model is also presented, where two more LSTMs were used concerning what was proposed by the authors, among other modifications. These changes represented a significant improvement in the results, with the best results being 100% accuracy, 100% recall, 100% precision and 100% AUC.

The literature has focused on the implementation of deep learning models in image processing of Crohn's disease, or ulcerative colitis disease. In this sense, the authors in the works [91, 111] developed a Recurrent attention neural network in the detection of the type of lesion caused by inflammation in the mucosa of Crohn's disease. Although the study objective is different, the teacher and student models implemented in this work describe a performance improvement.

When analysing the quantitative results obtained by ViTs and the interpretability of the models, it can be seen that transformers manage to stand out from CNNs in situations where noisy elements are visible in the images, due to the self-attention mechanism that helps improve the importance of some features in the function of others. Allowing the algorithm to abstract from additional features beyond the central feature of the image [112]. This functionality was also described by other authors who developed an investigation in the health area, where they compared CNNs and ViTs in image processing and demonstrated the importance of the self-attention mechanism through the results obtained [64].

The implementation of knowledge distillation is one of the important contributions of this work so that it is possible to use these architectures in the diagnosis of IBD patients. Comparing the results between the different techniques used, applied to CNN networks and ViTs. It is concluded that knowledge distillation in ViTs allowed a greater reduction in inference time, preserving a good performance of the models. On the other hand, in ViTs it is still verified that in experiments 3 and 5 the hybrid model R26+S/32 obtained a performance improvement, using the distillation technique through the attention (DeiT) when compared with knowledge distillation. In experiment four, the same situation is verified in the ViT-S/32 model. The implementation of DeiT in this work represents a performance improvement when compared with the implementation in other studies [60, 113].

8 CONCLUSION

It is concluded with this work that ViT models and CNN networks show a good performance in the recognition of inflammatory bowel disease. It is considered that the knowledge distillation through attention in the ViTs models allowed in some experiments performed to obtain a lower inference time concerning the knowledge distillation through the logits. Furthermore, in some experiments, such as the model in experiment 6, it was found that there was an improvement in the performance of the student model in the knowledge distillation through attention [43].

It is also evident in this work through the interpretability of the models that ViTs due to self-attention mechanism manage to be better than CNNs in the prediction of images that have elements external to the intestinal mucosa [54]. When analysing experiment 9 it is evident that both teacher and student models can demonstrate good performance and identify Crohn's disease lesions, ulcers and erythema in ulcerative colitis disease, ignoring the additional elements present in the images.

8.1 Strengths

This work stands out for the comparison it offers between CNNs and ViTs architectures to perform endoscopic image classification regarding inflammatory bowel disease types. It also stands out by implementing Knowledge distillation from large architectures to simpler architectures. In ViTs, two Knowledge distillation approaches were implemented: (i) based on the attention of the architectures; (ii) based on the distillation of the weights through temperature to scale them.

The methods used to pre-process the images to reduce some features that might throw off the classifier when determining the type of inflammatory bowel disease are also worth mentioning. This paper further offers the interpretability of the models with the use of some model-agnostic methods, to better understand which parts of the image the architectures relied on for making a decision. Some of the results obtained from experiment 3 are documented in [9]. Refer to Appendix A - Articles, where the full paper can be accessed.

8.2 Limitations

However, in this study, it is recognized as a limitation in the number of images used to train the ViTs. Because these deep-learning models need a large number of images to be able to learn the information well [34]. Another limitation of this work is that the images were not collected based on the same endoscopic tool, which consequently the images are different and in the future may not demonstrate the same performance in processing colonoscopy images regarding Crohn's disease.

Another limitation of this work was the limited computational resources, that did not allow testing other more robust deep learning architectures to classify the images.

8.3 Future work

As future work, it is suggested to use Multi-task algorithms to make the classification of the type of lesion in the case of Crohn's disease. As well as, the degree of severity of Ulcerative colitis disease based on the Mayo score. In addition to predicting the type of inflammatory bowel disease that the patient has, it is also important to predict the degree of the lesion in order to adapt and prescribe the best treatment. For this, ordinal techniques can be leveraged [114, 115].

Other deep learning architectures like MLP-Mixer [116] could be implemented to do endoscopic image classification to be able to compare with the results obtained with CNN and ViT architectures. Also suggested as future work, is to test the best architectures with new endoscopic images to evaluate the robustness of the trained models. Furthermore, it is also suggested that in this validation there should be a medical verification to ensure the veracity of the results.

REFERENCES

- [1] Y.-Z. Zhang, “Inflammatory bowel disease: Pathogenesis,” *World Journal of Gastroenterology*, vol. 20, no. 1, p. 91, 2014.
- [2] S. C. Ng, H. Y. Shi, N. Hamidi, F. E. Underwood, W. Tang, E. I. Benchimol, R. Panaccione, S. Ghosh, J. C. Y. Wu, F. K. L. Chan, J. J. Y. Sung, and G. G. Kaplan, “Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies,” *The Lancet*, vol. 390, no. 10114, pp. 2769–2778, Dec. 2017.
- [3] T. Sairenji, K. L. Collins, and D. V. Evans, “An Update on Inflammatory Bowel Disease,” *Primary Care: Clinics in Office Practice*, vol. 44, no. 4, pp. 673–692, Dec. 2017.
- [4] S. Saeid Seyedian, F. Nokhostin, and M. Dargahi Malamir, “A review of the diagnosis, prevention, and treatment methods of inflammatory bowel disease,” *Journal of Medicine and Life*, vol. 12, no. 2, pp. 113–122, Apr. 2019.
- [5] M. J. Rosen, A. Dhawan, and S. A. Saeed, “Inflammatory Bowel Disease in Children and Adolescents,” *JAMA Pediatrics*, vol. 169, no. 11, p. 1053, Nov. 2015.
- [6] J. P. Amorim, I. Domingues, P. H. Abreu, and J. A. Santos, “Interpreting deep learning models for ordinal problems.” in *European Symposium on Artificial Neural Networks (ESANN)*, 2018, pp. 373–378.
- [7] J. Maurício, I. Domingues, and J. Bernardino, “Comparing vision transformers and convolutional neural networks for image classification: A literature review,” *Applied Sciences*, vol. 13, no. 9, 2023.
- [8] J. Maurício and I. Domingues, “Deep Neural Networks to distinguish between Crohn’s disease and Ulcerative colitis,” in *11th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)*, 2023.
- [9] J. Maurício and I. Domingues, “Interpretability of deep neural networks to diagnose inflammatory bowel disease,” in *29th edition of the Portuguese Conference on Pattern Recognition*, 2023 (to be submitted).
- [10] —, “Knowledge distillation of vision transformers and convolutional networks to predict inflammatory bowel disease,” in *26th Iberoamerican Congress on Pattern Recognition*, 2023 (to be submitted).

- [11] “Doença inflamatória do intestino | CUF,” <https://www.cuf.pt/saude-a-z/doenca-inflamatoria-do-intestino> (accessed Nov. 02, 2022).
- [12] J. Wehkamp, M. Götz, K. Herrlinger, W. Steurer, and E. F. Stange, “Inflammatory Bowel Disease: Crohn’s disease and ulcerative colitis,” *Deutsches Ärzteblatt international*, Feb. 2016.
- [13] Y. A. Ghouri, V. Tahan, and B. Shen, “Secondary causes of inflammatory bowel diseases,” *World Journal of Gastroenterology*, vol. 26, no. 28, pp. 3998–4017, Jul. 2020.
- [14] S. Flynn and S. Eisenstein, “Inflammatory Bowel Disease Presentation and Diagnosis,” *Surgical Clinics of North America*, vol. 99, no. 6, pp. 1051–1062, Dec. 2019.
- [15] B. Moazzami, K. Moazzami, and N. Rezaei, “Early onset inflammatory bowel disease: manifestations, genetics and diagnosis,” *The Turkish Journal of Pediatrics*, vol. 61, no. 5, p. 637, 2019.
- [16] “What is inflammatory bowel disease (IBD)? | IBD,” <https://www.cdc.gov/ibd/what-is-IBD.htm> (accessed Nov. 15, 2022).
- [17] “Upper GI Endoscopy | NIDDK,” <https://www.niddk.nih.gov/health-information/diagnostic-tests/upper-gi-endoscopy> (accessed Nov. 15, 2022).
- [18] D. Bell and F. Gaillard, “Inflammatory bowel disease,” in *Radiopaedia.org*. Radiopaedia.org, Aug. 2009.
- [19] “MRI to evaluate small bowel Crohn’s disease,” <https://www.mri.melbourne/mri/mri-crohns-disease-and-small-bowel-2/> (accessed July 02, 2023).
- [20] “MRI - Mayo Clinic,” <https://www.mayoclinic.org/tests-procedures/mri/about/pac-20384768> (accessed Nov. 15, 2022).
- [21] “Contrast Radiography | Conditions & Treatments | UT Southwestern Medical Center,” <http://utswmed.org/conditions-treatments/contrast-radiography/> (accessed May. 30, 2023).
- [22] “Computed Tomography (CT) Scan,” <https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/computed-tomography-ct-scan> (accessed Nov. 15, 2022).
- [23] J. Melson, G. Trikudanathan, B. K. Abu Dayyeh, M. S. Bhutani, V. Chandrasekhara, P. Jirapinyo, K. Krishnan, N. A. Kumta, R. Pannala, M. A. Parsi, A. Sethi, A. J. Trindade, R. R. Watson, J. T. Maple, and D. R. Lichtenstein, “Video capsule endoscopy,” *Gastrointestinal Endoscopy*, vol. 93, no. 4, pp. 784–796, Apr. 2021.

- [24] B. L. Rodrigues, M. C. Mazzaro, C. K. Nagasako, M. D. L. S. Ayrizono, J. J. Fagundes, and R. F. Leal, "Assessment of disease activity in inflammatory bowel diseases: Non-invasive biomarkers and endoscopic scores," *World Journal of Gastrointestinal Endoscopy*, vol. 12, no. 12, pp. 504–520, Dec. 2020.
- [25] K. W. Schroeder, W. J. Tremaine, and D. M. Ilstrup, "Coated Oral 5-Aminosalicylic Acid Therapy for Mildly to Moderately Active Ulcerative Colitis," *New England Journal of Medicine*, vol. 317, no. 26, pp. 1625–1629, Dec. 1987.
- [26] K. Ikeya, H. Hanai, K. Sugimoto, S. Osawa, S. Kawasaki, T. Iida, Y. Maruyama, and F. Watanabe, "The Ulcerative Colitis Endoscopic Index of Severity More Accurately Reflects Clinical Outcomes and Long-term Prognosis than the Mayo Endoscopic Score," *Journal of Crohn's and Colitis*, vol. 10, no. 3, pp. 286–295, Mar. 2016.
- [27] G. Polat, H. T. Kani, I. Ergenc, Y. O. Alahdab, A. Temizel, and O. Atug, "Labeled Images for Ulcerative Colitis (LIMUC) Dataset," 2022.
- [28] M. Ferrante, J. Colombel, W. J. Sandborn, W. Reinisch, G. J. Mantzaris, A. Kornbluth, D. Rachmilewitz, S. Lichtiger, G. R. D'Haens, C. J. Van Der Woude, S. Danese, R. H. Diamond, A. F. Oortwijn, K. L. Tang, M. Miller, F. Cornillie, and P. J. Rutgeerts, "Validation of Endoscopic Activity Scores in Patients With Crohn's Disease Based on a Post Hoc Analysis of Data From SONIC," *Gastroenterology*, vol. 145, no. 5, pp. 978–986.e5, Nov. 2013.
- [29] M. Daperno, G. D'Haens, G. Van Assche, F. Baert, P. Bulois, V. Maunoury, R. Sostegni, R. Rocca, A. Pera, and A. Gevers, "Development and validation of a new, simplified endoscopic activity score for Crohn's disease: the SES-CD," *Gastrointestinal Endoscopy*, vol. 60, no. 4, pp. 505–512, Oct. 2004.
- [30] V. Annese, M. Daperno, M. D. Rutter, A. Amiot, P. Bossuyt, J. East, M. Ferrante, M. Götz, K. H. Katsanos, R. Kießlich, I. Ordás, A. Repici, B. Rosa, S. Sebastian, T. Kucharzik, and R. Eliakim, "European evidence based consensus for endoscopy in inflammatory bowel disease," *Journal of Crohn's and Colitis*, vol. 7, no. 12, pp. 982–1018, Dec. 2013.
- [31] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," *ArXiv*, 2015.
- [32] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 International Conference on Engineering and Technology (ICET)*. IEEE, Aug. 2017, pp. 1–6.
- [33] S. Saha, "A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way," Nov. 2022, <https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53> (accessed

Jan. 08, 2023).

- [34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *International Conference on Learning Representations*, 2020.
- [35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *ArXiv*, 2018, publisher: arXiv Version Number: 2.
- [36] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," *ArXiv*, 2015, publisher: arXiv Version Number: 1.
- [37] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge Distillation: A Survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021.
- [38] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for Thin Deep Nets," *ArXiv*, 2014, publisher: arXiv Version Number: 4.
- [39] L. Yuan, F. E. H. Tay, G. Li, T. Wang, and J. Feng, "Revisiting Knowledge Distillation via Label Smoothing Regularization," *ArXiv*, 2019, publisher: arXiv Version Number: 3.
- [40] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep Mutual Learning," *ArXiv*, 2017, publisher: arXiv Version Number: 1.
- [41] D. Chen, J.-P. Mei, C. Wang, Y. Feng, and C. Chen, "Online Knowledge Distillation with Diverse Peers," *ArXiv*, 2019, publisher: arXiv Version Number: 2.
- [42] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation," *ArXiv*, 2019, publisher: arXiv Version Number: 1.
- [43] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," *Arxiv*, 2020, publisher: arXiv Version Number: 2.
- [44] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," *Neural Information Processing Systems*, vol. abs/1706.03762, 2017.
- [45] X. Li, H. Xiong, X. Li, X. Wu, X. Zhang, J. Liu, J. Bian, and D. Dou, "Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond," *Knowledge and Information Systems*, vol. 64, no. 12, pp. 3197–3234, Dec. 2022.
- [46] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-CAM: Why did you say that?" *ArXiv*, 2016.

- [47] “A Complete Guide to SHAP - SHAPley Additive exPlanations for Practitioners,” <https://analyticsindiamag.com/a-complete-guide-to-shap-shapley-additive-explanations-for-practitioners/> (accessed June 15, 2023).
- [48] “Deep Learning Model Interpretation Using SHAP,” <https://towardsdatascience.com/deep-learning-model-interpretation-using-shap-a21786e91d16> (accessed June 15, 2023).
- [49] “Basic XAI with LIME for CNN Models | by Sahil Ahuja | DataDriven-Investor,” <https://medium.datadriveninvestor.com/xai-with-lime-for-cnn-models-5560a486578> (accessed June 15, 2023).
- [50] V. Petsiuk, A. Das, and K. Saenko, “RISE: Randomized Input Sampling for Explanation of Black-box Models,” *ArXiv*, 2018.
- [51] M. D. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks,” in *Computer Vision – ECCV 2014*. Cham: Springer International Publishing, 2014, vol. 8689, pp. 818–833.
- [52] P. Benz, S. Ham, C. Zhang, A. Karjauv, and I. S. Kweon, “Adversarial Robustness Comparison of Vision Transformer and MLP-Mixer to CNNs,” *British Machine Vision Conference*, 2021.
- [53] Y. Bai, J. Mei, A. Yuille, and C. Xie, “Are Transformers More Robust Than CNNs?” *Neural Information Processing Systems*, 2021.
- [54] K. Tyagi, G. Pathak, R. Nijhawan, and A. Mittal, “Detecting Pneumonia using Vision Transformer and comparing with other techniques,” in *5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. Coimbatore, India: IEEE, Dec. 2021, pp. 12–16.
- [55] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, “Do Vision Transformers See Like Convolutional Neural Networks?” *Neural Information Processing Systems*, 2021.
- [56] B. Gheflati and H. Rivaz, “Vision Transformer for Classification of Breast Ultrasound Images,” *44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021.
- [57] H.-Y. Zhou, C. Lu, S. Yang, and Y. Yu, “ConvNets vs. Transformers: Whose Visual Representations are More Transferable?” in *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Oct. 2021, pp. 2230–2238.
- [58] Y. Wu, S. Qi, Y. Sun, S. Xia, Y. Yao, and W. Qian, “A vision transformer for emphysema classification using CT images,” *Physics in Medicine & Biology*, vol. 66, no. 24, p. 245016, Dec. 2021.
- [59] M. Filipiuk and V. Singh, “Comparing Vision Transformers and Convolutional Nets for Safety Critical Systems,” *The AAAI Workshop on Artificial Intelligence*

Safety, p. 5, 2022.

- [60] A. Galdran, G. Carneiro, and M. A. G. Ballester, "Convolutional Nets Versus Vision Transformers for Diabetic Foot Ulcer Classification," *Diabetic Foot Ulcers Grand Challenge*, 2022.
- [61] S. Cuenat and R. Couturier, "Convolutional Neural Network (CNN) vs Vision Transformer (ViT) for Digital Holography," in *2nd International Conference on Computer, Control and Robotics (ICCCR)*. Shanghai, China: IEEE, Mar. 2022, pp. 235–240.
- [62] D. A. Coccomini, R. Caldelli, F. Falchi, C. Gennaro, and G. Amato, "Cross-Forgery Analysis of Vision Transformers and CNNs for Deepfake Image Detection," in *1st International Workshop on Multimedia AI against Disinformation*. Association for Computing Machinery, 2022, pp. 52–58.
- [63] H. Wang, "Traffic Sign Recognition with Vision Transformers," in *6th International Conference on Information System and Data Mining*. Association for Computing Machinery, 2022, pp. 55–61.
- [64] C. Xin, Z. Liu, K. Zhao, L. Miao, Y. Ma, X. Zhu, Q. Zhou, S. Wang, L. Li, F. Yang, S. Xu, and H. Chen, "An improved transformer network for skin cancer classification," *Computers in Biology and Medicine*, vol. 149, p. 105939, Oct. 2022.
- [65] Y. Peng and Y. Wang, "CNN and transformer framework for insect pest classification," *Ecological Informatics*, vol. 72, p. 101846, Dec. 2022.
- [66] A. Bakhtiarnia, Q. Zhang, and A. Iosifidis, "Single-layer vision transformers for more accurate early exits with less overhead," *Neural Networks*, vol. 153, pp. 461–473, Sep. 2022.
- [67] E. Asadi Shamsabadi, C. Xu, A. S. Rao, T. Nguyen, T. Ngo, and D. Dias-da Costa, "Vision transformer-based autonomous crack detection on asphalt and concrete surfaces," *Automation in Construction*, vol. 140, p. 104316, Aug. 2022.
- [68] R. Reedha, E. Dericquebourg, R. Canals, and A. Hafiane, "Vision Transformers For Weeds and Crops Classification Of High Resolution UAV Images," *Remote Sensing*, vol. 14, no. 3, p. 592, Jan. 2022.
- [69] L. Bottou and O. Bousquet, "The Tradeoffs of Large Scale Learning," in *Advances in Neural Information Processing Systems*, vol. 20. Curran Associates, Inc., 2007.
- [70] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-Aware Minimization for Efficiently Improving Generalization," *International Conference on Learning Representations*, 2020.
- [71] G. M. Korpelevich, "The extragradient method for finding saddle points and other problems," in *Matecon*, vol. 12, 1976.

- [72] R. W. Stidham, W. Liu, S. Bishu, M. D. Rice, P. D. R. Higgins, J. Zhu, B. K. Nalamotohu, and A. K. Waljee, "Performance of a Deep Learning Model vs Human Reviewers in Grading Endoscopic Disease Severity of Patients With Ulcerative Colitis," *JAMA Network Open*, vol. 2, no. 5, p. e193963, May 2019.
- [73] E. Klang, Y. Barash, R. Y. Margalit, S. Soffer, O. Shimon, A. Albshesh, S. Ben-Horin, M. M. Amitai, R. Eliakim, and U. Kopylov, "Deep learning algorithms for automated detection of Crohn's disease ulcers by video capsule endoscopy," *Gastrointestinal Endoscopy*, vol. 91, no. 3, pp. 606–613.e2, Mar. 2020.
- [74] K. Takenaka, K. Ohtsuka, T. Fujii, M. Negi, K. Suzuki, H. Shimizu, S. Oshima, S. Akiyama, M. Motobayashi, M. Nagahori, E. Saito, K. Matsuoka, and M. Watanabe, "Development and Validation of a Deep Neural Network for Accurate Evaluation of Endoscopic Images From Patients With Ulcerative Colitis," *Gastroenterology*, vol. 158, no. 8, pp. 2150–2157, Jun. 2020.
- [75] T. Majtner, J. B. Brodersen, J. Herp, J. Kjeldsen, M. L. Halling, and M. D. Jensen, "A deep learning framework for autonomous detection and classification of Crohn's disease lesions in the small bowel and colon with capsule endoscopy," *Endoscopy International Open*, vol. 09, no. 09, pp. E1361–E1370, Sep. 2021.
- [76] T.-Y. Huang, S.-Q. Zhan, P.-J. Chen, C.-W. Yang, and H. H.-S. Lu, "Accurate diagnosis of endoscopic mucosal healing in ulcerative colitis using deep learning and machine learning," *Journal of the Chinese Medical Association*, vol. 84, no. 7, pp. 678–681, Jul. 2021.
- [77] E. Klang, A. Grinman, S. Soffer, R. Margalit Yehuda, O. Barzilay, M. M. Amitai, E. Konen, S. Ben-Horin, R. Eliakim, Y. Barash, and U. Kopylov, "Automated Detection of Crohn's Disease Intestinal Strictures on Capsule Endoscopy Images Using Deep Neural Networks," *Journal of Crohn's and Colitis*, vol. 15, no. 5, pp. 749–756, May 2021.
- [78] A. L. Udristoiu, D. Stefanescu, G. Gruionu, L. G. Gruionu, A. V. Iacob, J. G. Karstensen, P. Vilman, and A. Saftoiu, "Deep Learning Algorithm for the Confirmation of Mucosal Healing in Crohn's Disease, Based on Confocal Laser Endomicroscopy Images," *Journal of Gastrointestinal and Liver Diseases*, vol. 30, no. 1, pp. 59–65, Mar. 2021.
- [79] M. J. Mascarenhas Saraiva, J. Afonso, T. Ribeiro, J. Ferreira, H. Cardoso, A. P. Andrade, M. Parente, R. Natal, M. Mascarenhas Saraiva, and G. Macedo, "Deep learning and capsule endoscopy: automatic identification and differentiation of small bowel lesions with distinct haemorrhagic potential using a convolutional neural network," *BMJ Open Gastroenterology*, vol. 8, no. 1, p. e000753, Sep. 2021.
- [80] M. Chierici, N. Puica, M. Pozzi, A. Capistrano, M. D. Donzella, A. Colangelo, V. Osmani, and G. Jurman, "Automatically detecting Crohn's disease and Ul-

- cerative Colitis from endoscopic imaging,” *BMC Medical Informatics and Decision Making*, vol. 22, no. S6, p. 300, Nov. 2022.
- [81] R. T. Sutton, O. R. Zaiane, R. Goebel, and D. C. Baumgart, “Artificial intelligence enabled automated diagnosis and grading of ulcerative colitis endoscopy images,” *Scientific Reports*, vol. 12, no. 1, p. 2748, Dec. 2022.
 - [82] G. Ruan, J. Qi, Y. Cheng, R. Liu, B. Zhang, M. Zhi, J. Chen, F. Xiao, X. Shen, L. Fan, Q. Li, N. Li, Z. Qiu, Z. Xiao, F. Xu, L. Lv, M. Chen, S. Ying, L. Chen, Y. Tian, G. Li, Z. Zhang, M. He, L. Qiao, Z. Zhang, D. Chen, Q. Cao, Y. Nian, and Y. Wei, “Development and Validation of a Deep Neural Network for Accurate Identification of Endoscopic Images From Patients With Ulcerative Colitis and Crohn’s Disease,” *Frontiers in Medicine*, vol. 9, p. 854677, Mar. 2022.
 - [83] L. Wang, L. Chen, X. Wang, K. Liu, T. Li, Y. Yu, J. Han, S. Xing, J. Xu, D. Tian, U. Seidler, and F. Xiao, “Development of a Convolutional Neural Network-Based Colonoscopy Image Assessment Model for Differentiating Crohn’s Disease and Ulcerative Colitis,” *Frontiers in Medicine*, vol. 9, p. 789862, Apr. 2022.
 - [84] X. Luo, J. Zhang, Z. Li, and R. Yang, “Diagnosis of ulcerative colitis from endoscopic images based on deep learning,” *Biomedical Signal Processing and Control*, vol. 73, p. 103443, Mar. 2022.
 - [85] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, “Dataset of breast ultrasound images,” *Data in Brief*, vol. 28, p. 104863, Feb. 2020.
 - [86] M. H. Yap, G. Pons, J. Marti, S. Ganau, M. Sentis, R. Zwigelaar, A. K. Davison, and R. Marti, “Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks,” *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 4, pp. 1218–1226, Jul. 2018.
 - [87] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. De Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt, M. Riegler, and P. Halvorsen, “KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection,” in *8th ACM on Multimedia Systems Conference*. Taipei Taiwan: ACM, Jun. 2017, pp. 164–169.
 - [88] H. Borgli, M. Riegler, V. Thambawita, D. Jha, S. Hicks, and P. Halvorsen, “The HyperKvasir Dataset,” 2019, publisher: OSF.
 - [89] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen, D. Johansen, C. Griwodz, H. K. Stensland, E. Garcia-Ceja, P. T. Schmidt, H. L. Hammer, M. A. Riegler, P. Halvorsen, and T. Lange, “HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy,” *Scientific Data*, vol. 7, p. 283, 2020.

- [90] “CrohnIPI,” <https://crohnipi.ls2n.fr/en/crohn-ipi-project/> (accessed Feb. 21, 2023).
- [91] A. Maissin, R. Vallée, M. Flamant, M. Fondain-Bossiere, C. L. Berre, A. Coutrot, N. Normand, H. Mouchère, S. Coudol, C. Trang, and A. Bourreille, “Multi-expert annotation of Crohn’s disease images of the small bowel for automatic detection using a convolutional recurrent attention neural network,” *Endoscopy Int Open*, vol. 09, pp. E1136–E1144, 2021.
- [92] R. Vallée, A. Coutrot, N. Normand, and H. Mouchère, “Influence of expertise on human and machine visual attention in a medical image classification task,” in *European Conference on Visual Perception*, 2021.
- [93] R. Vallée, A. Maissin, A. Coutrot, H. Mouchère, A. Bourreille, and N. Normand, “CrohnIPI: An endoscopic image database for the evaluation of automatic Crohn’s disease lesions recognition algorithms,” in *Medical Imaging: Biomedical Applications in Molecular, Structural, and Functional Imaging*. SPIE, 2020, p. 61.
- [94] S. H. Kassani, P. Hosseinzadeh Kassani, M. Wesolowski, K. Schneider, and R. Deters, “Classification of histopathological biopsy images using ensemble of deep learning networks,” in *arXiv preprint*, 11 2019.
- [95] C. Gamage, I. Wijesinghe, C. Chitraranjan, and I. Perera, “GI-Net: Anomalies Classification in Gastrointestinal Tract through Endoscopic Imagery with Deep Learning,” in *Moratuwa Engineering Research Conference (MERCon)*. Moratuwa, Sri Lanka: IEEE, Jul. 2019, pp. 66–71.
- [96] I. Shahzadi, T. B. Tang, F. Meriadeau, and A. Quyyum, “CNN-LSTM: Cascaded Framework For Brain Tumour Classification,” in *IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, 2018, pp. 633–637.
- [97] R. Vankdothu, M. A. Hameed, and H. Fatima, “A Brain Tumor Identification and Classification Using Deep Learning based on CNN-LSTM Method,” *Computers and Electrical Engineering*, vol. 101, p. 107960, 2022.
- [98] M. N. Khan, M. A. Hasan, and S. Anwar, “Improving the Robustness of Object Detection Through a Multi-Camera-Based Fusion Algorithm Using Fuzzy Logic,” *Frontiers in Artificial Intelligence*, vol. 4, p. 638951, 2021.
- [99] T. Zoumpiskas, M. Salamó, and A. Puig, “Effective Early Stopping of Point Cloud Neural Networks,” in *Modeling Decisions for Artificial Intelligence*. Springer International Publishing, 2022, vol. 13408, pp. 156–167, series Title: Lecture Notes in Computer Science.
- [100] L. Prechelt, “Early Stopping — But When?” in *Neural Networks: Tricks of the Trade*. Springer Berlin Heidelberg, 2012, vol. 7700, pp. 53–67, series Title: Lecture Notes in Computer Science.

- [101] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, "How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers," *arXiv Version Number: 2*, 2021.
- [102] D. Das, H. Massa, A. Kulkarni, and T. Rekatsinas, "An Empirical Analysis of the Impact of Data Augmentation on Knowledge Distillation," *ArXiv*, 2020, publisher: arXiv Version Number: 2.
- [103] W. Li, S. Shao, W. Liu, Z. Qiu, Z. Zhu, and W. Huan, "What Role Does Data Augmentation Play in Knowledge Distillation?" in *Computer Vision – ACCV 2022*. Springer Nature Switzerland, 2023, vol. 13842, pp. 507–525, series Title: Lecture Notes in Computer Science.
- [104] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, p. 6, 2020.
- [105] M. Turan and F. Durmus, "UC-NfNet: Deep learning-enabled assessment of ulcerative colitis from colonoscopy images," *Medical Image Analysis*, vol. 82, p. 102587, 2022.
- [106] "Understanding AUC - ROC Curve," <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5> (accessed June 06, 2021).
- [107] "Accuracy vs. F1-Score," <https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2> (accessed June 24, 2021).
- [108] "Classification: True vs. False and Positive vs. Negative | Machine Learning," <https://developers.google.com/machine-learning/crash-course/classification/true-false-positive-negative> (accessed July 04, 2023).
- [109] "Explaining Precision vs. Recall to Everyone," <https://towardsdatascience.com/explaining-precision-vs-recall-to-everyone-295d4848edaf> (accessed June 06, 2021).
- [110] "Matthews Correlation Coefficient is The Best Classification Metric You've Never Heard Of | Boaz Shmueli | Towards Data Science," <https://towardsdatascience.com/the-best-classification-metric-youve-never-heard-of-the-matthews-correlation-coefficient-3bf50a2f3e9a> (accessed June 22, 2023).
- [111] R. Vallée, A. Coutrot, N. Normand, and H. Mouchère, "Accurate small bowel lesions detection in wireless capsule endoscopy images using deep recurrent attention neural network," in *IEEE 21st Int WS on Multimedia Signal Proc (MMSP)*, 2019.
- [112] X. Qian, C. Zhang, L. Chen, and K. Li, "Deep Learning-Based Identification of Maize Leaf Diseases Is Improved by an Attention Mechanism: Self-Attention,"

Frontiers in Plant Science, vol. 13, p. 864486, Apr. 2022.

- [113] H. Zhu, W. Ke, D. Li, J. Liu, L. Tian, and Y. Shan, “Dual cross-attention learning for fine-grained visual categorization and object re-identification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4692–4702.
- [114] J. S. Cardoso, R. Sousa, and I. Domingues, “Ordinal data classification using kernel discriminant analysis: A comparison of three approaches,” in *11th Int Conference on Machine Learning and Applications*, vol. 1, 2012, pp. 473–477.
- [115] I. Domingues and J. S. Cardoso, “Max-ordinal learning,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 7, pp. 1384–1389, 2014.
- [116] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, “MLP-Mixer: An all-MLP Architecture for Vision,” in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, pp. 24 261–24 272.

APPENDICES

Appendix A - Articles

"Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review"

Published in Applied Sciences journal

Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review

José Maurício , Inês Domingues  and Jorge Bernardino * 

Polytechnic of Coimbra, Coimbra Institute of Engineering (ISEC), Rua Pedro Nunes, 3030-199 Coimbra, Portugal; a2018056151@isec.pt (J.M.); ines.domingues@isec.pt (I.D.)

* Correspondence: jorge@isec.pt

Abstract: Transformers are models that implement a mechanism of self-attention, individually weighting the importance of each part of the input data. Their use in image classification tasks is still somewhat limited since researchers have so far chosen Convolutional Neural Networks for image classification and transformers were more targeted to Natural Language Processing (NLP) tasks. Therefore, this paper presents a literature review that shows the differences between Vision Transformers (ViT) and Convolutional Neural Networks. The state of the art that used the two architectures for image classification was reviewed and an attempt was made to understand what factors may influence the performance of the two deep learning architectures based on the datasets used, image size, number of target classes (for the classification problems), hardware, and evaluated architectures and top results. The objective of this work is to identify which of the architectures is the best for image classification and under what conditions. This paper also describes the importance of the Multi-Head Attention mechanism for improving the performance of ViT in image classification.

Keywords: transformers; Vision Transformers (ViT); convolutional neural networks; multi-head attention; image classification



Citation: Maurício, J.; Domingues, I.; Bernardino, J. Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Appl. Sci.* **2023**, *13*, 5521. <https://doi.org/10.3390/app13095521>

Academic Editor: Yu-Dong Zhang

Received: 20 March 2023

Revised: 19 April 2023

Accepted: 26 April 2023

Published: 28 April 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Nowadays, transformers have become the preferred models for performing Natural Language Processing (NLP) tasks. They offer scalability and computational efficiency, allowing models to be trained with more than a hundred billion parameters without saturating model performance. Inspired by the success of the transformers applied to NLP and assuming that the self-attention mechanism could also be beneficial for image classification tasks, it was proposed to use the same architecture, with few modifications, to perform image classification [1]. The author's proposal was an architecture, called Vision Transformers (ViT), which consists of breaking the image into 2D patches and providing this linear sequence of patches as input to the model. Figure 1 presents the architecture proposed by the authors.

In contrast to this deep learning architecture, there is another very popular tool for processing large volumes of data called Convolutional Neural Networks (CNN). The CNN is an architecture that consists of multiple layers and has demonstrated good performance in various computer vision tasks such as object detection or image segmentation, as well as NLP problems [2]. The typical CNN architecture starts with convolutional layers that pass through the kernels or filters, from left to right of the image, extracting computationally interpretable features. The first layer extracts low-level features (e.g., colours, gradient orientation, edges, etc.), and subsequent layers extract high-level features. Next, the pooling layers reduce the information extracted by the convolutional layers, preserving the most important features. Finally, the fully-connected layers are fed with the flattened output of the convolutional and pooling layers and perform the classification. Its architecture is shown in Figure 2.

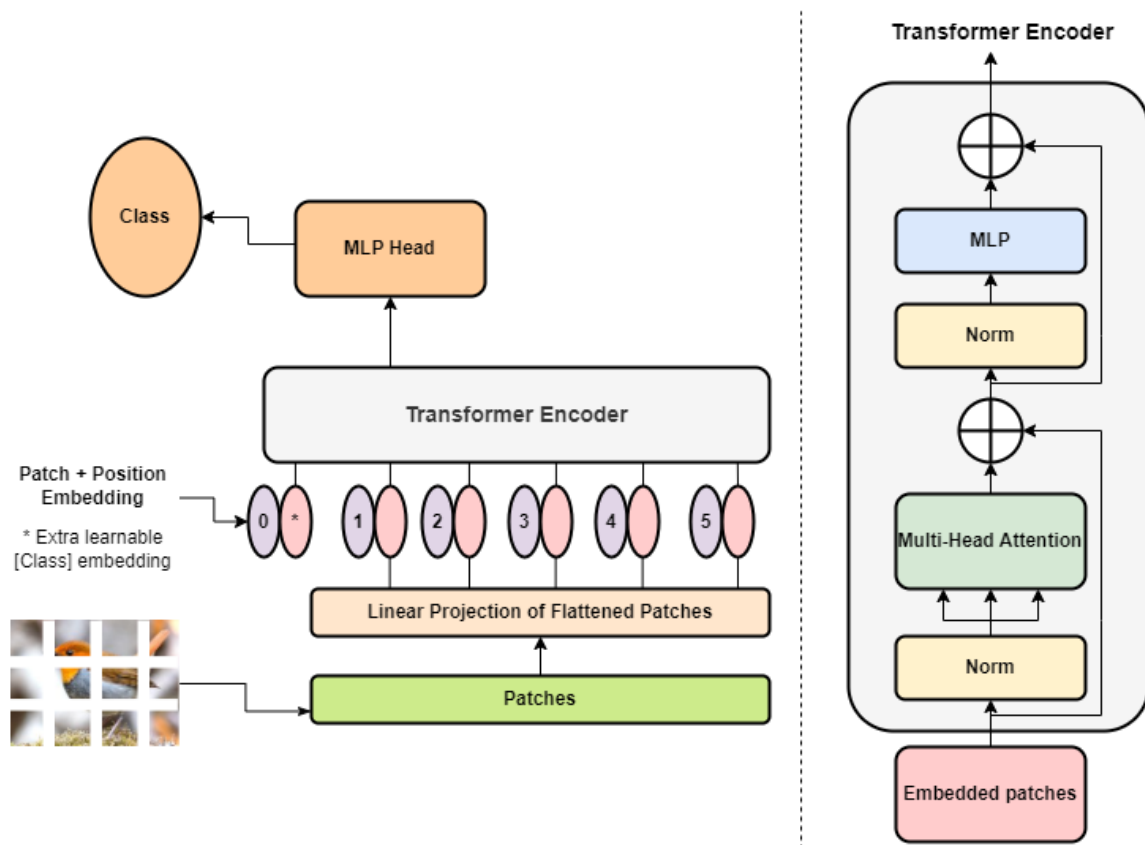


Figure 1. Example of an architecture of the ViT, based on [1].

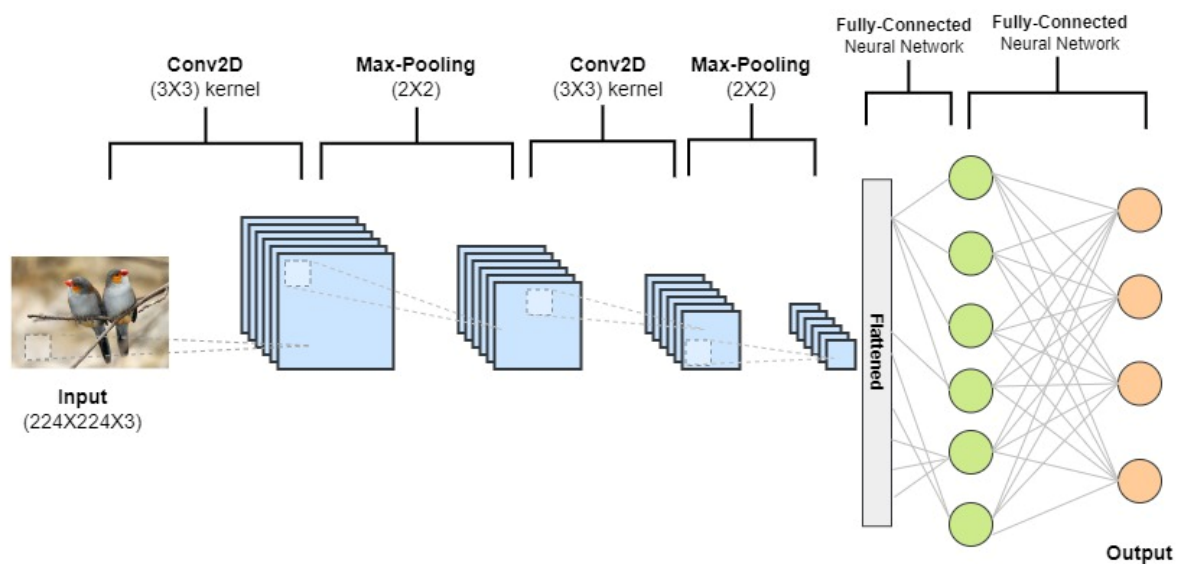


Figure 2. Example of an architecture of a CNN, based on [2].

With the increasing interest in Vision Transformers as a novel architecture for image recognition tasks, and the established success of CNNs in image classification, this work aims to review the state of the art in comparing Vision Transformers (ViT) and Convolutional Neural Networks (CNN) for image classification. Transformers offer advantages such as the ability to model long-range dependencies, adapt to different input sizes, and the potential for parallel processing, making them suitable for image tasks. However, Vision Transformers also face challenges such as computational complexity, model size, scalability

to large datasets, interpretability, robustness to adversarial attacks, and generalization performance. These points highlight the importance of comparing ViTs with older and established CNN models.

The overall goal of this work is to understand what conditions have the most influence on the performance of the two Deep Learning architectures, and what characteristics differ between the two architectures, that allow them to perform differently for the same objective. Some of the aspects that will be compared include datasets considerations, robustness, performance, evaluation, interpretability, and architecture. Specifically, we aim to answer the following research questions:

RQ1—Can the ViT architecture have a better performance than the CNN architecture, regardless of the characteristics of the dataset?

RQ2—What influences CNNs that do not perform as well as ViTs?

RQ3—How does the Multi-Head Attention mechanism, which is a key component of ViTs, influence the performance of these models in image classification?

In order to address these research questions, a literature review was conducted by searching various databases such as Google Scholar, Scopus, Web of Science, ACM Digital Library, and Science Direct using specific search terms. This paper presents the results of this review and analyses the methodologies and findings from the selected papers.

The rest of this paper is structured as follows. Section 2 describes the research methodology and search results. Section 3 presents the knowledge, methodology, and results found in the selected documents. Section 4 provides a brief overview of the reviewed papers and attempts to answer the three research questions. Section 5 discusses threats to the validity of the research. Section 6 overviews the strengths and weaknesses of each architecture and suggests future research directions, and Section 7 presents the main conclusions of this work.

2. Research Methodology

The purpose of a literature review is to evaluate, analyse and summarize the existing literature on a specific research topic, in order to facilitate the emergence of theoretical frameworks [3]. In this literature review, the aim is to synthesize the knowledge base, critically evaluate the methods used and analyze the results obtained in order to identify the shortcomings and improve the two aforementioned deep learning architectures for image classification. The methodology for conducting this literature review is based on the guidelines presented in [3,4].

2.1. Data Sources

ACM Digital Library, Google Scholar, Science Direct, Scopus, and Web of Science, were chosen as the data sources to extract the primary studies. The number of results found after searching papers in each of the data sources is shown in Table 1.

Table 1. Data sources and the number of obtained results.

Data Source	Number of Results	Number of Selected Papers
ACM Digital Library	19,159	1
Google Scholar	10,700	10
Science Direct	1437	3
Scopus	55	2
Web of Science	90	1

2.2. Search String

The research questions developed for this paper served as the basis for the search strings utilized in each of the data sources. Table 2 provides a list of the search strings used in each electronic database.

Table 2. Data sources and used search string.

Data Source	Search String
ACM Digital Library	((Vision Transformers) AND (convolutional neural networks) AND (images classification) AND (comparing))
Google Scholar	((ViT) AND (CNN) AND (Images Classification) OR (Comparing) OR (Vision Transformers) OR (convolutional neural networks) OR (differences))
Science Direct	((Vision Transformers) AND (convolutional neural networks) AND (images classification) AND (comparing))
Scopus	((ViT) AND (CNN) AND (comparing))
Web of Science	((ViT) AND (CNN) AND (comparing))

2.3. Inclusion Criteria

The inclusion criteria set to select the papers were that the studies were recent, had been written in English, and were published between January 2021 and December 2022. This choice of publication dates is based on the fact that ViTs were not proposed until the end of 2020 [1]. In addition, the studies had to demonstrate a comparison between CNNs and ViTs for image classification and could use any pre-trained model of the two architectures. Studies that presented a proposal for a hybrid architecture, where they combined the two architectures into one, were also considered. The dataset used during the studies did not have to be a specific one, but it had to be a dataset of images that allowed classification using both deep learning architectures.

2.4. Exclusion Criteria

Studies that oriented their research on using only one of the two deep learning architectures (i.e., Vision Transforms, or Convolutional Neural Networks) were excluded. Additionally, papers that were discovered to be redundant when searches were conducted throughout the chosen databases were eliminated. It was also defined that one of the exclusion criteria will be that the papers would have more than seven citations.

In summary, with the application of these criteria, 10,690 papers were excluded from the Google Scholar database, 89 papers from Web of Science, 53 papers from Scopus, 19,158 papers from ACM Digital Library, and 1434 papers from Science Direct.

2.5. Results

After applying the inclusion and exclusion criteria to the papers obtained in each of the electronic databases, seventeen (17) papers were selected for the literature review. Table 3 lists all the papers selected for this work, the year of publication and the type of publication.

Table 3. List of selected studies.

Ref.	Title	Year	Type
[5]	Adversarial Robustness Comparison of Vision Transformer and MLP-Mixer to CNNs	2021	Conference
[6]	Are Transformers More Robust Than CNNs?	2021	Conference
[7]	Detecting Pneumonia using Vision Transformer and comparing with other techniques	2021	Conference
[8]	Do Vision Transformers See Like Convolutional Neural Networks?	2021	Conference
[9]	Vision Transformer for Classification of Breast Ultrasound Images	2021	Conference
[10]	ConvNets vs. Transformers: Whose Visual Representations are More Transferable?	2021	Conference
[11]	A vision transformer for emphysema classification using CT images	2021	Journal
[12]	Comparing Vision Transformers and Convolutional Nets for Safety Critical Systems	2022	Conference
[13]	Convolutional Nets Versus Vision Transformers for Diabetic Foot Ulcer Classification	2022	Conference
[14]	Convolutional Neural Network (CNN) vs Vision Transformer (ViT) for Digital Holography	2022	Conference
[15]	Cross-Forgery Analysis of Vision Transformers and CNNs for Deepfake Image Detection	2022	Conference
[16]	Traffic Sign Recognition with Vision Transformers	2022	Conference
[17]	An improved transformer network for skin cancer classification	2022	Journal
[18]	CNN and transformer framework for insect pest classification	2022	Journal
[19]	Single-layer Vision Transformers for more accurate early exits with less overhead	2022	Journal
[20]	Vision transformer-based autonomous crack detection on asphalt and concrete surfaces	2022	Journal
[21]	Vision Transformers for Weeds and Crops Classification of High-Resolution UAV Images	2022	Journal

Figure 3 shows the distribution of the selected papers by year of publication.

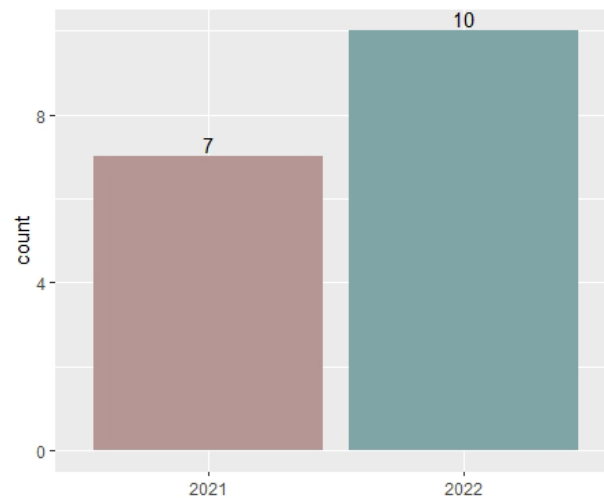


Figure 3. Distribution of the selected studies by years.

Figure 4 shows the distribution of the selected studies by application area. In the figure, most of the papers are generic in their application area. In these papers without a specific application area, the authors try to better understand the characteristics of the two architectures. For example, between CNNs and ViTs, the authors have tried to understand which of the architectures is more transferable. If architectures based on transformers are more robust than CNNs. And if the ViT will be able to see the same information as CNN with a different architecture. Within the health domain, some studies have been developed in different sub-areas, such as breast cancer, to show that ViT can be better than CNNs. The figure also shows that some work has been done, albeit to a lesser extent, in other application areas. Agriculture stands out with two papers comparing ViTs with CNNs.

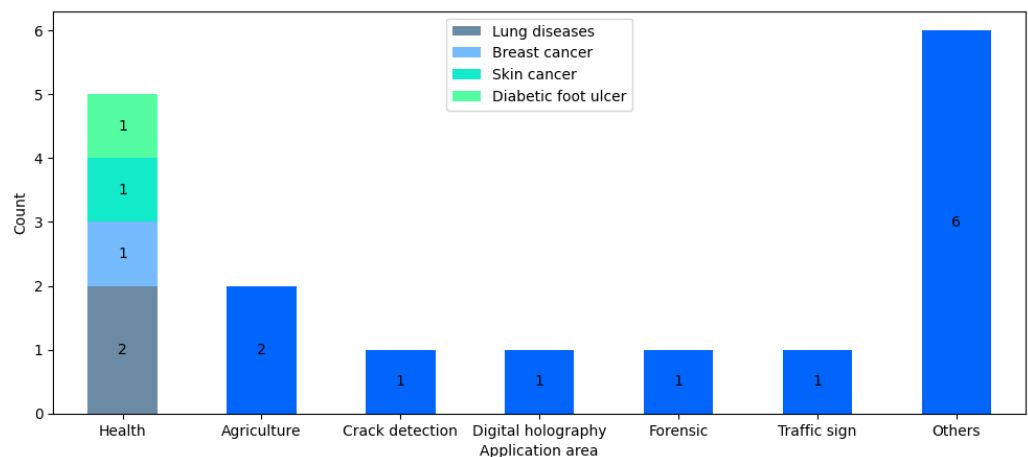


Figure 4. Distribution of the selected studies by application area.

3. Findings

An overview of the studies selected through the research methodology is shown in Table 4. This information summarizes the authors' approach, the findings, and other architectures that were used to build a comparative study. Therefore, to address the research questions, this section will offer an overview of the data found in the collected papers.

In the study developed in [12], the authors aimed to compare the two architectures (i.e., ViT and CNN), as well as the creation of a hybrid model that corresponded to the combination of the two. The experiment was conducted using the ImageNet dataset and

perturbations were applied to the dataset images. It was concluded that ViT can perform better and be more resilient on images with natural or adverse disturbances than CNN. It was also found in this work that the combination of the two architectures results in a 10% improvement in accuracy (Acc).

The work done in [14] aimed to compare Vision Transformers (ViT) with Convolutional Neural Networks (CNN) for digital holography, where the goal was to reconstruct amplitude and phase by extracting the distance of the object from the hologram. In this work, DenseNet201, DenseNet169, EfficientNetB4, EfficientNetB7, ViT-B/16, ViT-B32 and ViT-L/16 architectures were compared with a total of 3400 images. They were divided into four datasets, original images with or without filters, and negative images with or without filters. The authors concluded that ViT despite having an accuracy like CNN, was more robust because, due to the self-attention mechanism, it can learn the entire hologram rather than a specific area.

The authors in [7] studied the performance of ViT in comparison with other architectures to detect pneumonia, through chest X-ray images. Therefore, a ViT model, a CNN network developed by the authors and the VGG-16 network were used for the study which focussed on a dataset with 5856 images. After the experiments performed, the authors concluded that ViT was better than CNN with 96.45% accuracy, 86.38% validation accuracy, 10.8% loss and 18.25% validation loss. In this work, it was highlighted that ViT has a self-attention mechanism that allows splitting the image into small patches that are trainable, and each part of the image can be given an importance. However, the attention mechanism as opposed to the convolutional layers makes ViT's performance saturate fast when the goal is scalability.

In the study [21] the goal was to compare ViT with state-of-art CNN networks to classify UAV images to monitor crops and weeds. The authors compared the influence of the size of the training dataset on the performance of the architectures and found that ViT performed better with fewer images than CNN networks in terms of F1-Score. They concluded that ViT-B/16 was the best model to do crop and weed monitoring. In comparison with CNN networks, ViT could better learn the patterns of images in small datasets due to the self-attention mechanism.

In the scope of lung diseases, the authors in [11] investigated the performance of ViT models to automatically classify emphysema subtypes through Computed Tomography (CT) images in comparison with CNN networks. In this study, they performed a comparative study between the two architectures using a dataset collected by the authors (3192 patches) and a public dataset of 168 patches taken from 115 HRCT slides. In addition to this, they also verified the importance of pre-trained models. They concluded that ViT failed to generalize when trained with fewer images, because when comparing the pre-training accuracy with 91.27% on the training and 70.59% on the test.

In the work in [9], a comparison between state-of-the-art CNNs and ViT models for Breast ultrasound image classification was developed. The study was performed with two different datasets: the first containing 780 images and the second containing 163 images. The following architectures were selected for the study: ResNet50, VGG-16, Inception, NASNET, ViT-S/32, ViT-B/32, ViT-Ti/16, R + ViT-Ti/16 and R26 + ViT-S/16. ViT models were found to perform better than CNN networks for image classification. The authors also highlighted that ViT models could perform better when they were trained with a small dataset, because via the attention mechanism, it was possible to collect more information from different patches, instead of collecting information from the image.

Benz et al., in [5], compared ViT models, with the MLP-Mixer architecture and with CNNs. The goal was to evaluate which architecture was more robust in image classification. The study consisted of generating perturbations and adverse examples in the images and understanding which of the architectures was most robust. However, this study did not aim to analyse the causes. Therefore, the authors concluded that ViT were more robust than CNNs to adversarial attacks and from a features perspective CNN networks were more sensitive to high-frequency features. It was also described that the shift-variance property

of convolutional layers may be at the origin of the lack of robustness of the network in the classification of images that have been transformed.

The authors in [15] performed an analysis between ViT and CNN models aimed at detecting deepfake images. The experiment consisted in using the ForgeryNet dataset with 2.9 million images and 220 thousand video clips, together with three different image manipulation techniques, where they tried to train the models with real and manipulated images. By training the ViT-B model and the EfficientNetV2 network the authors demonstrated that the CNN network could generalize better and obtain higher training accuracy. However, ViT could have better generalization, reducing the bias in the identification of anomalies introduced by one or more different techniques to introduce anomalies.

Chao Xin et al. [17] aimed to compare their ViT model with CNN networks and with another ViT model to perform image classification to detect skin cancer. The experiment conducted by the authors used a public HAM10000 dataset with dermatoscopic skin cancer images and a clinical dataset collected through dermoscopy. In this study, a multi-scale image and the overlapping sliding window were used to serialize the images. They also used contrastive learning to improve the similarity of different labels and minimize the similarity in the same label. Thus, the ViT model developed was better for skin cancer classification using these techniques. However, the authors also demonstrated the effectiveness of balancing the dataset on the model performance, but they did not present the F1-Score values before the dataset is balanced to verify the improvement.

The authors in [19] aimed to study if ViT models could be an alternative to CNNs in time-critical applications. That is, for edge computing instances and IoT networks, applications using deep learning models consume multiple computational resources. The experiment used pre-trained networks such as ResNet152, DenseNet201, InceptionV3, and SL-ViT with three different datasets in the scope of images, audio, and video. They concluded that the ViT model introduced less overhead and performed better than the architectures used. It was also shown that increasing the kernel size of convolutional layers and using dilated convolutional caused a reduction in the accuracy of a CNN network.

In a study carried out in [20], the authors tried to find in ViTs an alternative solution to CNN networks for asphalt and concrete crack detection. The authors concluded that ViTs, due to the self-attention mechanism, had better performance in crack detection images with intense noise. CNN networks in the same images suffered from a high number of false negative rates, as well as the presence of biases in image classification.

Haolan Wang in [16] aimed to analyse eight different Vision Transformers and compare them with the performance of a pre-trained CNN network and without the pre-trained parameters to perform traffic signal recognition in autonomous driving systems. In this study, three different datasets with images of real-world traffic signals were used. This allowed the authors to conclude that the pre-trained DenseNet161 network had a higher accuracy than the ViT models to do traffic sign recognition. However, it was found in this work that ViT models performed better than the DenseNet161 network without the pre-trained parameters. From this work, it was also possible to conclude that the ViT models with a total number of parameters equal to or greater than the CNN networks, used during the experiment, had a shorter training time.

The work done in [13] compared CNN networks with Vision Transformers models for the classification of Diabetic Foot Ulcer images. For the study, the authors decided to use the following architectures: Big Image Transfer (BiT), EfficientNet, ViT-base and Data-efficient Image Transformers (DeiT) upon a dataset composed of 15,683 images. A further aim of this study to compare the performance of deep learning models using Stochastic Gradient Descent (SGD) [22] with Sharpness-Aware Optimization (SAM) [23,24]. These two tools are optimizers that seek to minimize the value of the loss function, improving the generalization ability of the model. However, SAM minimizes the value of the loss function and the sharpness loss, looking for parameters in the neighbourhood with a low loss. Therefore, this work concluded that the SAM optimizer originated an improvement in the values of F1-Score, AUC, Recall and Precision in all the architectures used. However, the authors did

not present the training and test values that allow for evaluating the improvement in the generalization of the models. Therefore, the BiT-ResNetX50 model with the SAM optimizer obtained the best performance for the classification of Diabetic Foot Ulcer images with F1-Score = 57.71%, AUC = 87.68%, Recall = 61.88%, and Precision = 57.74%.

The authors in [18] performed a comparative study between ViT models and CNN networks used in the state of the art with a model developed by them, where they combined CNN and transformers to perform insect pest recognition to protect agriculture worldwide. This study involved three public datasets: the IP102 dataset, the D0 dataset and Li's dataset. The algorithm created by the authors consisted of using the sequence of inputs formed by the CNN feature maps to make the model more efficient, and a flexible attention-based classification head was implemented to use the spatial information. Comparing the results obtained, the proposed model obtained a better performance in insect pest recognition with an accuracy of 74.897%. This work demonstrated that fine-tuning worked better on Vision Transformers than CNN, but on the other hand, this caused the number of parameters, the size, and the inference time of the model to increase significantly with respect to CNN networks. Through their experiments, the authors also demonstrated the advantage of using decoder layers in the proposed model to perform image classification. The greater the number of decoder layers, the greater the accuracy value of the model. However, this increase in the number of decoder layers increased the number of parameters, the size, and the inference time of the model. In other words, the architecture to process the images consumes far greater computational resources, which may not compensate for the increase in accuracy value with few layers. In the case of this study, the increase from one layer to three decoder layers represented only an increase of 0.478% in the accuracy value.

Several authors in [6,8,10] went deeper into the investigation and aimed to understand how the learning process of Vision Transformers works if ViT could be more transferable and better understand if the transform-based architecture were more robust than CNNs. In this sense, the authors in [8] intended to analyse the internal representations of ViT and CNN structures in image classification benchmarks and found differences between them. One of the differences was that ViT has greater similarity between high and low layers, while CNN architecture needs more low layers to compute similar representations in smaller datasets. This is due to the self-attention layers implemented in ViT, which allows it to aggregate information from other spatial locations, vastly different from the fixed field sizes in CNNs. They also observed that ViTs in the lower, self-attention layers can access information from local heads (small distances) and global heads (large distances). Whereas CNNs have access to information locally in the lower layers. On the other hand, the authors in [10] systematically analysed the transfer learning capacity in the two architectures. The study was conducted by comparing the performance of the two architectures on single-task and multi-task learning problems, using the ImageNet dataset. Through this study, the authors concluded that the transform-based architecture contained more transferable representations compared to convolutional networks for fine-tuning, presenting better performance and robustness in multi-task learning problems.

In another study carried out in [6], the goal was to prove if ViT were more robust than CNN as the most recent studies have shown. The authors developed their work comparing the robustness of the two architectures using two different types of perturbations: adversarial samples, which consists in evaluating the robustness of deep learning architectures in images with human-caused perturbations (i.e., data augmentation) and out-of-distribution samples, which consists in evaluating the robustness of the architectures in benchmarks of classification images. Through this experiment, it was demonstrated that by replacing the activation function ReLU by the activation function of transformer-based architecture (i.e., GELU) the CNN network was more robust than ViT in adversarial samples. In this study, it was also demonstrated that CNN networks were more robust than ViT in patch-based attacks. However, the authors concluded that the self-attention mechanism was the key to the robustness of the transformer-based architecture in most of the experiments performed.

Table 4. Overview of selected studies.

Ref.	Datasets	Images Size	Number of Classes	Hardware	Evaluated Architectures	Best Architecture	Best Results
[5]	ImageNet-1K (more than 1.431 M images) for training and ImageNet-C for validation	224 × 224	2	N/A	ViT-B/16, ViT-L/16, Mixer-B/16, Mixer-B/16, RN18 (SWSL), RN50 (SWSL), RN18 and RN50	ViT-L/16	82.89% of Acc
[6]	ImageNet-A; ImageNet-C and Stylized ImageNet	224 × 224	N/A	N/A	ResNet50 and DeiT-S	N/A	N/A
[7]	5856 images collected by X-ray	250 × 250 to ViT 224 × 224 to CNN	2	Intel Core i5-8300H 2.30 GHz	ViT, CNN and VGG16	ViT	96.45% Acc, 86.38% val. Acc, 10.92% loss and 18.25% val. Loss
[8]	ImageNetILSVRC 2012 (1.78 M images)	224 × 224	1000	N/A	ViT-B/32, ViT-L/16, ViT-H/14, ResNet50 and ResNet152	N/A	N/A
[9]	Public dataset 1 [25] with 780 images; Public dataset 2 [26] with 163 images	224 × 224	3	N/A	ViT-S/32, ViT-B/32, ViT-Ti/16, R26 + S/16, R + Ti/16, VGG, Inception and NASNET	ViT-B/32	86.7% Acc and 95% AUC
[10]	Flower 102 (4080 to 11,016 images); CUB 200 (11,788 images); Indoor 67 (15,620 images); NY Depth V2 (1449 images); WikiArt; COVID-19 Image Data Collection (700 images); Caltech101 (9146 images); FG-NET (1002 images)	384 × 384; 224 × 224; 300 × 300	40 to 102	N/A	R-101 × 3, R-152 × 4, ViT-B/16, ViT-L/16, and Swim-B	N/A	N/A
[11]	3192 images collected by CT and 160 images of a public dataset with CT biomarkers	61 × 61	4	Intel Core i7-9700 3.0 GHz, 326 GB RAM; NVIDIA GeForce RTX 2080 Ti (116 GB DDR6)	AlexNet, VGG-16, InceptionV3, MobileNetV2, ResNet34, ResNet50 and ViT	ViT	95.95% Acc
[12]	ImageNet-C benchmark	224 × 224	2	NVIDIA Quadro A6000	ViT-L/16, CNN, hybrid model (BiT-M + ResNet152 × 4)	Hybrid model	99.20% Acc
[13]	Dataset provided in DFUC 2021 challenge (15,683 images)	224 × 224	4	NVIDIA GeForce RTX 3080, 10 GB memory	EfficientNetB3, BiT-ResNeXt50, ViT-B/16 and DeiT-S/16	BiT-ResNeXt50	88.49% AUC, 61.53% F1-Score, 65.59% recall and 60.53% precision
[14]	3400 images collected by holographic camera	512 × 512	10	NVIDIA V100	EfficientNetB7, Densenet169 and ViT-B/16	ViT-B/16	99% Acc
[15]	ForgeryNet with 2.9 M images	N/A	2	N/A	EfficientNetV2 and ViT-B	N/A	N/A
[16]	German dataset (51,830 images); Indian dataset (1976 images); Chinese dataset (18,168 images)	128 × 128	15, 43 and 103	AMD Ryzen 7 5800H; NVIDIA GeForce RTX 3070	DenseNet161, ViT, DeepViT, MLP-Mixer, CvT, PiT, CaiT, CCT, CrossViT and Twins-SVT	CCT	99.04% Acc

Table 4. Cont.

Ref.	Datasets	Images Size	Number of Classes	Hardware	Evaluated Architectures	Best Architecture	Best Results
[17]	HAM10000 dataset (10,015 images); 1016 images collected by dermoscopy	224 × 224	3	Intel i7; 2x NVIDIA RTX 3060, 12 GB	MobileNetV2, ResNet50, InceptionV2, ViT and Proposed ViT model	Proposed ViT model	94.10% Acc, 94.10% precision and 94.10% F1-Score
[18]	IP102 dataset (75,222 images); D0 dataset (4508 images); Li's dataset (5629 images)	224 × 224; 480 × 480	10 to 102	Train: Intel Xeon; 8x NVIDIA Tesla V100, 256 GB. Test: Intel Core; NVIDIA GTX 1060 Ti, 16 GB	ResNet, EfficientNetB0, EfficientB1, RepVGG, VGG-16, ViT-L/16 and hybrid model proposed	Proposed hybrid model	99.47% Acc on the D0 dataset and 97.94% Acc on the Li's dataset
[19]	CIFAR-10 and CIFAR-100 (6000 images); Speech commands (100,503 1-second audio clips); GTZAM (1,00,030-second audio clips); DISCO (1935 images)	224 × 224 px; 1024 × 576 px. Spectrograms: 229 × 229 samples; 512 × 256 samples	10 to 100	4x NVIDIA 2080 Ti	SL-ViT, ResNet152, DenseNet201 and InceptionV3	SL-ViT	71.89% Acc
[20]	CrackTree260 (260 images); Ozegenel (458 images); Lab's on dataset (80,000 images)	256 × 256; 448 × 448	2	N/A	TransUNet, U-Net, DeepLabv3+ and CNN + ViT	CNN + ViT	99.55% Acc and 99.57% precision
[21]	10,265 images collected by Pilgrim technologies UAV with Sony ILCE-7R-36 mega pixels	64 × 64	5	Intel Xeon E5-1620 V4 3.50 GHz with 8 processor, 16 GB RAM; NVIDIA Quadro M2000	ViT-B/16, ViT-B/32, EfficientNetB, EfficientNetB1 and ResNet50	ViT-B/16	99.8% Acc

4. Discussion

The results can be summarized as follows. In [12], ViTs were found to perform better and be more resilient to images with natural or adverse disturbances compared to CNNs. Another study [14] concluded that ViTs are more robust in digital holography because they can access the entire hologram rather than just a specific area, giving them an advantage. ViTs have also been found to outperform CNNs in detecting pneumonia in chest X-ray images [7] and in classifying UAV images for crop and weed monitoring with small datasets [21]. However, it has been noted that ViT performance may saturate if scalability is the goal [7]. In a study on classifying emphysema subtypes in CT images [11], ViTs were found to struggle with generalization when trained on fewer images. Nevertheless, ViTs were found to outperform CNNs in breast ultrasound image classification, especially with small datasets [9]. Another study [5] found that ViTs are more robust to adversarial attacks and that CNNs are more sensitive to high-frequency features. The authors in [15] found that CNNs had higher training accuracy and better generalization, but ViTs showed potential to reduce bias in anomaly detection. In [17], the authors claimed that the ViT model showed better performance for skin cancer classification. ViTs have also been shown to introduce less overhead and perform better for time-critical applications in edge computing and IoT networks [19]. In [20], the authors investigated the use of ViTs for asphalt and concrete crack detection and found that ViTs performed better due to the self-attention mechanism, especially in images with intense noise and biases. Wang [16] found that a pre-trained CNN network had higher accuracy, but the ViT models performed better than the non-pre-trained CNN network and had a shorter training time. The authors in [13] used several models for diabetic foot ulcer image classification and compared SGD and SAM optimizers, concluding that the SAM optimizer improved several evaluation metrics. In [18], the authors showed that fine-tuning performed better on ViT models than CNNs for insect pest recognition.

Therefore, based on the information gathered from the selected papers, we attempt to answer the research questions posed in Section 1:

RQ1—Can the ViT architecture have a better performance than the CNN architecture, regardless of the characteristics of the dataset?

The literature review shows that ViT in image processing can be more efficient in smaller datasets due to the increase of relations created between images through the self-attention mechanism. However, it is also shown that if ViT trained with little data will have less generalization ability and worse performance compared to CNN's.

RQ2—What influences the CNNs that do not allow them to perform as well as the ViTs?

Shift-invariance is a limitation of CNN that makes the same architecture not have a satisfactory performance because the introduction of noise in the input images makes the same architecture unable to get the maximum information from the central pixels. However, the authors in [27] propose the addition of an anti-aliasing filter which combines blurring with subsampling in the Convolutional, MaxPooling and AveragePooling layers. Demonstrating through the experiment carried out that the application of this filter originates a greater generalization capacity and an increase in the accuracy of CNN. Furthermore, increasing the kernel size in convolutional layers and using dilated convolution have been shown as limitations that deteriorate the performance of CNNs against ViTs.

RQ3—How does the Multi-Head Attention mechanism, which is a key component of ViTs, influence the performance of these models in image classification?

The Attention mechanism is described as the mapping of a query and a set of key-value pairs to an output, the output being the result of a weighted sum of the values, in which the weight given is calculated, through the query with the corresponding key by a compatibility function. Multi-head Attention mechanism instead of performing a single attention function will perform multiple projections of attention functions [28]. This mechanism improves the ViT architecture because it allows it to extract more information from each pixel of the images that have been placed inside the embedding. In addition, this

mechanism can have better performance if the images have more secondary elements that illustrate the central element. And since this mechanism performs several computations in parallel, it reduces the computational cost [29].

Overall, ViTs have shown promising performance compared to CNNs in various applications, but there are limitations and factors that can affect their performance, such as dataset size, scalability, and pre-training accuracy.

5. Threats to Validity

This section discusses internal and external validity threats. The validity of the entire process performed in this study is demonstrated and how the results of this study can be replicated in other future experiments.

In this literature review, different search strings were used in each of the selected data sources, resulting in different results from each source. This approach may introduce a bias into the validation of the study, as it makes it difficult to draw conclusions about the diversity of studies obtained by replicating the same search. In addition, the maturity of the work was identified as an internal threat to validity, as the ViT architecture is relatively new and only a limited number of research projects have been conducted using it. In order to draw more comprehensive conclusions about the robustness of ViT compared to CNN, it is imperative that this architecture is further disseminated and deployed, thereby making more research available for analysis.

In addition to these threats, this study did not use methods that would allow to quantitatively and qualitatively analyse the results obtained in the selected papers. This may bias the validity of this review in demonstrating which of the deep learning architectures is more efficient in image processing.

The findings obtained in this study could be replicated in other future research in image classification. However, the results obtained may not be the same as those described by the selected papers because it has been proven that for different problems and different methodologies used, the results are different. In addition, the authors do not describe in sufficient detail all the methodologies they used, nor the conditions under which the experiment was performed.

6. Strengths, Limitations, and Future Research Directions

The review made it possible to identify not only the strengths of each architecture (outlined in Section 6.1), but also their potential for improvement (described in Section 6.2). Future research directions were also derived from this and are presented in Section 6.3.

6.1. Strengths

Both CNNs and ViTs have their own advantages, and some common ones. This section will explore these in more detail, including considerations on the Datasets, Robustness, Performance optimization, Evaluation, Explainability and Interpretability, and Architectures.

6.1.1. Dataset Considerations

CNNs have been widely used and extensively studied for image-related tasks, resulting in a rich literature, established architectures, and pre-trained models, making them accessible and convenient for many datasets. On the other hand, ViTs can process patches in parallel, which can lead to efficient computation, especially for large-scale datasets, and allow faster training and inference. ViTs can also handle images of different sizes and aspect ratios without losing resolution, making them more scalable and adaptable to different datasets and applications.

6.1.2. Robustness

CNNs are inherently translation-invariant, making them robust to small changes in object position or orientation within an image. The main advantage of ViTs is their ability

to effectively capture global contextual information through the self-attention mechanism, enabling them to model long-range dependencies and contextual relationships, which can improve robustness in tasks that require understanding global context. ViTs can also adaptively adjust the receptive fields of the self-attention mechanism based on input data, allowing them to better capture both local and global features, making them more robust to changes in scale, rotation, or perspective of objects.

Both architectures can be trained using data augmentation techniques, such as random cropping, flipping, and rotation, which can help improve robustness to changes in input data and reduce overfitting. Another technique is known as adversarial training, where they are trained on adversarial examples: perturbed images designed to confuse the model, to improve its ability to handle input data with adversarial perturbations. Combining models, using ensemble methods, such as bagging or boosting, can also improve robustness by exploiting the diversity of multiple models, which can help mitigate the effects of individual model weaknesses.

6.1.3. Performance Optimization

CNNs can be effectively compressed using techniques such as pruning, quantization, and knowledge distillation, reducing model size and improving inference efficiency without significant loss of performance. They are also well-suited for hardware optimization, with specialized hardware accelerators (e.g., GPUs, TPUs) designed to perform convolutional operations efficiently, leading to optimized performance in terms of speed and energy consumption.

ViTs can efficiently scale to handle high-resolution images or large-scale datasets because they operate on the entire image at once and do not require processing of local receptive fields at multiple spatial scales, potentially resulting in improved performance in terms of scalability.

Transfer Learning for pre-train on large-scale datasets and fine-tune on smaller datasets can potentially lead to improved performance with limited available data and can be used with both architectures.

6.1.4. Evaluation

CNNs have been widely used in image classification tasks for many years, resulting in well-established benchmarks and evaluation metrics that allow meaningful comparison and evaluation of model performance. The standardized evaluation protocols, such as cross-validation or hold-out validation, which provide a consistent framework for evaluating and comparing model performance across different datasets and tasks, are applicable for both architectures.

6.1.5. Explainability and Interpretability

CNNs produce feature maps that can be visualized, making it possible to interpret the behaviour of the model by visualizing the learned features or activations in different layers. They capture local features in images, such as edges or textures, which can lead to interpretable features that are visually meaningful and can provide insight into how the model is processing the input images. ViTs, on the other hand, are designed to capture global contextual information, making them potentially more interpretable in tasks that require an understanding of long-range dependencies or global context. They have a hierarchical structure with self-attention heads that can be visualized and interpreted individually, providing insights into how different heads attend to different features or regions in the input images.

6.1.6. Architecture

CNNs have a wide range of established architecture variants, such as VGG, ResNet, and Inception, with proven effectiveness in various image classification tasks. These architectures are well-tested and widely used in the deep learning community. ViTs can be

easily modified to accommodate different input sizes, patch sizes, and depth, providing flexibility in architecture design and optimization.

6.2. Limitations

Despite their many advantages and the breakthroughs made over the years. There are still some drawbacks to the architectures studied. This section focuses on these.

6.2.1. Dataset Considerations

CNNs can be susceptible to biases present in training datasets, such as biased sampling or label noise, which can affect the validity of training results. They typically operate on fixed input spatial resolutions, which may not be optimal for images of varying size or aspect ratio, resulting in information loss or distortion. While pre-trained models for CNNs are well-established, pre-trained models for ViT are (still) less common for some datasets, which may affect the ease of use in some situations.

6.2.2. Robustness

CNNs may struggle to capture long-range contextual information, as they focus primarily on local feature extraction, which may limit the ability to understand global context, leading to reduced robustness in tasks that require global context, such as scene understanding, image captioning or fine-grained recognition.

Both architectures can be prone to overfitting, especially when the training data is limited or noisy, which can lead to reduced robustness to input data outside the training distribution. Adversarial attacks can also pose a challenge to the robustness of both architectures. In particular, ViTs do not have an inherent spatial inductive bias like CNNs, which are specifically designed to exploit the spatial locality of images. This can make them more vulnerable to certain types of adversarial attacks that rely on spatial information, such as spatially transformed adversarial examples.

6.2.3. Performance Optimization

CNNs can suffer from reduced performance and increased memory consumption when applied to high-resolution images or large-scale datasets, as they require processing of local receptive fields at multiple spatial scales, leading to increased computational requirements. Compared to CNNs, ViTs are computationally expensive, especially as the image size increases or model depth increases, which may limit their use in certain resource-constrained environments. Reduced computational complexity can sometimes result in decreased robustness, as models may not have the ability to learn complex features that can help generalize well to adversarial examples.

6.2.4. Evaluation

As mentioned above, CNNs are primarily designed for local feature extraction and may struggle to capture long-range contextual dependencies, which can limit the evaluation performance in tasks that require understanding of global context or long-term dependencies. ViTs are relatively newer than CNNs, and as such, may lack well-established benchmarks or evaluation metrics for specific tasks or datasets, which can make performance evaluation difficult and less standardized.

6.2.5. Explainability and Interpretability

Despite well-established methods for model interpretation, CNNs still lack full interpretability because the complex interactions between layers and neurons can make it difficult to fully understand the model's decision-making process, particularly in deeper layers of the network.

While ViTs produce attention maps for interpretability, the complex interactions between self-attention heads can still present challenges in accurately interpreting the model's behaviour. ViTs can have multiple heads attending to different regions, which can make it

difficult to interpret the interactions between different attention heads and to understand the reasoning behind model predictions.

6.2.6. Architecture

CNNs typically have fixed, predefined model architectures, which may limit the flexibility to adapt to specific task requirements or incorporate domain-specific knowledge, potentially affecting performance optimization. For ViTs, the availability of established architecture variants is still limited, which may require more experimentation and exploration to find optimal architectures for specific tasks.

6.3. Future Research Directions

As future research, meta-analysis or systematic reviews should be conducted within the scope of this review to provide the scientific community with more detail on which of the architectures is more effective at image classification, in addition to specifying under what conditions a particular architecture stands out from the others. It is therefore necessary to facilitate the choice of the deep learning architecture to be used in future image classification problems. This section aims to provide guidelines for future research in this area.

6.3.1. Dataset Considerations

The datasets used in most studies may not be representative of real-world scenarios. Future research should consider using more diverse datasets that better reflect the complexity and variability of real-world images. As an example, it would be interesting to study the impact that image resolution might have on the performance of deep learning architectures. That is, it would be important to find out in which of the architectures (i.e., ViT, CNN, and MLP-Mixer) the resolution of the images will influence their performance, as well as what impact it will have on the processing time of the deep learning architectures.

6.3.2. Robustness

As documented in [5], deep learning models are typically vulnerable to adversarial attacks, where small perturbations to an input image can cause the model to misclassify it. Future research should focus on developing architectures that are more robust to adversarial attacks (for example by further augmenting the robustness of ViTs), as well as exploring ways to detect and defend against these attacks.

Beyond that, most studies (as the ones reviewed in this work) have focused on the performance of deep learning architectures on image classification tasks, but there are many other image processing tasks (such as object detection, segmentation, and captioning) that could benefit from the use of these architectures. Future research should further explore the effectiveness of these architectures on these tasks.

6.3.3. Performance Optimization

Deep learning architectures require substantial amounts of labelled data to achieve high performance. However, labelling data is time-consuming and expensive. Future research should explore ways to improve the efficiency of deep learning models, such as developing semi-supervised learning methods or transfer learning (following up on the finding in [10]) that can leverage pre-trained models.

In addition, the necessity of large amounts of labelled data requires significant computational resources, which limits the deployment on resource-constrained devices. Future research should focus on developing architectures that are optimized for deployment on these devices, as well as exploring ways to reduce the computational cost of existing architectures. It should explore the advantages of the implementation of the knowledge distillation of deep learning architectures to reduce computational resources.

6.3.4. Evaluation

The adequacy of the metrics to the task and problem at hand is also another suggested line of future research. Most studies have used standard performance metrics (such as accuracy and F1-score) to evaluate the performance of deep learning architectures. Future research should consider using more diverse metrics that better capture the strengths and weaknesses of different architectures.

6.3.5. Explainability and Interpretability

Deep learning models are often considered as black boxes because they do not provide insight into the decision-making process. This may prevent the usage of the models in certain areas such as justice and healthcare [30], among others. Future research should focus on making these models more interpretable and explainable. For example, by designing transformer architectures that provide visual explanations of their decisions or by developing methods for extracting features that are easily interpretable.

6.3.6. Architecture

In future investigations, it will be necessary to study the impact of the MLP-Mixer deep learning architecture in image processing, what are the characteristics that allow it to have a performance superior to CNNs, but inferior to the performance obtained by the ViT architecture [5]. Future research should also focus on developing novel architectures that can achieve high performance with fewer parameters or that are more efficient in terms of computation and memory usage.

7. Conclusions

This work has reviewed recent studies done in image processing to give more information about the performance of the two architectures and what distinguishes them. A common feature across all papers is that transformer-based architecture or the combination of ViTs with CNN allows for better accuracy compared to CNN networks. It has also been shown that this new architecture, even with hyperparameters fine-tuning, can be lighter than the CNN, consuming fewer computational resources and taking less training time as demonstrated in the works [16,19].

In summary, the ViT architecture is more robust than CNN networks for images that have noise or are augmented. It manages to perform better compared to CNN due to the self-attention mechanism because it makes the overall image information accessible from the highest to the lowest layers [12]. On the other hand, CNN's can generalize better with smaller datasets and get better accuracy than ViTs, but in contrast, ViTs have the advantage of learning information better with fewer images. This is because the images are divided into small patches, so there is a greater diversity of relationships between them.

Author Contributions: Conceptualization, J.B.; Methodology, J.M. and J.B.; Software, J.M.; Validation, J.M., I.D. and J.B.; Formal analysis, J.M., I.D. and J.B.; Investigation, J.M.; Resources, J.M.; Data curation, J.M.; Writing—original draft preparation, J.M.; Writing—review and editing, J.M., I.D. and J.B.; Supervision, J.B. and I.D.; Project administration, J.B. and I.D.; Funding acquisition, J.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929. [CrossRef]
2. Saha, S. A Comprehensive Guide to Convolutional Neural Networks—The ELI5 Way. Available online: <https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53> (accessed on 8 January 2023).

3. Snyder, H. Literature Review as a Research Methodology: An Overview and Guidelines. *J. Bus. Res.* **2019**, *104*, 333–339. [\[CrossRef\]](#)
4. Matloob, F.; Ghazal, T.M.; Taleb, N.; Aftab, S.; Ahmad, M.; Khan, M.A.; Abbas, S.; Soomro, T.R. Software Defect Prediction Using Ensemble Learning: A Systematic Literature Review. *IEEE Access* **2021**, *9*, 98754–98771. [\[CrossRef\]](#)
5. Benz, P.; Ham, S.; Zhang, C.; Karjauv, A.; Kweon, I.S. Adversarial Robustness Comparison of Vision Transformer and MLP-Mixer to CNNs. *arXiv* **2021**, arXiv:2110.02797. [\[CrossRef\]](#)
6. Bai, Y.; Mei, J.; Yuille, A.; Xie, C. Are Transformers More Robust Than CNNs? *arXiv* **2021**, arXiv:2111.05464. [\[CrossRef\]](#)
7. Tyagi, K.; Pathak, G.; Nijhawan, R.; Mittal, A. Detecting Pneumonia Using Vision Transformer and Comparing with Other Techniques. In Proceedings of the 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), IEEE, Coimbatore, India, 2 December 2021; pp. 12–16.
8. Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; Dosovitskiy, A. Do Vision Transformers See Like Convolutional Neural Networks? *arXiv* **2021**, arXiv:2108.08810. [\[CrossRef\]](#)
9. Gheflati, B.; Rivaz, H. Vision Transformer for Classification of Breast Ultrasound Images. *arXiv* **2021**, arXiv:2110.14731. [\[CrossRef\]](#)
10. Zhou, H.-Y.; Lu, C.; Yang, S.; Yu, Y. ConvNets vs. Transformers: Whose Visual Representations Are More Transferable? In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), IEEE, Montreal, BC, Canada, 17 October 2021; pp. 2230–2238.
11. Wu, Y.; Qi, S.; Sun, Y.; Xia, S.; Yao, Y.; Qian, W. A Vision Transformer for Emphysema Classification Using CT Images. *Phys. Med. Biol.* **2021**, *66*, 245016. [\[CrossRef\]](#)
12. Filipiuk, M.; Singh, V. Comparing Vision Transformers and Convolutional Nets for Safety Critical Systems. *AAAI Workshop Artif. Intell. Saf.* **2022**, *3087*, 1–5.
13. Galdran, A.; Carneiro, G.; Ballester, M.A.G. Convolutional Nets Versus Vision Transformers for Diabetic Foot Ulcer Classification. *arXiv* **2022**, arXiv:2111.06894. [\[CrossRef\]](#)
14. Cuenat, S.; Couturier, R. Convolutional Neural Network (CNN) vs Vision Transformer (ViT) for Digital Holography. In Proceedings of the 2022 2nd International Conference on Computer, Control and Robotics (ICCCR), IEEE, Shanghai, China, 18 March 2022; pp. 235–240.
15. Coccomini, D.A.; Caldelli, R.; Falchi, F.; Gennaro, C.; Amato, G. Cross-Forgery Analysis of Vision Transformers and CNNs for Deepfake Image Detection. In Proceedings of the 1st International Workshop on Multimedia AI against Disinformation, Newark, NJ, USA, 27–30 June 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 52–58.
16. Wang, H. Traffic Sign Recognition with Vision Transformers. In Proceedings of the 6th International Conference on Information System and Data Mining, Silicon Valley, CA, USA, 27–29 May 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 55–61.
17. Xin, C.; Liu, Z.; Zhao, K.; Miao, L.; Ma, Y.; Zhu, X.; Zhou, Q.; Wang, S.; Li, L.; Yang, F.; et al. An Improved Transformer Network for Skin Cancer Classification. *Comput. Biol. Med.* **2022**, *149*, 105939. [\[CrossRef\]](#)
18. Peng, Y.; Wang, Y. CNN and Transformer Framework for Insect Pest Classification. *Ecol. Inform.* **2022**, *72*, 101846. [\[CrossRef\]](#)
19. Bakhtiarnia, A.; Zhang, Q.; Iosifidis, A. Single-Layer Vision Transformers for More Accurate Early Exits with Less Overhead. *Neural Netw.* **2022**, *153*, 461–473. [\[CrossRef\]](#)
20. Asadi Shamsabadi, E.; Xu, C.; Rao, A.S.; Nguyen, T.; Ngo, T.; Dias-da-Costa, D. Vision Transformer-Based Autonomous Crack Detection on Asphalt and Concrete Surfaces. *Autom. Constr.* **2022**, *140*, 104316. [\[CrossRef\]](#)
21. Reedha, R.; Dericquebourg, E.; Canals, R.; Hafiane, A. Vision Transformers for Weeds and Crops Classification of High Resolution UAV Images. *Remote Sens.* **2022**, *14*, 592. [\[CrossRef\]](#)
22. Bottou, L.; Bousquet, O. The Tradeoffs of Large Scale Learning. In *Advances in Neural Information Processing Systems*; Platt, J., Koller, D., Singer, Y., Roweis, S., Eds.; Curran Associates, Inc.: Vancouver, BC, Canada, 2007; Volume 20.
23. Foret, P.; Kleiner, A.; Mobahi, H.; Neyshabur, B. Sharpness-Aware Minimization for Efficiently Improving Generalization. *arXiv* **2020**, arXiv:2010.01412. [\[CrossRef\]](#)
24. Korpelevich, G.M. The Extragradient Method for Finding Saddle Points and Other Problems. *Ekon. Mat. Metod.* **1976**, *12*, 747–756.
25. Al-Dhabyani, W.; Gomaa, M.; Khaled, H.; Fahmy, A. Dataset of Breast Ultrasound Images. *Data Brief* **2020**, *28*, 104863. [\[CrossRef\]](#)
26. Yap, M.H.; Pons, G.; Marti, J.; Ganau, S.; Sentis, M.; Zwiggelaar, R.; Davison, A.K.; Marti, R. Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks. *IEEE J. Biomed. Health Inform.* **2018**, *22*, 1218–1226. [\[CrossRef\]](#)
27. Zhang, R. Making Convolutional Networks Shift-Invariant Again. *arXiv* **2019**, arXiv:1904.11486. [\[CrossRef\]](#)
28. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *Neural Inf. Process. Syst.* **2017**, *30*, 3762. [\[CrossRef\]](#)
29. Zhou, D.; Kang, B.; Jin, X.; Yang, L.; Lian, X.; Jiang, Z.; Hou, Q.; Feng, J. DeepViT: Towards Deeper Vision Transformer. *arXiv* **2021**, arXiv:2103.11886. [\[CrossRef\]](#)
30. Amorim, J.P.; Domingues, I.; Abreu, P.H.; Santos, J.A.M. Interpreting Deep Learning Models for Ordinal Problems. In Proceedings of the European Symposium on Artificial Neural Networks, Bruges, Belgium, 25–27 April 2018.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

**"Deep Neural Networks to distinguish between Crohn's disease and
Ulcerative colitis"**

Published in IbPRIA 2023: 11th Iberian Conference on Pattern Recognition and Image
Analysis

Deep Neural Networks to distinguish between Crohn's disease and Ulcerative colitis

José Maurício¹[0000–0001–8234–9481] and Inês Domingues^{1,2}[0000–0002–2334–7280]

¹ Instituto Politécnico de Coimbra, Instituto Superior de Engenharia, Rua Pedro Nunes - Quinta da Nora, 3030-199 Coimbra, Portugal

² Centro de Investigação do Instituto Português de Oncologia do Porto (CI-IPOP): Grupo de Física Médica, Radiobiologia e Protecção Radiológica
{a2018056151, ines.domingues}@isec.pt

Abstract. The number of patients with inflammatory bowel disease (IBD) has been increasing. The diagnosis is a difficult task for the gastroenterologist performing the endoscopic examination. However, in order to prescribe medical treatment and provide quality of life to the patient, the diagnosis must be quick and accurate. Therefore, it is important to develop tools that, based on the characteristics of the inflammatory present in the mucosa, automatically recognise the type of inflammatory bowel disease. This paper presents a study where the objective was to collect and analyse endoscopic images referring to Crohn's disease and Ulcerative colitis using six deep learning architectures: AlexNet, ResNet50, InceptionV3, VGG-16, ResNet50+MobileNetV2, and a hybrid model. The hybrid model consists of the combination of two architectures, a CNN and a LSTM. This work also presents techniques that can be used to pre-process the images before the training to remove accessory elements from the image. The obtained results demonstrate that it is possible to automate the process of diagnosing patients with IBD using convolutional networks for processing images collected during an endoscopic examination, and thus develop tools to help the medical specialist diagnose the disease.

Keywords: Inflammatory Bowel Disease · CNN · image classification · Crohn's disease · Ulcerative colitis.

1 Introduction

Inflammatory bowel disease (IBD) is characterized as a disease of unknown cause that results from the interaction between genetic and environmental factors, triggering an immune response that causes digestive disorders and inflammation in the gastrointestinal tract. These types of diseases are divided into Ulcerative colitis and Crohn's disease [17]. However, some studies have shown that some of the diagnoses of inflammatory bowel disease refer to a different category of disease, called indeterminate colitis. This pathology refers to colitis with an extreme degree of severity that has features of both diseases (Ulcerative colitis and Crohn's disease), making it difficult to distinguish between them [16].

It is estimated that in Portugal it already affects 7,000 to 15,000 people, with 2.9 cases per 100,000 inhabitants per year with Ulcerative colitis and 2.4 cases per 100,000 inhabitants per year with Crohn's disease [7]. In Europe, the prevalence of the disease is more than 1 in 198 people for Ulcerative colitis and 1 in 310 people for Crohn's disease [26]. And if the comparison is made worldwide, the prevalence of Ulcerative colitis is 7.6 in 246 cases per 100,000 people per year, preceded by Crohn's disease with 3.6 in 214 cases per 100,000 people per year. This disease has a higher incidence in North America and Europe. However, in recent decades, it has seen an increase in countries such as China and India [11].

As the number of patients worldwide with IBD has been increasing, its diagnosis has become an arduous task for medical specialists and may have some limitations in recognising the disease. To automate the diagnosis of this chronic disease and thus reduce the patient's waiting time in prescribing medical treatment. This work seeks, through the selection of six deep learning architectures, to classify images related to the type of disease (i.e., Crohn's disease and Ulcerative colitis) that were previously collected using tests such as endoscopy and colonoscopy and from the characteristics present in the mucosa according to the type of inflammation, to predict the type of inflammatory bowel disease.

The proposed methodology allows the pre-processing of images to treat the observations made by gastroenterologists during the endoscopic examination, allowing the CNN architectures to focus on the part of the image where the intestinal mucosa is located. In addition, this study presents the development of two hybrid architectures: (i) the junction of a CNN with an LSTM; (ii) the combination of convolutional layers from the ResNet50 network with fully connected layers from the MobileNetV2 network. Which, as demonstrated by the obtained results, are accurate in predicting the type of inflammatory bowel disease.

In summary, the contributions of this work include:

- The derivation of two versions of a new database for the distinction of Crohn's disease from Ulcerative colitis
- The proposal of three pre-processing techniques to minimize the impact of artifacts in the training data
- The development of two new Hybrid Deep Learning Architectures
- Extensive testing and validation (two versions of the database, combined with three pre-processing methods, combined with six architectures, totalling 36 experimental settings)

This paper is divided into four sections: section 2 compiles some works of interest; section 3 presents the methodology developed during this work and the datasets that were used for the classification; section 4 collects the quantitative results; and section 5 summarises the conclusions and directions for future work.

2 Related work

In this section, a summary of the findings made by some related work within the scope of this study are presented. Along the years, the use of computer

vision tools in the diagnosis of inflammatory bowel disease has become essential. Since the evaluation of images obtained through imaging examinations is done by professional gastroenterologists, the criteria used to determine the patient’s diagnosis is subjective and may differ among doctors [19]. This process becomes a consequence for the patient as it results in a delay in prescribing an effective treatment to combat the disease.

The authors in [13] developed a study using the CrohnIPI dataset. It consisted of 2124 images without pathology, 1360 images with pathology, and 14 inconclusive images. The goal was to develop a recurrent attention neural network for diagnosing mucosal lesions caused by inflammation. In addition, they also tested networks such as ResNet34, VGGNet16, and VGGNet19 under the same conditions. The experiment was conducted in three rounds: in the first round, the images were labelled by a medical expert, and the networks were trained based on those images; in the second round, the images identified in the first round were validated by three experts; and in the third round, the four medical experts met in several sessions to reduce the number of inconclusive images. They concluded that the proposed network achieved its best performance in the last round with 93.70% accuracy.

Other authors developed a study in [22], where they propose a recurrent attention neural network to classify images with lesions. In this study, the authors used the CrohnIPI dataset to train the proposed network and then used images of several lesions from the CAD-CAP dataset presented during the Giana competition to test [9]. Based on the results presented, the proposed network obtained 90.85% accuracy with the CrohnIPI dataset, and when it was tested with the images from the CAD-CAP dataset, it obtained an accuracy of 99.67%.

In the study developed in [14], the authors sought to develop a system that, through deep learning, could reduce subjectivity and improve confidence in the diagnosis of Ulcerative colitis (UC) disease. Therefore, the authors used the LIMUC dataset to train a proposed deep learning regression-based approach to assess the severity of UC according to the Mayo endoscopic score (MEC). They conclude that this proposed approach after training on $k = 10$ folds and compared to state-of-the-art convolutional neural networks, was more robust in classifying images from the test dataset. This is based on the results obtained on macro F1-score metrics and quadratic weighted Kappa.

Based on the findings made by other authors, this work seeks to stand out by proposing to pool the images by the type of inflammatory bowel disease to distinguish between Crohn’s disease and Ulcerative colitis. A methodology that includes some image pre-processing techniques for removing redundant features from endoscopic exam images is also suggested. Moreover, a set of six classification metrics allow the performance of the deep learning architectures proposed in this work to be critically evaluated.

3 Methodology

In order to improve the diagnosis of inflammatory bowel disease, a methodology based on six phases is proposed. In the first phase, the authors collected images of

the two inflammatory bowel diseases (i.e. Ulcerative colitis and Crohn’s disease). In the second phase, frames were extracted from the videos collected during an endoscopic examination performed on the patients. In the third phase, was applied a pre-processing to the images. In the fourth phase, data augmentation was applied to the training dataset. In the fifth phase, the convolutional networks were implemented and configured. Finally, the performance of the CNNs was evaluated using the classification metrics: Accuracy, Loss, Precision, Recall, F1-Score, Area Under Curve and Mathew’s Correlation Coefficient. This methodology is illustrated in Figure 1 and will be further explained in the following sections.

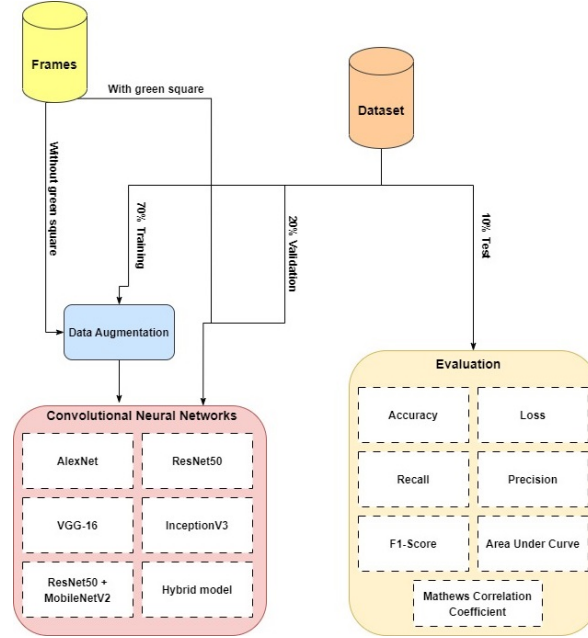


Fig. 1: Experimental setup

3.1 Dataset

Since we were not able to find a dataset with examples from both Crohn’s disease and Ulcerative colitis, for the development of this study, three datasets of images referring to Crohn’s disease and Ulcerative colitis were used, with the goal of creating a single dataset involving images of both diseases. As illustrated in Figure 2, HyperKvasir [1, 2] and LIMUC [15] datasets were used to get images of the Ulcerative colitis disease, while the CrohnIPI [6] dataset was selected to

collect images of Crohn’s disease. This last dataset has already been used in other works to develop tools for diagnosing the disease [23, 24].



Fig. 2: Ulcerative colitis image from the HyperKvasir dataset (left); Ulcerative colitis image from the LIMUC dataset (center); Crohn’s disease image from the CrohnIPI dataset (right).

We started by gathering the images present in the CrohnIPI dataset that were divided by lesion type and assigning them to Crohn-disease class. Then we also collected the ulcerative colitis disease images from the HyperKvasir dataset, which were divided by severity according to the Mayo endoscopic score (MAE) and assign them to ulcerative-colitis class. Figure 3 (left) demonstrates the initial distribution of the classes in the created dataset. This dataset does not contain the class “Normal” included in the CrohnIPI dataset, which was excluded. Fourteen images were also ignored due to the lack of labelling.

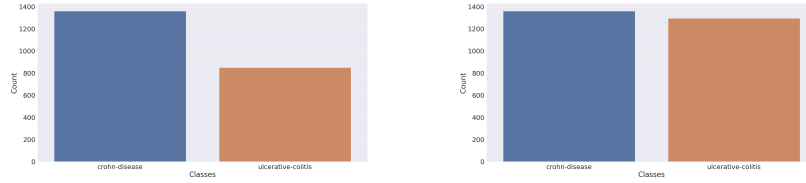


Fig. 3: Class distribution of the imbalanced dataset (left) and the balanced dataset (right). The blue bar represents Crohn’s disease, while the orange bar refers to Ulcerative colitis.

After a first inspection of the distribution of the dataset, it was concluded that the classes are not balanced. The ulcerative-colitis class is in the minority when compared to the Crohn’s disease class. To mitigate this problem, an additional dataset was created that maintained the basis of the first dataset with the addition of 444 more images from the LIMUC dataset. Since the LIMUC dataset was divided by the degree of severity according to the Mayo score, the images were collected, taking into account 25% of each severity grade. 111 images from Mayo 0, 111 images from Mayo 1, 111 images from Mayo 2, and 111 images from Mayo 3.

The HyperKvasir dataset, besides the above-mentioned images, contains videos identified by gastroenterologists. With these videos, the goal is to extract more images related to ulcerative colitis disease to balance the classes of the created

dataset. We proceeded to extract 20 frames per second with the same dimension as the dataset images (572×531). This procedure resulted in the extraction of a total of 64 frames. Figure 3 (right) demonstrates the distribution of the classes of this more balanced dataset.

3.2 Pre-Processing

In the images and frames extracted from a video of the HyperKvasir dataset, there was a green square in the lower left corner, which represents the position of the endoscope during the diagnostic examination. This may become a problem for classifying the type of inflammatory bowel disease present in the images provided to CNN. Therefore, experiments were performed where some transformations were applied when this type of situation was verified.

One of the applied transformations was a horizontal crop across the entire width of the image to remove the green square in the lower left corner. Then a few more experiments were performed, where a vertical crop was applied over the entire height of the image to remove that part of the image. However, during these transformations, some information regarding the mucosa of the intestine is lost. Some more experiments were performed in which a Gaussian blur of $2/16$ was applied only to the area where the green square was. The goal was to preserve as much of the image information as possible but to disguise that feature in some images. An illustration of these pre-processing methods is available in Figure 4.

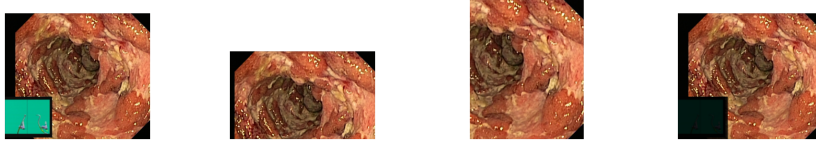


Fig. 4: Pre-processing examples: Original image (left); Horizontal crop (center-left); Vertical crop (center-right); Gaussian blur (right).

3.3 Experimental setup

In conducting this study, Tensorflow, version 2.11.0, was used. The programming environment for importing the libraries was Google Colab with the NVIDIA T4 Tensor Core GPU.

Data augmentation was performed on the train set only with horizontal and vertical Random Flip, Random Contrast with a factor of 0.15, Random Rotation with a factor of 0.2 and a Random Zoom with a portion of -0.2 for height and -0.3 for width.

The presence of the green square in the lower-left corner of the images obtained from the HyperKvasir dataset, rose the concern that this could lead the

models to identify the green square as a disease feature. Therefore, six experiments were performed to test these hypotheses, by combining the three pre-processing techniques presented in Section 3.2 with the two dataset versions described in 3.1. We refer to Table 1 for the configurations of each experiment.

Table 1: Configurations of the six performed experiments

Pre-processing	Dataset	
	Not balanced	Balanced
Horizontal crop	Exp 1	Exp 2
Vertical crop	Exp 3	Exp 4
Gaussian blur	Exp 5	Exp 6

3.4 Convolutional Neural Networks

The following architectures were used to classify images of inflammatory bowel diseases: AlexNet, ResNet50, VGG-16, and InceptionV3. These architectures have been pre-trained with the ImageNet dataset. Besides these architectures, another architecture was built that combines a ResNet50 with a MobileNetV2 network. Finally, a hybrid model was built where a CNN is combined with an LSTM.

The architecture that combines the ResNet50 network and the MobileNetV2 network consists of freezing the weights of all layers of the ResNet50 network and extracting the convolutional layers from it [10]. Then the classification layers of the MobileNetV2 network are extracted to classify the features extracted by the convolutional layers of the ResNet50 network. This architecture is illustrated in Figure 5.

The hybrid model built in this work was based on a similar architecture developed by other authors [18, 21, 25]. This architecture is based on: 8 Conv2D layers, 2 BatchNormalization layers, 4 MaxPooling2D layers, 1 Flatten layer, 7 Dense layers, 3 Dropout layers and 3 LSTM layers. The hybrid model is illustrated in Figure 6. LSTM networks contain a forget gate that learns which information should be forgotten and which should be remembered, sending this information to the next layers [21].

The architectures used to classify the images were trained for 200 epochs, using SparseCategoricalCrossentropy as the loss function, and Adam with a learning rate of 1E-05 was set as the optimizer [12]. During training, an EarlyStopping callback was used with patience of 5 to monitor the validation accuracy.

3.5 Evaluation

In all experiments, the dataset was split into 70% for training, 20% for validation, and 10% for testing [5]. In total, of the 64 frames extracted, 30 frames belonged to the same video in which the green square in the lower left corner was found.

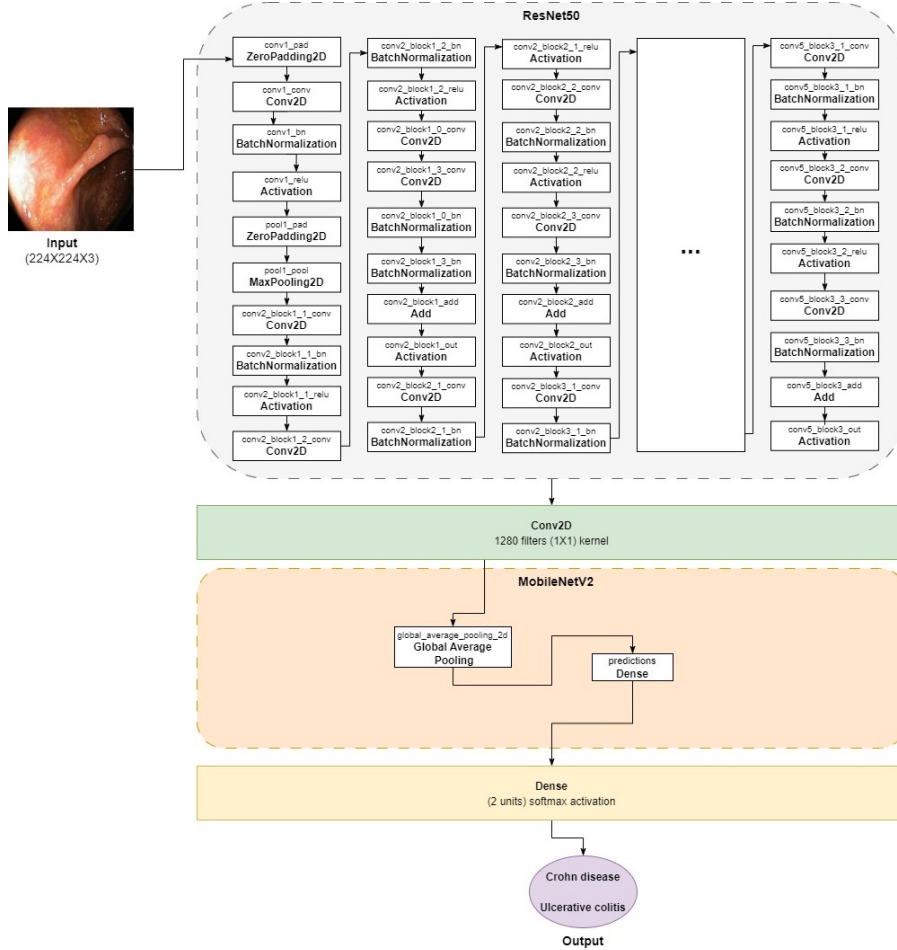


Fig. 5: Architecture of the ResNet50+MobileNetV2 model

These frames were attached to the images in the validation dataset. On the other hand, the remaining frames that belonged to another video and did not have the green square were introduced in the training dataset. Both of the gathered frames have visible ulcerative colitis lesions.

Classification metrics were selected to evaluate the network’s performance on the test dataset. These include Accuracy, Loss, Precision, Recall, F1-Score, Area Under Curve (AUC) and Mathew’s Correlation Coefficient (MCC) [4, 20].

$$Loss = - \sum_{i=1}^n y_i \cdot \log \hat{y}_i$$

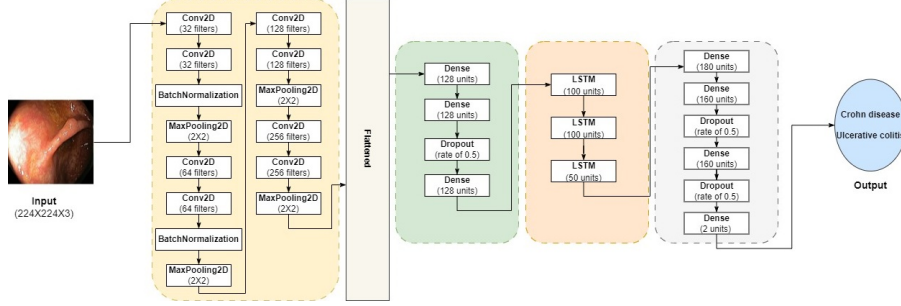


Fig. 6: Architecture of the hybrid model

4 Results

Tables 2 and 3 show the results obtained in experiments 1 and 2, respectively (we refer back to Table 1 for the configurations of each experiment). In Table 3, it is possible to observe the impact that the balanced dataset causes on the results.

Tables 4 and 5 show the results obtained in experiments 3 and 4, respectively. The analysis of the results shows a drop in the results of some architectures when compared to the results of Tables 2 and 3. However, in experiment 4, the hybrid model shows an improvement of 0.70% in Accuracy, 4.46% in Recall, 2.23% in AUC and 4.84% in MCC when comparing the results to experiment 2.

Tables 6 and 7 show the results obtained by the six architectures in experiments 5 and 6, respectively. As can be seen from the results, the chosen architectures perform better in experiment 6 where the classes were balanced.

The experiments performed in this paper serve as a reference for gastroenterologist doctors to improve the process of diagnosing the type of inflammatory bowel disease in patients. Based on the results obtained in the classification metrics, the ResNet50+MobileNetV2 architecture stands out from the others in most of the performed experiments. It is also verified from the values obtained in the AUC metric that the selected architectures demonstrate a good ability to distinguish the classes present in the created datasets, and through the values present in the MCC metric, it is observed that the number of false positives and false negatives is very close to zero or even manages to be zero. Since the MCC metric in most of the experiments is equal to 1.0 and in the remaining ones is very close to 1.0.

Table 2: Results of experiment 1

	Accuracy	Loss	Precision	Recall	F1-Score	AUC	MCC
AlexNet	0.9674	0.1186	0.9733	0.9733	0.9733	0.9795	0.9590
ResNet50	1.0000	0.6359	1.0000	1.0000	1.0000	1.0000	1.0000
VGG-16	0.9767	0.0849	1.0000	0.9333	0.9655	0.9667	0.9493
InceptionV3	0.9907	0.6280	1.0000	0.9733	0.9865	0.9867	0.9796
ResNet50+MobileNetV2	0.9488	0.6428	1.0000	0.8133	0.8971	0.9067	0.8599
Hybrid model	0.9907	0.0770	0.9868	1.0000	0.9934	0.9964	0.9898

Table 3: Results of experiment 2

	Accuracy	Loss	Precision	Recall	F1-Score	AUC	MCC
AlexNet	0.9648	0.0990	0.9932	0.9236	0.9571	0.9578	0.9109
ResNet50	1.0000	0.6326	1.0000	1.0000	1.0000	1.0000	1.0000
VGG-16	0.9894	0.0855	1.0000	0.9427	0.9705	0.9713	0.9382
InceptionV3	1.0000	0.6278	1.0000	1.0000	1.0000	1.0000	1.0000
ResNet50+MobileNetV2	1.0000	0.6427	1.0000	1.0000	1.0000	1.0000	1.0000
Hybrid model	0.9824	0.1000	0.9933	0.9490	0.9707	0.9706	0.9374

Table 4: Results of experiment 3

	Accuracy	Loss	Precision	Recall	F1-Score	AUC	MCC
AlexNet	0.9907	0.0963	0.9762	1.0000	0.9880	0.9925	0.9806
ResNet50	1.0000	0.6324	1.0000	0.9867	0.9933	0.9933	0.9898
VGG-16	0.9628	0.1576	1.0000	0.8800	0.9362	0.9400	0.9093
InceptionV3	0.9953	0.6303	1.0000	1.0000	1.0000	1.0000	1.0000
ResNet50+MobileNetV2	0.9116	0.6553	1.0000	0.8000	0.8889	0.9000	0.8500
Hybrid model	0.9953	0.0814	0.9868	1.0000	0.9934	0.9964	0.9898

Table 5: Results of experiment 4

	Accuracy	Loss	Precision	Recall	F1-Score	AUC	MCC
AlexNet	0.9824	0.0661	0.9805	0.9618	0.9711	0.9691	0.9363
ResNet50	0.9930	0.6395	0.9936	0.9873	0.9904	0.9897	0.9787
VGG-16	0.9824	0.1038	0.9935	0.9745	0.9839	0.9833	0.9647
InceptionV3	0.9930	0.6342	1.0000	1.0000	1.0000	1.0000	1.0000
ResNet50+MobileNetV2	1.0000	0.6457	1.0000	1.0000	1.0000	1.0000	1.0000
Hybrid model	0.9894	0.0715	0.9936	0.9936	0.9936	0.9929	0.9858

Table 6: Results of experiment 5

	Accuracy	Loss	Precision	Recall	F1-Score	AUC	MCC
AlexNet	0.9767	0.0711	1.0000	0.9467	0.9726	0.9733	0.9594
ResNet50	1.0000	0.6261	1.0000	1.0000	1.0000	1.0000	1.0000
VGG-16	0.9814	0.0553	1.0000	0.9600	0.9796	0.9800	0.9695
InceptionV3	0.9953	0.6287	1.0000	0.9733	0.9865	0.9867	0.9796
ResNet50+MobileNetV2	0.9953	0.6386	1.0000	1.0000	1.0000	1.0000	1.0000
Hybrid model	0.9488	0.2109	0.9857	0.9200	0.9517	0.9564	0.9285

Table 7: Results of experiment 6

	Accuracy	Loss	Precision	Recall	F1-Score	AUC	MCC
AlexNet	0.9859	0.0503	1.0000	0.9618	0.9805	0.9809	0.9583
ResNet50	1.0000	0.6312	1.0000	1.0000	1.0000	1.0000	1.0000
VGG-16	0.9542	0.1613	1.0000	0.8662	0.9283	0.9331	0.8622
InceptionV3	1.0000	0.6219	1.0000	1.0000	1.0000	1.0000	1.0000
ResNet50+MobileNetV2	1.0000	0.6372	1.0000	1.0000	1.0000	1.0000	1.0000
Hybrid model	0.9754	0.0830	0.9935	0.9745	0.9839	0.9833	0.9647

5 Conclusion

This study concludes that the use of deep learning architectures is useful in assisting gastroenterologists in the early diagnosis of inflammatory bowel disease. Based on the results of the classification metrics, it can be concluded that the architectures used achieved good accuracy in classifying the disease.

Experiments in which transformations were applied to the images succeeded in maintaining the performance of the architectures. Moreover, this work demonstrates, through the experiments carried out, that applying transformations to the images adds advantages to the classification made by CNNs. It has been discovered that horizontal cropping or the use of a Gaussian blur improves the performance of the architectures.

In future work, it would be important to perform a medical validation with new images of both types of the disease to understand which of the architectures used will be the most accurate in recognising the type of inflammatory bowel disease based on the results obtained and which of the experiments performed will result in better processing.

Also, it is important to use in future work multi-task models to not only predict the type of inflammatory bowel disease but what is the degree of severity according to the Mayo score of ulcerative colitis and what type of lesion a patient with Crohn's disease has in the intestinal mucosa. For this, ordinal techniques can be leveraged [3, 8].

References

1. Borgli, H., Riegler, M., Thambawita, V., Jha, D., Hicks, S., Halvorsen, P.: The HyperKvasir Dataset (2019), publisher: OSF
2. Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., Johansen, D., Griwodz, C., Stensland, H.K., Garcia-Ceja, E., Schmidt, P.T., Hammer, H.L., Riegler, M.A., Halvorsen, P., Lange, T.: HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data* **7**, 283 (2020)
3. Cardoso, J.S., Sousa, R., Domingues, I.: Ordinal data classification using kernel discriminant analysis: A comparison of three approaches. In: 11th Int Conference on Machine Learning and Applications. vol. 1, pp. 473–477 (2012)
4. Chicco, D., Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020)
5. Chierici, M., Puica, N., Pozzi, M., Capistrano, A., Donzella, M.D., Colangelo, A., Osmari, V., Jurman, G.: Automatically detecting Crohn's disease and Ulcerative Colitis from endoscopic imaging. *BMC Medical Informatics and Decision Making* **22**, 300 (2022)
6. CrohniPI, <https://crohniipi.ls2n.fr/en/crohn-ipi-project/> (accessed Feb. 21, 2023)
7. Doença inflamatória do intestino CUF, <https://www.cuf.pt/saude-a-z/doenca-inflamatoria-do-intestino> (accessed Feb. 21, 2023)
8. Domingues, I., Cardoso, J.S.: Max-ordinal learning. *IEEE Transactions on Neural Networks and Learning Systems* **25**(7), 1384–1389 (2014)
9. Dray, X., Li, C., Saurin, J., Cholet, F., Rahmi, G., Mouel, J.L., Leandri, C., Lecleire, S., Amiot, X., Delvaux, J., Duburque, C., Gérard, R., Leenhardt, R., Mesli, F., Vanbiervliet, G., Larmurier, I.N., Sacher-Huvelin, S., Simon-Chane, C., Olivier, R., Histace, A.: CAD-CAP: une base de données française à vocation internationale, pour le développement et la validation d'outils de diagnostic assisté par ordinateur en vidéocapsule endoscopique du grêle. In: *Endoscopy*. p. s-0038–1623358 (2018)
10. Fatema, K., Montaha, S., Rony, M.A.H., Azam, S., Hasan, M.Z., Jonkman, M.: A Robust Framework Combining Image Processing and Deep Learning Hybrid Model to Classify Cardiovascular Diseases Using a Limited Number of Paper-Based Complex ECG Images. *Biomedicines* **10**, 2835 (2022)
11. Ghouri, Y.A., Tahan, V., Shen, B.: Secondary causes of inflammatory bowel diseases. *World J of Gastroenterology* **26**, 3998–4017 (2020)

12. Khan, M.N., Hasan, M.A., Anwar, S.: Improving the Robustness of Object Detection Through a Multi-Camera-Based Fusion Algorithm Using Fuzzy Logic. *Frontiers in Artificial Intelligence* **4**, 638951 (2021)
13. Maissin, A., Vallée, R., Flamant, M., Fondain-Bossiere, M., Berre, C.L., Coutrot, A., Normand, N., Mouchère, H., Coudol, S., Trang, C., Bourreille, A.: Multi-expert annotation of Crohn's disease images of the small bowel for automatic detection using a convolutional recurrent attention neural network. *Endoscopy Int Open* **09**, E1136–E1144 (2021)
14. Polat, G., Kani, H.T., Ergenc, I., Alahdab, Y.O., Temizel, A., Atug, O.: Improving the Computer-Aided Estimation of Ulcerative Colitis Severity According to Mayo Endoscopic Score by Using Regression-Based Deep Learning. *Inflammatory Bowel Diseases* p. izac226 (2022)
15. Polat, G., Kani, H.T., Ergenc, I., Alahdab, Y.O., Temizel, A., Atug, O.: Labeled Images for Ulcerative Colitis (LIMUC) Dataset (2022)
16. Sairenji, T., Collins, K.L., Evans, D.V.: An Update on Inflammatory Bowel Disease. Primary Care: Clinics in Office Practice **44**, 673–692 (2017)
17. Seyedian, S.S., Nokhostin, F., Malamir, M.D.: A review of the diagnosis, prevention, and treatment methods of inflammatory bowel disease. *J of Medicine and Life* **12**, 113–122 (2019)
18. Shahzadi, I., Tang, T.B., Meriadeau, F., Quyyum, A.: CNN-LSTM: Cascaded Framework For Brain Tumour Classification. In: *IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*. pp. 633–637 (2018)
19. Stidham, R.W., Liu, W., Bishu, S., Rice, M.D., Higgins, P.D.R., Zhu, J., Nallamothu, B.K., Waljee, A.K.: Performance of a Deep Learning Model vs Human Reviewers in Grading Endoscopic Disease Severity of Patients With Ulcerative Colitis. *JAMA Network Open* **2**(5), e193963 (2019)
20. Turan, M., Durmus, F.: UC-NfNet: Deep learning-enabled assessment of ulcerative colitis from colonoscopy images. *Medical Image Analysis* **82**, 102587 (2022)
21. Udristoiu, A.L., Stefanescu, D., Gruionu, G., Gruionu, L.G., Iacob, A.V., Karstensen, J.G., Vilman, P., Saftoiu, A.: Deep Learning Algorithm for the Confirmation of Mucosal Healing in Crohn's Disease, Based on Confocal Laser Endomicroscopy Images. *J of Gastrointestinal and Liver Diseases* **30**, 59–65 (2021)
22. Vallée, R., Coutrot, A., Normand, N., Mouchère, H.: Accurate small bowel lesions detection in wireless capsule endoscopy images using deep recurrent attention neural network. In: *IEEE 21st Int WS on Multimedia Signal Proc (MMSP)* (2019)
23. Vallée, R., Coutrot, A., Normand, N., Mouchère, H.: Influence of expertise on human and machine visual attention in a medical image classification task. In: *European Conference on Visual Perception* (2021)
24. Vallée, R., Maissin, A., Coutrot, A., Mouchère, H., Bourreille, A., Normand, N.: CrohnIPI: An endoscopic image database for the evaluation of automatic Crohn's disease lesions recognition algorithms. In: *Medical Imaging: Biomedical Applications in Molecular, Structural, and Functional Imaging*. p. 61. SPIE (2020)
25. Vankdothu, R., Hameed, M.A., Fatima, H.: A Brain Tumor Identification and Classification Using Deep Learning based on CNN-LSTM Method. *Computers and Electrical Engineering* **101**, 107960 (2022)
26. Wehkamp, J., Götz, M., Herrlinger, K., Steurer, W., Stange, E.F.: Inflammatory Bowel Disease: Crohn's disease and ulcerative colitis. *Deutsches Ärzteblatt Int* (2016)

Diagnose of the IBD using Deep Learning Models with Interpretability

"Interpretability of Deep Neural Networks to diagnose Inflammatory Bowel Disease"

Submitted in RECPAD 2023: 29th Portuguese Conference on Pattern Recognition

Interpretability of Deep Neural Networks to diagnose Inflammatory Bowel Disease

José Maurício
a2018056151@isec.pt

Inês Domingues
ines.domingues@isec.pt

Instituto Politécnico de Coimbra,
Instituto Superior de Engenharia,
Rua Pedro Nunes - Quinta da Nora,
3030-199 Coimbra, Portugal
Centro de Investigação do Instituto Português
de Oncologia do Porto (CI-IPOP):
Grupo de Física Médica, Radiobiologia
e Protecção Radiológica

Abstract

The number of patients with inflammatory bowel disease (IBD) has been increasing. The diagnosis is a difficult task for the gastroenterologist performing the endoscopic examination. However, in order to prescribe medical treatment and provide quality of life to the patient, the diagnosis must be quick and accurate. This paper presents a study where the objective is to collect and analyse endoscopic images referring to Crohn's disease and Ulcerative colitis using four deep neural networks: ResNet50, InceptionV3, VGG16, and a hybrid model. The hybrid model consists of the combination of two architectures, a CNN and an LSTM. The main focus is on the understanding of the networks, offering through this paper a comparative study of five interpretability models, Grad-CAM, LIME, SHAP values, RISE, and Occlusion sensitivity. The obtained results demonstrate that it is possible to automate the process of diagnosing patients with IBD using deep networks for processing images collected during an endoscopic examination, and thus develop tools that, with the help of interpretability models, help the medical specialist diagnose the disease, by understanding which region of mucosal the network considered to make a decision.

1 Introduction

Inflammatory bowel disease (IBD) is characterized as a disease of unknown cause that results from the interaction between genetic and environmental factors, triggering an immune response that causes digestive disorders and inflammation in the gastrointestinal tract. These types of diseases are divided into Ulcerative colitis and Crohn's disease [11]. However, some studies have shown that some of the diagnoses of inflammatory bowel disease refer to a different category of disease, called indeterminate colitis. This pathology refers to colitis with an extreme degree of severity that has features of both diseases (Ulcerative colitis and Crohn's disease), making it difficult to distinguish between them [10]. It is estimated that in Portugal it already affects 7,000 to 15,000 people, with 2.9 cases per 100,000 inhabitants per year with Ulcerative colitis and 2.4 cases per 100,000 inhabitants per year with Crohn's disease [4].

The methodology proposed in the present work allows the classification of images through deep neural networks to predict the type of inflammatory bowel disease present in that particular image. In addition, it is also suggested the use of interpretability algorithms of the networks so that the medical specialist can evaluate the region of the mucosal that was considered to make the decision. Towards this end, this study demonstrates a comparison between five interpretability algorithms, namely, Grad-CAM, LIME, SHAP values, RISE and Occlusion sensitivity.

This paper is organized into three sections: section 2 presents the methodology developed during this work and the dataset that was used for the classification; section 3 collects the quantitative results and shows the interpretability of the neural networks; and section 4 summarises the conclusions and directions for future work.

2 Methodology and Data

In order to improve the diagnosis of inflammatory bowel disease, a methodology based on five phases is proposed. In the first phase, the authors collected images of the inflammatory bowel disease (i.e. Ulcerative colitis and Crohn's disease). In the second phase, data augmentation was applied to the training dataset. In the third phase, the convolutional networks were implemented and configured. In the fourth phase, the performance of the CNNs was evaluated using eight different metrics. Finally, the interpretability of the models used was implemented. This methodology is illustrated in Figure 1 and is further explained next.

Dataset: For the development of this study, three datasets of images referring to Crohn's disease and Ulcerative colitis were used, with the goal of creating a single dataset involving images of both diseases. LIMUC [9] dataset was used to get images of the Ulcerative colitis disease, while the CrohniPI [3] dataset was selected to collect images of Crohn's disease. This last dataset has already been used in other works to develop tools for diagnosing the disease [6, 13, 14, 15].

Data Augmentation: Before classifying the images, data augmentation was performed on the train set with horizontal and vertical Random Flip, Random Con-

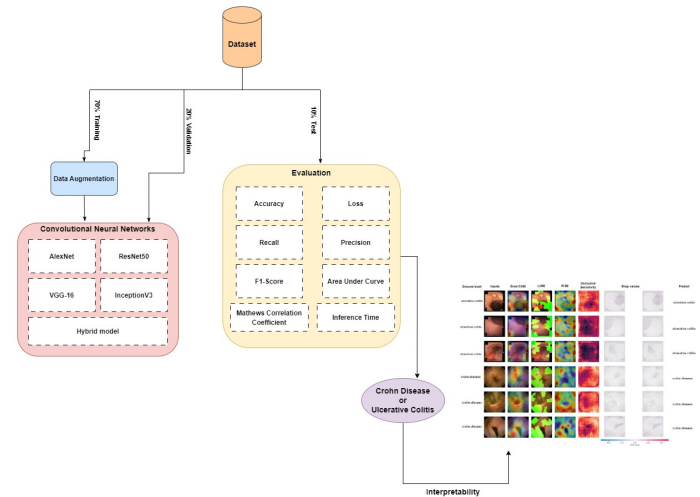


Figure 1: Experimental setup.

trast with a factor of 0.15, Random Rotation with a factor of 0.2 and a Random Zoom with a portion of -0.2 for height and -0.3 for width.

Deep Neural Networks The following architectures were used to classify images of inflammatory bowel disease: ResNet50, VGG16, and InceptionV3. These architectures have been pre-trained with the ImageNet dataset. Besides these architectures, a hybrid model was built where a CNN is combined with an LSTM [7]. The architectures used to classify the images were trained for 200 epochs, using SparseCategoricalCrossentropy as the loss function, a batch size of 32, and Adam with a learning rate of $1E-05$ was set as the optimizer [5]. During training, an EarlyStopping callback was used with the patience of 5 to monitor the validation accuracy. The images were initially resized to 224×224 pixels.

Evaluation In all experiments, the dataset was split into 70% for training, 20% for validation, and 10% for testing [2]. Classification metrics were selected to evaluate the network's on the test dataset. These include Accuracy, Loss, Precision, Recall, F1-Score, Area Under Curve (AUC), Mathew's Correlation Coefficient (MCC) and Inference time [1, 12].

Interpretability In order to interpret the output of the neural networks and gain insights into the key features considered important for predicting outcomes, several algorithms were chosen. These include Grad-CAM, LIME, SHAP values, RISE, and Occlusion sensitivity. These algorithms play a vital role in analyzing and determining which specific parts of the images contribute significantly to the predicted output.

3 Results

In conducting this study, Tensorflow, version 2.8.0, Tfimm, version 0.6.13, SHAP, version 0.40.0, were used. The programming environment for importing the libraries was Google Colab with the NVIDIA A100 GPU. Table 1 shows the results obtained by all architectures used in this study. Figures 2- 5 illustrate the results of each interpretability method for each CNN architecture.

Table 1 shows that the hybrid model was the architecture with the worst results among all the networks selected for this study. Even so, based on the AUC value, it shows a good ability to distinguish classes. It is also concluded that due to the complexity of the networks used in this study, the inference time is greater than 2 seconds.

After analysing the result of the network's interpretability, in Figures 2- 5, it can be observed that they tend to interpret some characteristics of the images as decisive elements to predict the disease. The LIME algorithm and the Occlusion sensitivity, for the networks InceptionV3 and VGG16 have the ability to recognise in ulcerative colitis images zones of bleeding, as well as ulcers. In Crohn's disease, it can recognise the lesions caused by the disease. Shap values and the RISE model on the same networks have also shown some evidence of this effect.

Table 1: Results obtained through the experience.

	ResNet50	VGG16	InceptionV3	Hybrid model
Acc	1.0000	1.0000	1.0000	0.9926
Loss	0.0044	0.0000	0.0067	0.3328
Recall	1.0000	1.0000	1.0000	1.0000
Precision	1.0000	1.0000	1.0000	0.9852
F1-Score	1.0000	1.0000	1.0000	0.9925
AUC	1.0000	1.0000	1.0000	0.9927
MCC	1.0000	1.0000	1.0000	0.9853
Inference time	2.44s	3.43s	2.94s	2.23s

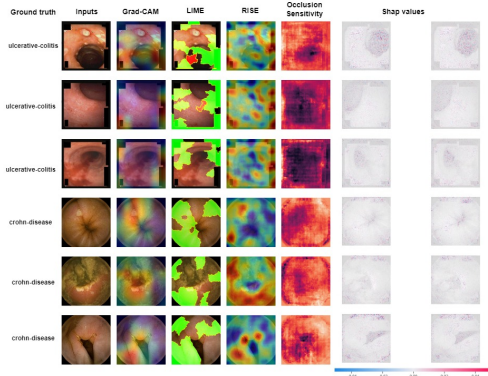


Figure 2: Interpretability and predicted output by the ResNet50 network.

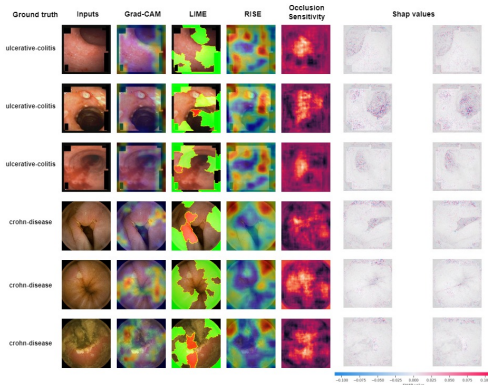


Figure 3: Interpretability and predicted output by the VGG16 network.

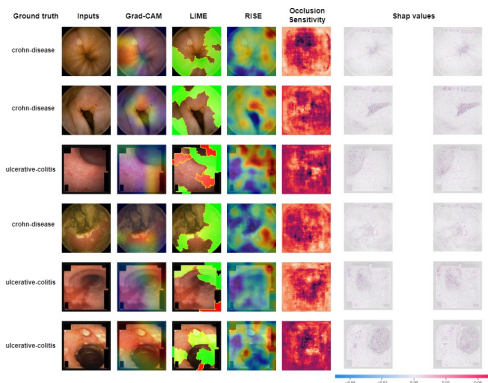


Figure 4: Interpretability and predicted output by the InceptionV3 network.

4 Conclusions

The findings of this study highlight the utility of deep learning architectures in assisting gastroenterologists with the early diagnosis of inflammatory bowel disease. The classification metrics demonstrate that the employed architectures achieved a good level of accuracy in disease classification.

However, upon analyzing the interpretability of the architectures utilized in this study, it becomes evident that certain interpretability models partially or even completely disregard the mucosal part of the intestine.

Future research will prioritize medical validation using new images encompassing both types of disease. This validation will enable a comprehensive under-

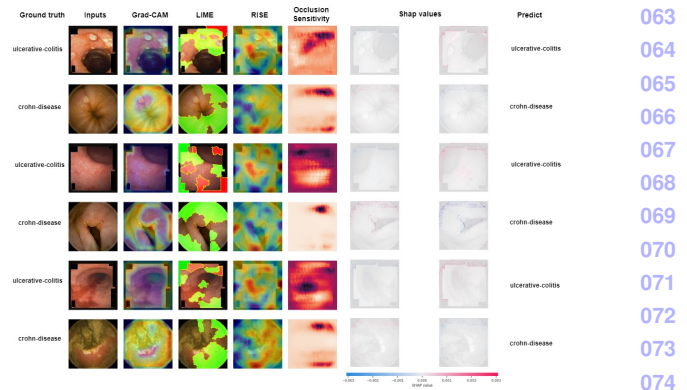


Figure 5: Interpretability and predicted output by the hybrid model.

standing of which architectures prove most interpretable in identifying the specific type of inflammatory bowel disease, based on the obtained results and the insights of experienced gastroenterologists. Also, they could develop studies where vision transformers are applied to the classification of images as an alternative to CNNs. These models demonstrate to be accurate and have a good performance in the classification of images [8].

References

- [1] D Chicco and G Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21:6, 2020.
- [2] M Chierici, N Puica, M Pozzi, A Capistrano, MD Donzella, A Colangelo, V Osmani, and G Jurman. Automatically detecting Crohn's disease and Ulcerative Colitis from endoscopic imaging. *BMC Medical Informatics and Decision Making*, 22(S6):300, 2022.
- [3] CrohnIPI. CrohnIPI. <https://crohnipi.ls2n.fr/en/crohn-ipi-project/> (accessed Feb. 21, 2023).
- [4] Doença inflamatória do intestino CUF. Doença inflamatória do intestino CUF. <https://www.cuf.pt/saude-a-z/doenca-inflamatoria-do-intestino> (accessed Feb. 21, 2023).
- [5] MN Khan, MA Hasan, and S Anwar. Improving the Robustness of Object Detection Through a Multi-Camera-Based Fusion Algorithm Using Fuzzy Logic. *Frontiers in Artificial Intelligence*, 4:638951, 2021.
- [6] A Maissin, R Vallée, M Flamant, M Fondain-Bossiere, CL Berre, A Coutrot, N Normand, H Mouchère, S Coudol, C Trang, and A Bourreille. Multi-expert annotation of Crohn's disease images of the small bowel for automatic detection using a convolutional recurrent attention neural network. *Endoscopy Int Open*, 09:E1136–E1144, 2021.
- [7] J Mauricio and I Domingues. Deep Neural Networks to distinguish between Crohn's disease and Ulcerative colitis. In *11th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)*, 2023.
- [8] J Mauricio, I Domingues, and J Bernardino. Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences*, 13(9):5521, 2023.
- [9] G Polat, HT Kani, IE, YO Alahdab, A Temizel, and O Atug. Labeled Images for Ulcerative Colitis (LIMUC) Dataset, 2022.
- [10] T Sairenji, KL Collins, and DV Evans. An Update on Inflammatory Bowel Disease. *Primary Care: Clinics in Office Practice*, 44:673–692, 2017.
- [11] SS Seyedian, F Nokhostin, and MD Malamir. A review of the diagnosis, prevention, and treatment methods of inflammatory bowel disease. *J of Medicine and Life*, 12:113–122, 2019.
- [12] M Turan and F Durmus. UC-NfNet: Deep learning-enabled assessment of ulcerative colitis from colonoscopy images. *Medical Image Analysis*, 82: 102587, 2022.
- [13] R Vallée, A Coutrot, N Normand, and H Mouchère. Accurate small bowel lesions detection in wireless capsule endoscopy images using deep recurrent attention neural network. In *IEEE 21st Int WS on Multimedia Signal Proc (MMSP)*, 2019.
- [14] R Vallée, A Maissin, A Coutrot, H Mouchère, A Bourreille, and N Normand. CrohnIPI: An endoscopic image database for the evaluation of automatic Crohn's disease lesions recognition algorithms. In *Medical Imaging: Biomedical Applications in Molecular, Structural, and Functional Imaging*, page 61. SPIE, 2020. ISBN 9781510634015.
- [15] R Vallée, A Coutrot, N Normand, and H Mouchère. Influence of expertise on human and machine visual attention in a medical image classification task. In *European Conference on Visual Perception*, 2021.

José Carlos Nunes Maurício

"Knowledge Distillation of Vision Transformers and Convolutional Networks to Predict Inflammatory Bowel Disease"

Submitted in CIARP 2023: 26th Iberoamerican Congress on Pattern Recognition

Knowledge Distillation of Vision Transformers and Convolutional Networks to Predict Inflammatory Bowel Disease

José Maurício¹[0000–0001–8234–9481] and Inês Domingues^{1,2}[0000–0002–2334–7280]

¹ Instituto Politécnico de Coimbra, Instituto Superior de Engenharia, Rua Pedro Nunes - Quinta da Nora, 3030-199 Coimbra, Portugal

² Centro de Investigação do Instituto Português de Oncologia do Porto (CI-IPOP): Grupo de Física Médica, Radiobiologia e Protecção Radiológica
{a2018056151, ines.domingues}@isec.pt

Abstract. Inflammatory bowel disease is a chronic disease of unknown cause that can affect the entire gastrointestinal tract, from the mouth to the anus. It is important for patients with this pathology that a good diagnosis is made as early as possible, so that the inflammation present in the mucosa intestinal is controlled and the most severe symptoms are reduced, thus offering the quality of life to people. Therefore, through this comparative study, we seek to find a way of automating the diagnosis of these patients during the endoscopic examination, reducing the subjectivity that is subject to the observation of a gastroenterologist, using six CNNs: AlexNet, ResNet50, VGG16, ResNet50-MobileNetV2 and Hybrid model. Also, five ViTs were used in this study: ViT-B/32, ViT-S/32, ViT-B/16, ViT-S/16 and R26+S/32. This comparison also consists in applying knowledge distillation to build simpler models, with fewer parameters, based on the learning of the pre-trained architectures on large volumes of data. It is concluded that in the ViTs framework, it is possible to reduce 25x the number of parameters by maintaining good performance and reducing the inference time by 5.32 seconds. For CNNs the results show that it is possible to reduce 107x the number of parameters, reducing consequently the inference time in 3.84 seconds.

Keywords: Inflammatory Bowel Disease · Knowledge Distillation · Vision Transformers · CNN · DeiT.

1 Introduction

Crohn's disease and ulcerative colitis are two separate pathologies that are joined by inflammatory bowel disease, which is characterised by chronic inflammation throughout the intestinal tract. With 2.9 cases of ulcerative colitis per 100,000 residents per year and 2.4 cases of Crohn's disease per 100,000 residents per year in Portugal, it is estimated that it already affects 7,000 to 15,000 people there [8].

Crohn's disease affects any part of the gastrointestinal tract, from the mouth to the anus. The severity of the disease and its location determine the signs and

symptoms, which have a wide spectrum of clinical presentations [11]. Ulcerative colitis is restricted to the large intestine and involves a superficial inflammatory process. This disease only affects the colon and rectum. Whereas Crohn’s disease affects any part of the gastrointestinal tract. Normally, the inflammation starts in the rectum and can be continuously extended to the terminal ileum, where it is terminated by backwash ileitis [25].

The use of deep learning models to help the gastroenterologist in the diagnosis of patients with inflammatory bowel disease is important and, with time, more necessary. Not only in Portugal but worldwide the number of people who have this chronic disease is increasing and its diagnosis depends on the evaluation that the gastroenterologist makes of the endoscopic images, which makes this diagnosis subject to a great subjectivity [24]. Therefore, it is crucial to find tools that automate the diagnosis of the disease to prescribe the most suitable treatment for the patient.

These tools need to be fast in recognising the disease through endoscopic images, as well as required to be light and less complex to occupy fewer computational resources when deployed in medical systems. But that they guarantee the same accuracy in predicting inflammatory bowel disease. In this sense, this study seeks to make a comparative study, where six CNNs and five ViTs are used to process images collected using colonoscopy and video capsule endoscopic to recognise which type of inflammatory bowel disease is present: Crohn’s disease and Ulcerative Colitis. Together with knowledge distillation to simpler models that guarantee similar performance [1] as pre-trained models on large volumes of data.

The organization this paper is divided into four sections: section 2 describes the created datasets for this study and the methodology implemented; section 3 presents the quantitative results obtained by the teachers and student’s models; section 4 compare the results obtained in this study with the results of the other authors and critically assesses the findings made; and section 5 summarises the findings and suggests some future research directions.

2 Methodology and Data

Aiming to facilitate the deployment of deep models in embedded systems to automate and improve the diagnosis of inflammatory bowel disease, this work suggests a methodology based on six phases.

In the first phase, the authors collected images of the inflammatory bowel disease (i.e. Ulcerative colitis and Crohn’s disease). The videos captured during the patients’ endoscopic examination were used in the second phase to extract frames. In the third phase, the images were pre-processed to fill in some undesired characteristics. In the fourth phase, different data augmentation techniques were applied to the training dataset. In the fifth phase, the convolutional networks and vision transformers were implemented and configured. In the sixth phase, the performance of the CNNs and ViTs was evaluated using six classification

metrics, the Loss and the Inference time. Figure 1 illustrates the methodology proposed in this paper and in the following subsections will be explained.

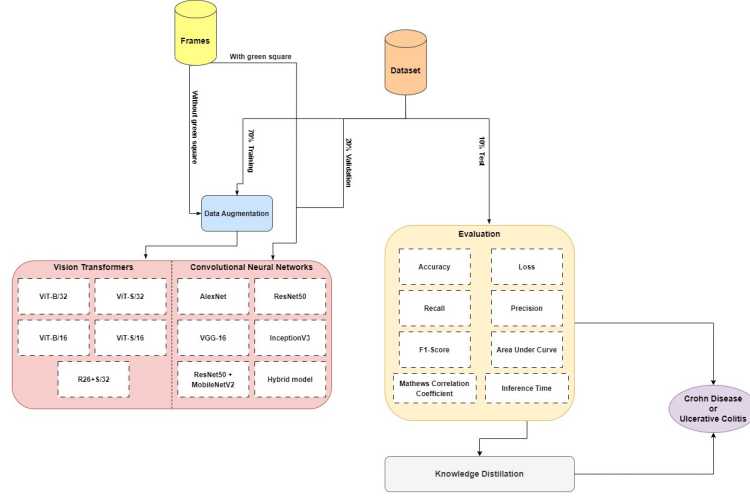


Fig. 1: Proposed methodology.

2.1 Dataset

Three datasets of images related to Crohn’s disease and Ulcerative colitis were used to conduct this study, with the aim of combining the images from the two diseases into a single dataset. HyperKvasir [2, 3] and LIMUC [23] datasets were used to get images of the Ulcerative colitis disease, while the CrohnIPI [7, 32, 33] dataset was selected to collect images of Crohn’s disease.

Figure 2 shows an example of the images that exist in each collected dataset. Therefore, two different datasets were created:

- **Dataset 1:** this dataset compiles 850 images of ulcerative colitis from the HyperKvasir dataset and 1360 images of Crohn’s disease from the CrohnIPI dataset. It also includes 64 video frames that were taken from the HyperKvasir dataset’s labelled videos. Figure 3a shows the class distribution of this dataset;
- **Dataset 2:** this dataset consists of 446 instances of the LIMUC dataset, the instances are equally distributed by severity degree. That is, 25% of the total images referring to the severity degree were extracted from the original dataset; 64 video frames were extracted from the labelled videos from the HyperKvasir dataset; 850 instances from the HyperKvasir dataset; and 1360 instances from the CrohnIPI dataset. Figure 3b shows the class distribution of this dataset.

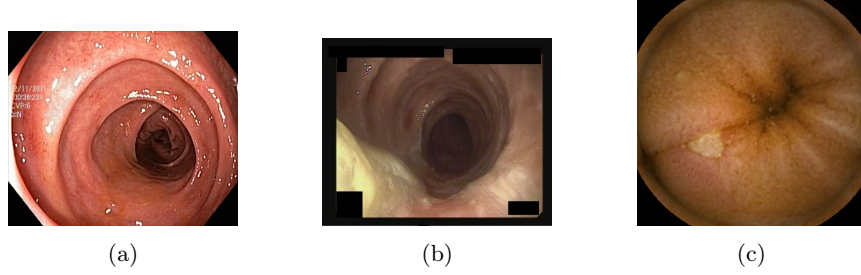


Fig. 2: Image Samples: (a) Ulcerative colitis image from the HyperKvasir dataset; (b) Ulcerative colitis image from the LIMUC dataset; (c) Crohn's disease image from the CrohnIPI dataset.

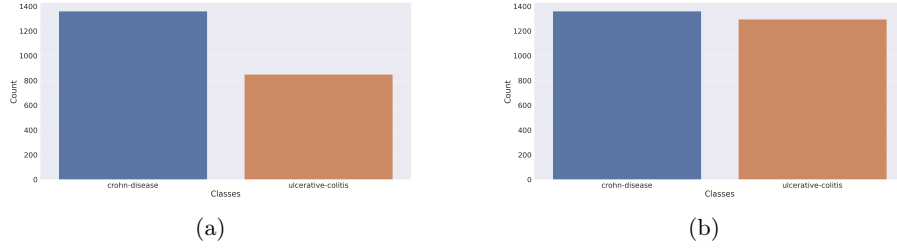


Fig. 3: Class distribution: (a) Dataset 1; (b) Dataset 2.

2.2 Experimental setup

Tensorflow, version 2.8.0, Tfimm, version 0.6.13, and Shap, version 0.40.0, were utilised to carry out this investigation. The NVIDIA A100 GPU was used in conjunction with Google Colab as the programming environment for importing the libraries.

Based on the datasets created, a total of six experiments were realized, allowing different combinations. Therefore, it started with developing three experiments, where the first experiment consisted in using dataset 1. The second experiment had the aim of balancing the classes and diversifying the images regarding class ulcerative colitis, for which was used the dataset 2. Table 1 summarizes the experiments realized, as well as pre-processing techniques (see section 2.3) and dataset used.

2.3 Pre-processing

The position of the endoscope during the diagnostic examination is indicated by a green square in the lower-left corner of the HyperKvasir dataset's images and frames extracted from a video. This might make it difficult to identify the specific type of inflammatory bowel disease depicted in the images given to CNN. When this kind of circumstance was confirmed, experiments were conducted where

Table 1: Configurations of the six performed experiments

Experiment	Pre-processing	Dataset
Experiment 1	Horizontal crop	Dataset 1
Experiment 2	Horizontal crop	Dataset 2
Experiment 3	Vertical crop	Dataset 1
Experiment 4	Vertical crop	Dataset 2
Experiment 5	Gaussian blur	Dataset 1
Experiment 6	Gaussian blur	Dataset 2

some transformations were applied. One of the applied transformations was a horizontal crop across the entire width of the image to remove the green square in the lower left corner. Then a few more experiments were performed, where a vertical crop was applied over the entire height of the image to remove that part of the image. However, during these transformations, some information regarding the mucosa of the intestine is lost. Some more experiments were performed in which a Gaussian blur of 2/16 was applied only to the area where the green square was.

The goal was to preserve as much of the image information as possible but to disguise that feature in some images. An illustration of these pre-processing methods is available in Figure 4.

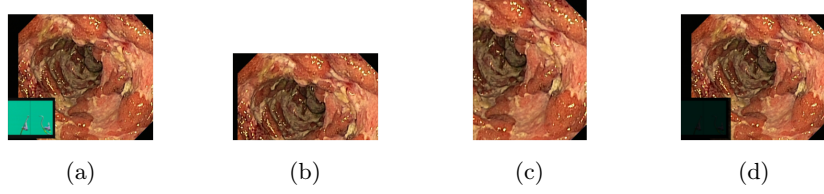


Fig. 4: Pre-processing examples: (a) Original image; (b) Horizontal crop; (c) Vertical crop; (d) Gaussian blur.

2.4 Data augmentation

Before classifying the images, data augmentation was performed on the train set with horizontal and vertical Random Flip, Random Contrast with a factor of 0.15, Random Rotation with a factor of 0.2 and a Random Zoom with a portion of -0.2 for height and -0.3 for width.

2.5 Deep Learning models

The CNN networks selected for this study were the following: ResNet50, VGG16, and InceptionV3. These architectures have been pre-trained with the ImageNet

dataset. Besides these architectures, an ensemble model was built that combines a ResNet50 with a MobileNetV2 network. A hybrid model that combines a CNN with an LSTM was also considered, as well as AlexNet was also used.

The architecture that combines the ResNet50 network and the MobileNetV2 network consists in an ensemble model, where the two architectures were combined. That is, the input images will be directed to the two architectures that will process these images without the last classification layer and the output of the two will be concatenated to be classified by a Dense layer with 2 neurons referring to class Crohn disease and ulcerative colitis. The construction of this model was inspired by [15, 13].

The hybrid model built in this work was based on a similar architecture developed by other authors [26, 30, 34]. This architecture is based on: 8 Conv2D layers, 2 BatchNormalization layers, 4 MaxPooling2D layers, 1 Flatten layer, 7 Dense layers, 3 Dropout layers and 3 LSTM layers. The hybrid model is illustrated in Figure 5.

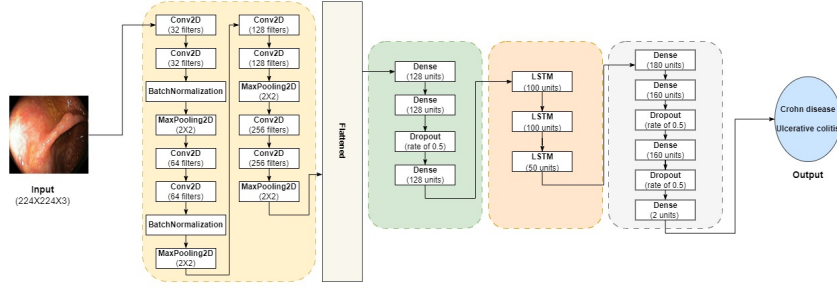


Fig. 5: The hybrid model's architecture [20].

Within the vision transformers, the following models were used: ViT-B/32, ViT-S/32, ViT-B/16, ViT-S/16 and R26+S/32. These models were pre-trained with the Imagenet21k dataset. The R26+S/32 model was created by Google and made available, as well as other versions very similar to this one, which consists of combining a ResNet network with a vision transformer [27]. That is, the first part of the model name (R26) represents the number of convolutional layers that were used from the ResNet network and the second part of the name (S/32) means that those layers extracted from the ResNet network were combined with a small vision transformer of 32 patches.

Using SparseCategoricalCrossentropy as the loss function, a batch size of 32, and Adam with a learning rate of 1E-05 as the optimizer, the models used to classify the pictures were trained for 200 epochs [16]. An EarlyStopping callback with patience of 5 was utilized during training to track the validity accuracy. The images were resized to 224x224 pixels in all experiments.

2.6 Knowledge Distillation

Deploying large architectures into medical systems to assist gastroenterologists in diagnosing the disease can be very complex and even impractical due to the time it takes to process the images, as well as the computational resources required to store these models. Therefore, this sub-section describes the process of distilling knowledge from core architectures to simpler and lighter architectures.

For the CNNs, a student with a total of 1,254,194 parameters was created, with the Adam optimizer using a learning rate of $1e-04$ and the SparseCategoricalCrossentropy loss function. During distillation, the KLDivergence function was set as the loss function, with an alpha of 0.8 and a temperature of 105 degrees. For student training, 100 epochs and an EarlyStopping callback with the patience of 5 were set to monitor the validation accuracy.

In the case of ViTs, the distillation process was done in two different ways: i) using the temperature technique to scale the weights returned by the teacher's Softmax function, as shown in Figure 6; ii) using the DeiT technique. In the implementation of the first technique, an architecture with a total of 2,300,162 parameters was built to distil architectures with 32 patches and an architecture with a total of 3,442,274 parameters to distil architectures with 16 patches. The same configurations used in the distillation of the CNNs were considered.

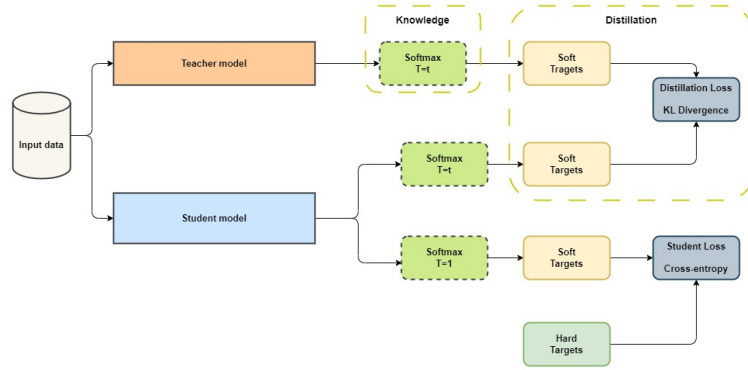


Fig. 6: Example of the Response-base knowledge, based on [14].

The second method involved building an architecture with 3,554,884 total parameters to distil the architectures with 32 patches and an architecture with 3,112,516 total parameters to distil the architectures with 16 patches. These students had the AdamW optimizer set up with a weight decay of 0.0001 and a scaled learning rate of $6.25e-07$, as determined by the following formula:

$$lr_{scaled} = \frac{lr}{512} \times batchsize$$

The learning rate (lr) is equal to $1.00e-05$ and the batch size is 32. The student’s loss function and the distillation function were both used to define the SparseCategoricalCrossentropy function. Using an EarlyStopping callback and a patience of 5, the students were trained for 100 epochs. Data augmentation was also used in both of the ViTs’ distillation methods and in CNNs’ knowledge distillation [9, 17].

2.7 Evaluation

The dataset used in all experiments was divided into three sets: training (70%), validating (20%) and testing (10%) [6].

As described in section 2.1, the datasets include 64 video frames taken from the HyperKvasir dataset’s labelled videos. Thirty of these 64 extracted frames present a green square in the lower left corner. All of these frames are connected to the validation dataset. The remaining frames from the extra video that were missing the green square were included in the training dataset. We used classification metrics to assess the performance of the CNNs and ViTs on the test dataset, as well as the performance of the students. These include Accuracy, Precision, Recall, F1-Score, Area Under Curve (AUC), Mathew’s Correlation Coefficient (MCC), the inference time [5, 29], and the Loss:

$$Loss = - \sum_{i=1}^n y_i \cdot \log \hat{y}_i$$

3 Results

This section shows the results obtained from all experiments. Table 2 and 3 show the results obtained by the CNNs and ViTs in experiment 1, respectively.

Table 2: Results of the CNNs in experiment 1.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	AlexNet	1.0000	0.0031	1.0000	1.0000	1.0000	1.0000	1.0000	0.75s	46,760,706
	ResNet50	1.0000	0.0297	1.0000	1.0000	1.0000	1.0000	1.0000	3.65s	23,591,810
	VGG16	1.0000	0.0016	1.0000	1.0000	1.0000	1.0000	1.0000	3.80s	134,268,738
	InceptionV3	1.0000	0.0596	1.0000	1.0000	1.0000	1.0000	1.0000	2.95s	21,806,882
	ResNet50+MobileNetV2	1.0000	0.0429	1.0000	1.0000	1.0000	1.0000	1.0000	4.20s	25,852,354
Knowledge Distillation	Hybrid model	0.9953	0.4380	0.9877	1.0000	0.9938	0.9938	0.9901	2.31s	1,930,966
	AlexNet student	1.0000	0.0026	1.0000	1.0000	1.0000	1.0000	1.0000	2.51s	1,254,194
	ResNet50 student	0.3991	1.3687	1.0000	0.3876	0.5586	0.5152	0.1084	2.06s	1,254,194
	VGG16 student	0.4272	2.6307	1.0000	0.3990	0.5704	0.5379	0.1739	1.61s	1,254,194
	InceptionV3 student	1.0000	0.0702	1.0000	1.0000	1.0000	1.0000	1.0000	1.45s	1,254,194
	ResNet50+MobileNetV2 student	0.9953	0.0476	1.0000	0.9878	0.9939	0.9962	0.9901	1.69s	1,254,194

In the tables, it can be observed that both deep learning models managed to obtain a good performance in image classification and through knowledge distillation it was possible to reduce the inference time in most models, except in the AlexNet network. However, in this experiment, it is verified that the ResNet50 and VGG16 students were not able to correctly learn the information distilled by the teacher model.

Table 3: Results of the ViTs in experiment 1.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	ViT-B/32	1.0000	0.0011	1.0000	1.0000	1.0000	1.0000	1.0000	3.07s	87,417,602
	ViT-S/32	1.0000	0.0007	1.0000	1.0000	1.0000	1.0000	1.0000	2.63s	22,475,138
	ViT-B/16	1.0000	0.0029	1.0000	1.0000	1.0000	1.0000	1.0000	5.25s	86,417,594
	ViT-S/16	0.9953	0.0097	1.0000	0.9878	0.9939	0.9962	0.9901	3.26s	21,590,402
	R26+S/32	0.9953	0.0169	1.0000	0.9878	0.9939	0.9962	0.9901	5.88s	36,028,098
Knowledge Distillation	ViT-B/32 student	0.9624	0.1138	0.9012	1.0000	0.9481	0.9506	0.9218	1.81s	2,300,162
	ViT-S/32 student	0.9859	0.0251	1.0000	0.9643	0.9818	0.9886	0.9708	1.31s	2,300,162
	ViT-B/16 student	0.9953	0.0132	1.0000	0.9878	0.9939	0.9962	0.9901	1.74s	3,442,274
	ViT-S/16 student	0.9906	0.0147	0.9753	1.0000	0.9875	0.9877	0.9802	0.91s	3,442,274
	R26+S/32 student	0.9953	0.0097	1.0000	0.9878	0.9939	0.9962	0.9901	1.42s	2,300,162
DeiT	ViT-B/32 student	0.9671	0.0636	0.9506	0.9625	0.9565	0.9639	0.9302	0.93s	3,554,884
	ViT-S/32 student	0.9671	0.0983	0.9136	1.0000	0.9548	0.9568	0.9314	0.91s	3,554,884
	ViT-B/16 student	0.9577	0.1023	0.8889	1.0000	0.9412	0.9444	0.9122	1.03s	3,112,516
	ViT-S/16 student	0.9624	0.0849	0.9012	1.0000	0.9481	0.9506	0.9218	1.01s	3,112,516
	R26+S/32 student	0.9531	0.1419	0.8765	1.0000	0.9342	0.9383	0.9027	0.92s	3,554,884

Tables [4 - 5] present the results of experiment 2 regarding CNN networks and ViT models, in the same order. Based on obtained the results, it is perceived by the value of the AUC metric, that the used architectures present a good ability to distinguish the images of the classes under study. Comparing the distillation results obtained by CNNs and ViTs it is concluded that the students of CNNs can perform better in classifying images of inflammatory bowel disease. And when we compare the results between the two distillation techniques of the ViT models we realize that the results worsened in the distillation through attention, having for example the model ViT-S/16 reduced 4.96% in Accuracy, 5.19% in AUC and 9.67% in MCC metric. But inference time was shorter using the DeiT distillation technique.

Table 4: Results of the CNNs in experiment 2.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	AlexNet	0.9858	0.0230	0.9852	0.9852	0.9852	0.9858	0.9716	1.08s	46,760,706
	ResNet50	0.9965	0.0339	1.0000	0.9926	0.9963	0.9966	0.9929	2.68s	23,591,810
	VGG16	0.9965	0.0038	0.9926	1.0000	0.9963	0.9963	0.9929	5.60s	134,268,738
	InceptionV3	0.9929	0.0408	1.0000	0.9854	0.9926	0.9932	0.9859	3.96s	21,806,882
	ResNet50+MobileNetV2	1.0000	0.0177	1.0000	1.0000	1.0000	1.0000	1.0000	4.18s	25,852,354
Knowledge Distillation	Hybrid model	0.9929	0.3326	1.0000	0.9854	0.9926	0.9932	0.9859	3.31s	1,930,960
	AlexNet student	0.9752	0.0998	0.9926	0.9571	0.9745	0.9759	0.9509	2.10s	1,254,194
	ResNet50 student	0.9610	0.1058	0.9185	1.0000	0.9575	0.9593	0.9244	1.73s	1,254,194
	VGG16 student	0.9929	0.1006	0.9926	0.9926	0.9926	0.9929	0.9858	1.76s	1,254,194
	InceptionV3 student	0.9823	0.1043	1.0000	0.9643	0.9818	0.9830	0.9651	3.21s	1,254,194
	ResNet50+MobileNetV2 student	0.9894	0.0738	0.9926	0.9853	0.9889	0.9895	0.9787	1.69s	1,254,194

Table 5: Results of the ViTs in experiment 2.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	ViT-B/32	0.9965	0.0116	0.9926	1.0000	0.9963	0.9963	0.9929	3.42s	87,417,602
	ViT-S/32	0.9965	0.0070	1.0000	0.9926	0.9963	0.9966	0.9929	2.87s	22,475,138
	ViT-B/16	1.0000	0.0004	1.0000	1.0000	1.0000	1.0000	1.0000	6.49s	86,417,594
	ViT-S/16	1.0000	0.0065	1.0000	1.0000	1.0000	1.0000	1.0000	3.67s	21,590,402
	R26+S/32	1.0000	0.0042	1.0000	1.0000	1.0000	1.0000	1.0000	6.45s	36,028,098
Knowledge Distillation	ViT-B/32 student	0.9539	0.1103	0.9704	0.9357	0.9527	0.9546	0.9083	1.76s	2,300,162
	ViT-S/32 student	0.9574	0.1615	0.9111	1.0000	0.9535	0.9556	0.9178	1.00s	2,300,162
	ViT-B/16 student	0.9894	0.0413	0.9778	1.0000	0.9888	0.9889	0.9789	1.22s	3,442,274
	ViT-S/16 student	0.9929	0.0376	1.0000	0.9854	0.9926	0.9932	0.9859	1.04s	3,442,274
	R26+S/32 student	0.9504	0.2228	0.9926	0.9116	0.9504	0.9521	0.9042	1.06s	2,300,162
DeiT	ViT-B/32 student	0.9362	0.1776	0.8963	0.9680	0.9308	0.9345	0.8739	1.08s	3,554,884
	ViT-S/32 student	0.9468	0.1330	0.9111	0.9762	0.9425	0.9454	0.8950	1.10s	3,554,884
	ViT-B/16 student	0.9397	0.2261	0.8815	0.9917	0.9333	0.9373	0.8837	1.16s	3,112,516
	ViT-S/16 student	0.9433	0.2013	0.8963	0.9837	0.9380	0.9413	0.8892	1.16s	3,112,516
	R26+S/32 student	0.9291	0.2303	0.8667	0.9832	0.9213	0.9265	0.8629	1.04s	3,554,884

Tables [6 - 13] show the results obtained by the teacher and student models from experiments 3 to 6. The results represented in the tables express a behaviour similar to the one described during experiment 2. Except that during experiment 3 the model R26+S/32 had a performance improvement with the distillation of knowledge through attention. The same was described by models ViT-B/32 and ViT-S/32 in experiment 5. Also in the experiment, it is verified that the hybrid model of CNNs did not obtain a good performance. In addition, in each experiment using the same distillation technique and when comparing inference time between ViTs and CNNs, it was possible to substantially reduce the inference time in the ViTs models. Although the student models present a higher number of parameters than the CNNs' student model.

Table 6: Results of the CNNs in experiment 3.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	AlexNet	1.0000	0.0024	1.0000	1.0000	1.0000	1.0000	1.0000	0.78s	46,760,706
	ResNet50	1.0000	0.0376	1.0000	1.0000	1.0000	1.0000	1.0000	2.27s	23,591,810
	VGG16	1.0000	0.0000	1.0000	1.0000	1.0000	1.0000	1.0000	3.56s	134,268,738
	InceptionV3	0.9906	0.0644	1.0000	0.9759	0.9878	0.9924	0.9804	2.94s	21,806,882
	ResNet50+MobileNetV2	1.0000	0.0338	1.0000	1.0000	1.0000	1.0000	1.0000	3.61s	25,852,354
	Hybrid model	1.0000	0.3582	1.0000	1.0000	1.0000	1.0000	1.0000	2.43s	1,930,966
Knowledge Distillation	AlexNet student	0.9906	0.0207	0.9753	1.0000	0.9875	0.9877	0.9802	2.15s	1,254,194
	ResNet50 student	0.9859	0.0625	0.9877	0.9756	0.9816	0.9863	0.9702	1.72s	1,254,194
	VGG16 student	1.0000	0.0006	1.0000	1.0000	1.0000	1.0000	1.0000	1.43s	1,254,194
	InceptionV3 student	0.9906	0.0711	1.0000	0.9759	0.9878	0.9924	0.9804	1.70s	1,254,194
	ResNet50+MobileNetV2 student	0.9906	0.0758	1.0000	0.9759	0.9878	0.9924	0.9804	1.90s	1,254,194

Table 7: Results of the ViTs in experiment 3.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	ViT-B/32	1.0000	0.0003	1.0000	1.0000	1.0000	1.0000	1.0000	3.21s	87,417,602
	ViT-S/32	1.0000	0.0040	1.0000	1.0000	1.0000	1.0000	1.0000	2.91s	22,475,138
	ViT-B/16	1.0000	0.0035	1.0000	1.0000	1.0000	1.0000	1.0000	5.62s	86,417,594
	ViT-S/16	1.0000	0.0008	1.0000	1.0000	1.0000	1.0000	1.0000	3.33s	21,590,402
	R26+S/32	1.0000	0.0033	1.0000	1.0000	1.0000	1.0000	1.0000	6.29s	36,028,098
	ViT-B/32 student	0.9953	0.0038	0.9877	1.0000	0.9938	0.9938	0.9901	1.70s	2,300,162
Knowledge Distillation	ViT-S/32 student	0.9906	0.0395	0.9753	1.0000	0.9875	0.9877	0.9802	1.58s	2,300,162
	ViT-B/16 student	0.9906	0.0509	0.9753	1.0000	0.9875	0.9877	0.9802	1.15s	3,442,274
	ViT-S/16 student	0.9906	0.0399	0.9753	1.0000	0.9875	0.9877	0.9802	1.65s	3,442,274
	R26+S/32 student	0.9061	0.3198	0.7531	1.0000	0.8592	0.8765	0.8087	0.97s	2,300,162
DeiT	ViT-B/32 student	0.9671	0.1054	0.9136	1.0000	0.9548	0.9568	0.9314	1.02s	3,554,884
	ViT-S/32 student	0.9718	0.0793	0.9383	0.9870	0.9620	0.9653	0.9404	1.28s	3,554,884
	ViT-B/16 student	0.9718	0.0738	0.9259	1.0000	0.9615	0.9630	0.9411	1.54s	3,112,516
	ViT-S/16 student	0.9718	0.0964	0.9259	1.0000	0.9615	0.9630	0.9411	1.10s	3,112,516
	R26+S/32 student	0.9671	0.0660	0.9383	0.9744	0.9560	0.9616	0.9302	1.03s	3,554,884

Table 8: Results of the CNNs in experiment 4.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	AlexNet	0.9858	0.0354	0.9926	0.9781	0.9853	0.9861	0.9717	1.07s	46,760,706
	ResNet50	1.0000	0.0262	1.0000	1.0000	1.0000	1.0000	1.0000	3.70s	23,591,810
	VGG16	1.0000	0.0007	1.0000	1.0000	1.0000	1.0000	1.0000	5.59s	134,268,738
	InceptionV3	0.9965	0.0369	1.0000	0.9926	0.9963	0.9966	0.9929	3.87s	21,806,882
	ResNet50+MobileNetV2	1.0000	0.0179	1.0000	1.0000	1.0000	1.0000	1.0000	4.96s	25,852,354
	Hybrid model	0.9823	0.3591	1.0000	0.9643	0.9818	0.9830	0.9651	2.67s	1,930,966
Knowledge Distillation	AlexNet student	0.9894	0.0467	0.9926	0.9853	0.9889	0.9895	0.9787	2.50s	1,254,194
	ResNet50 student	0.9929	0.0754	1.0000	0.9854	0.9926	0.9932	0.9859	1.72s	1,254,194
	VGG16 student	0.9929	0.0862	1.0000	0.9854	0.9926	0.9932	0.9859	1.88s	1,254,194
	InceptionV3 student	0.9929	0.0422	1.0000	0.9854	0.9926	0.9932	0.9859	1.66s	1,254,194
	ResNet50+MobileNetV2 student	0.9929	0.0592	1.0000	0.9854	0.9926	0.9932	0.9859	1.80s	1,254,194

Table 9: Results of the ViTs in experiment 4.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	ViT-B/32	1.0000	0.0028	1.0000	1.0000	1.0000	1.0000	1.0000	4.52s	87,417,602
	ViT-S/32	0.9965	0.0051	1.0000	0.9926	0.9963	0.9966	0.9929	2.85s	22,475,138
	ViT-B/16	1.0000	0.0019	1.0000	1.0000	1.0000	1.0000	1.0000	7.72s	86,417,594
	ViT-S/16	1.0000	0.0018	1.0000	1.0000	1.0000	1.0000	1.0000	4.87s	21,590,402
	R26+S/32	0.9787	0.0480	1.0000	0.9574	0.9783	0.9796	0.9583	8.11s	36,028,098
Knowledge Distillation	ViT-B/32 student	0.9716	0.0697	0.9630	0.9774	0.9701	0.9713	0.9432	1.53s	2,300,162
	ViT-S/32 student	0.8794	0.6815	0.9926	0.8024	0.8874	0.8841	0.7808	1.35s	2,300,162
	ViT-B/16 student	0.9645	0.1487	0.9259	1.0000	0.9615	0.9630	0.9311	1.42s	3,442,274
	ViT-S/16 student	0.9787	0.0924	0.9556	1.0000	0.9773	0.9778	0.9582	1.33s	3,442,274
	R26+S/32 student	0.9645	0.0733	0.9333	0.9921	0.9618	0.9633	0.9303	1.56s	2,300,162
DeiT	ViT-B/32 student	0.9574	0.1114	0.9333	0.9767	0.9545	0.9565	0.9154	1.14s	3,554,884
	ViT-S/32 student	0.9468	0.1741	0.9037	0.9839	0.9421	0.9450	0.8958	1.10s	3,554,884
	ViT-B/16 student	0.9433	0.2008	0.8963	0.9837	0.9380	0.9413	0.8892	1.37s	3,112,516
	ViT-S/16 student	0.9397	0.1884	0.8963	0.9758	0.9344	0.9379	0.8815	1.31s	3,112,516
	R26+S/32 student	0.9397	0.1539	0.9111	0.9609	0.9354	0.9385	0.8801	1.11s	3,554,884

Table 10: Results of the CNNs in experiment 5.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	AlexNet	1.0000	0.0007	1.0000	1.0000	1.0000	1.0000	1.0000	0.74s	46,760,706
	ResNet50	1.0000	0.0051	1.0000	1.0000	1.0000	1.0000	1.0000	2.58s	23,591,810
	VGG16	1.0000	0.0002	1.0000	1.0000	1.0000	1.0000	1.0000	3.65s	134,268,738
	InceptionV3	0.9953	0.0638	1.0000	0.9878	0.9939	0.9962	0.9901	3.48s	21,806,882
	ResNet50+MobileNetV2	1.0000	0.0356	1.0000	1.0000	1.0000	1.0000	1.0000	3.58s	25,852,354
Knowledge Distillation	Hybrid model	0.4413	0.6797	1.0000	0.4050	0.5765	0.5492	0.1997	2.40s	1,930,966
	AlexNet student	1.0000	0.0044	1.0000	1.0000	1.0000	1.0000	1.0000	2.15s	1,254,194
	ResNet50 student	1.0000	0.0448	1.0000	1.0000	1.0000	1.0000	1.0000	1.69s	1,254,194
	VGG16 student	0.9906	0.1358	1.0000	0.9759	0.9878	0.9924	0.9804	1.67s	1,254,194
	InceptionV3 student	0.9953	0.0457	0.9877	1.0000	0.9938	0.9938	0.9901	1.68s	1,254,194
	ResNet50+MobileNetV2 student	1.0000	0.0489	1.0000	1.0000	1.0000	1.0000	1.0000	2.22s	1,254,194

Table 11: Results of the ViTs in experiment 5.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	ViT-B/32	1.0000	0.0024	1.0000	1.0000	1.0000	1.0000	1.0000	3.11s	87,417,602
	ViT-S/32	1.0000	0.0031	1.0000	1.0000	1.0000	1.0000	1.0000	2.76s	22,475,138
	ViT-B/16	1.0000	0.0012	1.0000	1.0000	1.0000	1.0000	1.0000	5.11s	86,417,594
	ViT-S/16	1.0000	0.0017	1.0000	1.0000	1.0000	1.0000	1.0000	3.26s	21,590,402
	R26+S/32	1.0000	0.0007	1.0000	1.0000	1.0000	1.0000	1.0000	5.86s	36,028,098
Knowledge Distillation	ViT-B/32 student	0.8779	0.4485	0.6790	1.0000	0.8088	0.8395	0.7532	2.66s	2,300,162
	ViT-S/32 student	0.9484	0.1526	0.8642	1.0000	0.9272	0.9321	0.8932	1.18s	2,300,162
	ViT-B/16 student	1.0000	0.0011	1.0000	1.0000	1.0000	1.0000	1.0000	1.76S	3,442,274
	ViT-S/16 student	1.0000	0.0035	1.0000	1.0000	1.0000	1.0000	1.0000	1.81s	3,442,274
	R26+S/32 student	0.8404	0.7652	0.5802	1.0000	0.7344	0.7901	0.6793	1.66s	2,300,162
DeiT	ViT-B/32 student	0.9484	0.1210	0.8765	0.9861	0.9281	0.9345	0.8918	0.97s	3,554,884
	ViT-S/32 student	0.9531	0.1131	0.8765	1.0000	0.9342	0.9383	0.9027	1.00s	3,554,884
	ViT-B/16 student	0.8779	0.3492	0.6790	1.0000	0.8088	0.8395	0.7532	1.53s	3,112,516
	ViT-S/16 student	0.9249	0.2402	0.8025	1.0000	0.8904	0.9012	0.8460	1.12s	3,112,516
	R26+S/32 student	0.9577	0.0981	0.9012	0.9865	0.9419	0.9468	0.9111	1.29s	3,554,884

Table 12: Results of the CNNs in experiment 6.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	AlexNet	0.9858	0.0323	0.9778	0.9925	0.9851	0.9855	0.9717	1.04s	46,760,706
	ResNet50	1.0000	0.0104	1.0000	1.0000	1.0000	1.0000	1.0000	2.63s	23,591,810
	VGG16	1.0000	0.0016	1.0000	1.0000	1.0000	1.0000	1.0000	5.58s	134,268,738
	InceptionV3	1.0000	0.0271	1.0000	1.0000	1.0000	1.0000	1.0000	3.68s	21,806,882
	ResNet50+MobileNetV2	0.9965	0.0216	1.0000	0.9926	0.9963	0.9966	0.9929	4.27s	25,852,354
Knowledge Distillation	Hybrid model	0.9787	0.3919	1.0000	0.9574	0.9783	0.9796	0.9583	2.72s	1,930,966
	AlexNet student	0.9645	0.1238	0.9926	0.9371	0.9640	0.9657	0.9306	2.04s	1,254,194
	ResNet50 student	0.9787	0.0611	1.0000	0.9574	0.9783	0.9796	0.9583	1.74s	1,254,194
	VGG16 student	0.9858	0.3479	1.0000	0.9712	0.9854	0.9864	0.9720	3.20s	1,254,194
	InceptionV3 student	0.9752	0.0860	1.0000	0.9507	0.9747	0.9762	0.9515	1.76s	1,254,194
	ResNet50+MobileNetV2 student	0.9907	0.2484	1.0000	0.8282	0.9060	0.9048	0.8188	2.00s	1,254,194

Table 13: Results of the ViTs in experiment 6.

	Models	Acc	Loss	Recall	Precision	F1-Score	AUC	MCC	Inference time	Total of params
Main architectures	ViT-B/32	1.0000	0.0025	1.0000	1.0000	1.0000	1.0000	1.0000	3.46s	87,417,602
	ViT-S/32	1.0000	0.0048	1.0000	1.0000	1.0000	1.0000	1.0000	4.00s	22,475,138
	ViT-B/16	1.0000	0.0006	1.0000	1.0000	1.0000	1.0000	1.0000	7.26s	86,417,594
	ViT-S/16	1.0000	0.0013	1.0000	1.0000	1.0000	1.0000	1.0000	4.80s	21,590,402
	R26+S/32	0.9965	0.0195	1.0000	0.9926	0.9963	0.9966	0.9929	6.74s	36,028,098
Knowledge Distillation	ViT-B/32 student	0.9681	0.1090	0.93333	1.0000	0.9655	0.9667	0.9378	1.45s	2,300,162
	ViT-S/32 student	0.9787	0.0897	0.9926	0.9640	0.9781	0.9793	0.9578	1.44s	2,300,162
	ViT-B/16 student	0.9716	0.0937	0.9778	0.9635	0.9706	0.9719	0.9433	1.94s	3,442,274
	ViT-S/16 student	0.9929	0.0215	0.9926	0.9926	0.9926	0.9929	0.9858	1.44s	3,442,274
	R26+S/32 student	0.9823	0.0506	0.9630	1.0000	0.9811	0.9815	0.9650	2.30s	2,300,162
DeiT	ViT-B/32 student	0.9681	0.1108	0.9481	0.9846	0.9660	0.9673	0.9366	1.46s	3,554,884
	ViT-S/32 student	0.9397	0.1967	0.8889	0.9836	0.9339	0.9376	0.8825	1.46s	3,554,884
	ViT-B/16 student	0.9255	0.2187	0.8741	0.9672	0.9183	0.9234	0.8539	1.68s	3,112,516
	ViT-S/16 student	0.9007	0.3263	0.8222	0.9652	0.8880	0.8975	0.8081	1.20s	3,112,516
	R26+S/32 student	0.9327	0.1409	0.9037	0.9683	0.9349	0.9382	0.8807	1.49s	3,554,884

4 Discussion

This study serves as a reference for hospitals and gastroenterology doctors to automate the process of diagnosing patients with inflammatory bowel disease. The AlexNet network in experiment 1 was the algorithm that achieved the shortest inference time (0.75 seconds), with a good performance: 100% Accuracy, 100% Recall, 100% Precision, 100% F1-Score, 100% AUC, 100% F1-Score and 0.0031 Loss. However, due to the self-attention mechanism and in the performance improvement that ViTs demonstrate over CNNs [22], ViT-S/16 with knowledge distillation could be a very promising alternative with: 99.06% accuracy, 97.53% recall, 100% precision, 98.75% f1-Score, 98.77% AUC, 98.02% MCC, 0.0147 Loss and took 0.91 seconds to infer the test results.

The models used in this study prove to be accurate in the diagnosis of inflammatory bowel disease, and through the analysis of the AUC and MCC metrics, the models show a good capacity in distinguish the classes, and the number of false positives and false negatives is very close to zero or even manages to be zero, respectively. The results obtained are superior to those presented by the authors [35] who used a ResNetXt-101 network. Furthermore, if we compare with the results presented in the same study corresponding to the accuracy of clinicians in diagnosing the disease, the accuracy of the models used manages to be much higher. However, the number of images used by the authors was higher than the one used in this study, as well as the images were collected using the same endoscopic tool.

The authors [30] in their study proposed a hybrid model, where they combine a CNN with an LSTM to perform the confirmation of mucosal healing in Crohn's disease. The best result achieved by the authors was 95.3% accuracy, 92.78% sensitivity, 94.6% specificity and an AUCROC of 0.98. In this paper, a hybrid model is also presented, where two more LSTMs were used concerning what was proposed by the authors, among other modifications. These changes represented a significant improvement in the results, with the best results being 100% accuracy, 100% recall, 100% precision and 100% AUC.

The literature has focused on the implementation of deep learning models in image processing of Crohn's disease, or ulcerative colitis disease. In this sense, the authors in the works [18, 31] developed a Recurrent attention neural network in the detection of the type of lesion caused by inflammation in the mucosa of Crohn's disease. Although the study objective is different, the teacher and student models implemented in this work describe a performance improvement.

The implementation of knowledge distillation is one of the important contributions of this work so that it is possible to use these architectures in the diagnosis of IBD patients. Comparing the results between the different techniques used, applied to CNN networks and ViTs. It is concluded that knowledge distillation in ViTs allowed a greater reduction in inference time, preserving a good performance of the models. On the other hand, in ViTs it is still verified that in experiments 3 and 5 the hybrid model R26+S/32 obtained a performance improvement, using the distillation technique through the attention (DeiT) when compared with knowledge distillation. In experiment four the same situation is

verified in the ViT-S/32 model. The implementation of DeiT in this work represents a performance improvement when compared with the implementation in another study [12].

5 Conclusion

It is concluded with this work that ViT models and CNN networks show a good performance in the recognition of inflammatory bowel disease. It is considered that the knowledge distillation through attention in the ViTs models allowed in some experiments performed to obtain a lower inference time concerning the knowledge distillation through the logits. Furthermore, in some experiments, such as the R26+S/32 model in experiment 3, it was found that there was an improvement in the performance of the student model in the knowledge distillation through attention [28].

However, in this study, it is recognized as a limitation in the number of images used to train the ViTs. Because these deep-learning models need a large number of images to be able to learn the information well [10]. Another limitation of this work is that the images were not collected based on the same endoscopic tool, which consequently the images are different and in the future may not demonstrate the same performance in processing colonoscopy images regarding Crohn's disease.

In future research, the interpretability of the deep learning models should be performed to understand which parts of the mucosa the models are using to predict inflammatory bowel disease [21]. This will ensure that the performance described by the algorithms corresponds to a correct prediction of the disease based on the inflammation caused. Another line of research lies in leveraging ordinal techniques [4, 19] to predict the degree of severity according to the Mayo score of ulcerative colitis. Furthermore, it is important in the future to develop efforts to explore and implement deep learning models that process video from endoscopic examinations in real time and recognise the area affected by inflammation, predicting the type of inflammatory bowel disease.

References

1. Amorim, J.P., Domingues, I., Abreu, P.H., Santos, J.A.: Interpreting deep learning models for ordinal problems. In: European Symposium on Artificial Neural Networks (ESANN). pp. 373–378 (2018)
2. Borgli, H., Riegler, M., Thambawita, V., Jha, D., Hicks, S., Halvorsen, P.: The HyperKvasir Dataset (2019), publisher: OSF
3. Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., Johansen, D., Griwodz, C., Stensland, H.K., Garcia-Ceja, E., Schmidt, P.T., Hammer, H.L., Riegler, M.A., Halvorsen, P., Lange, T.: HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data* **7**, 283 (2020)

4. Cardoso, J.S., Sousa, R., Domingues, I.: Ordinal data classification using kernel discriminant analysis: A comparison of three approaches. In: 11th International Conference on Machine Learning and Applications. vol. 1, pp. 473–477 (2012)
5. Chicco, D., Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020)
6. Chierici, M., Puica, N., Pozzi, M., Capistrano, A., Donzella, M.D., Colangelo, A., Osmani, V., Jurman, G.: Automatically detecting Crohn’s disease and Ulcerative Colitis from endoscopic imaging. *BMC Medical Informatics and Decision Making* **22**, 300 (2022)
7. CrohnIPI, <https://crohnipi.ls2n.fr/en/crohn-ipi-project/> (accessed Feb. 21, 2023)
8. Doença inflamatória do intestino | CUF, <https://www.cuf.pt/saude-a-z/doenca-inflamatoria-do-intestino> (accessed Nov. 02, 2022)
9. Das, D., Massa, H., Kulkarni, A., Rekatsinas, T.: An Empirical Analysis of the Impact of Data Augmentation on Knowledge Distillation. *Arxiv* (2020), publisher: arXiv Version: 2
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations* (2020)
11. Flynn, S., Eisenstein, S.: Inflammatory Bowel Disease Presentation and Diagnosis. *Surgical Clinics of North America* **99**(6), 1051–1062 (Dec 2019)
12. Galdran, A., Carneiro, G., Ballester, M.A.G.: Convolutional Nets Versus Vision Transformers for Diabetic Foot Ulcer Classification. *Diabetic Foot Ulcers Grand Challenge* (2022)
13. Gamage, C., Wijesinghe, I., Chitraranjan, C., Perera, I.: GI-Net: Anomalies Classification in Gastrointestinal Tract through Endoscopic Imagery with Deep Learning. In: *Moratuwa Engineering Research Conference (MERCon)*. pp. 66–71. IEEE, Moratuwa, Sri Lanka (Jul 2019)
14. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge Distillation: A Survey. *International Journal of Computer Vision* **129**(6), 1789–1819 (Jun 2021)
15. H. Kassani, S., Hosseinzadeh Kassani, P., Wesolowski, M., Schneider, K., Deters, R.: Classification of histopathological biopsy images using ensemble of deep learning networks. In: *arXiv preprint* (11 2019)
16. Khan, M.N., Hasan, M.A., Anwar, S.: Improving the Robustness of Object Detection Through a Multi-Camera-Based Fusion Algorithm Using Fuzzy Logic. *Frontiers in Artificial Intelligence* **4**, 638951 (2021)
17. Li, W., Shao, S., Liu, W., Qiu, Z., Zhu, Z., Huan, W.: What Role Does Data Augmentation Play in Knowledge Distillation? In: *Computer Vision – ACCV 2022*, vol. 13842, pp. 507–525. Springer Nature Switzerland (2023), series Title: Lecture Notes in Computer Science
18. Maissin, A., Vallée, R., Flamant, M., Fondain-Bossiere, M., Berre, C.L., Coutrot, A., Normand, N., Mouchère, H., Coudol, S., Trang, C., Bourreille, A.: Multi-expert annotation of Crohn’s disease images of the small bowel for automatic detection using a convolutional recurrent attention neural network. *Endoscopy Int Open* **09**, E1136–E1144 (2021)
19. Marques, F., Duarte, H., Santos, J.a., Domingues, I., Amorim, J.P., Abreu, P.H.: An iterative oversampling approach for ordinal classification. In: *34th ACM/SIGAPP Symposium on Applied Computing*. p. 771–774 (2019)

20. Maurício, J., Domingues, I.: Deep Neural Networks to distinguish between Crohn's disease and Ulcerative colitis. In: 11th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA) (2023)
21. Maurício, J., , Domingues, I.: Interpretability of deep neural networks to diagnose inflammatory bowel disease. In: 29th edition of the Portuguese Conference on Pattern Recognition (2023 (to appear))
22. Maurício, J., Domingues, I., Bernardino, J.: Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences* **13**(9) (2023)
23. Polat, G., Kani, H.T., Ergenc, I., Alahdab, Y.O., Temizel, A., Atug, O.: Labeled Images for Ulcerative Colitis (LIMUC) Dataset (2022)
24. Sairenji, T., Collins, K.L., Evans, D.V.: An Update on Inflammatory Bowel Disease. *Primary Care: Clinics in Office Practice* **44**, 673–692 (2017)
25. Seyedian, S.S., Nokhostin, F., Malamir, M.D.: A review of the diagnosis, prevention, and treatment methods of inflammatory bowel disease. *J of Medicine and Life* **12**, 113–122 (2019)
26. Shahzadi, I., Tang, T.B., Meriadeau, F., Quyyum, A.: CNN-LSTM: Cascaded Framework For Brain Tumour Classification. In: IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES). pp. 633–637 (2018)
27. Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., Beyer, L.: How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. arXiv Version: 2 (2021)
28. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. Arxiv (2020), publisher: arXiv Version: 2
29. Turan, M., Durmus, F.: UC-NfNet: Deep learning-enabled assessment of ulcerative colitis from colonoscopy images. *Medical Image Analysis* **82**, 102587 (2022)
30. Udristoiu, A.L., Stefanescu, D., Gruionu, G., Gruionu, L.G., Iacob, A.V., Karstensen, J.G., Vilman, P., Saftoiu, A.: Deep Learning Algorithm for the Confirmation of Mucosal Healing in Crohn's Disease, Based on Confocal Laser Endomicroscopy Images. *J of Gastrointestinal and Liver Diseases* **30**, 59–65 (2021)
31. Vallée, R., Coutrot, A., Normand, N., Mouchère, H.: Accurate small bowel lesions detection in wireless capsule endoscopy images using deep recurrent attention neural network. In: IEEE 21st Int WS on Multimedia Signal Proc (MMSP) (2019)
32. Vallée, R., Coutrot, A., Normand, N., Mouchère, H.: Influence of expertise on human and machine visual attention in a medical image classification task. In: European Conference on Visual Perception (2021)
33. Vallée, R., Maissin, A., Coutrot, A., Mouchère, H., Bourreille, A., Normand, N.: CrohnIPI: An endoscopic image database for the evaluation of automatic Crohn's disease lesions recognition algorithms. In: Medical Imaging: Biomedical Applications in Molecular, Structural, and Functional Imaging. p. 61. SPIE (2020)
34. Vankdothu, R., Hameed, M.A., Fatima, H.: A Brain Tumor Identification and Classification Using Deep Learning based on CNN-LSTM Method. *Computers and Electrical Engineering* **101**, 107960 (2022)
35. Wang, L., Chen, L., Wang, X., Liu, K., Li, T., Yu, Y., Han, J., Xing, S., Xu, J., Tian, D., Seidler, U., Xiao, F.: Development of a Convolutional Neural Network-Based Colonoscopy Image Assessment Model for Differentiating Crohn's Disease and Ulcerative Colitis. *Frontiers in Medicine* **9**, 789862 (Apr 2022)

Appendix B - Supplementary material

This chapter serves as supplementary material, where it presents the students' interpretability images of the implemented deep learning architectures. With this, it is intended to demonstrate on which image characteristics the students relied to achieve the results demonstrated in chapter 5.

Figures [8.1 - 8.9] present the results of the interpretability of the CNNs students in all experiments. Figures [8.10 - 8.18] present the result of the interpretability of the ViTs students from experiment 1 to experiment 9, respectively. Figures [8.19 - 8.27] present the result of the interpretability of the ViTs students, whose knowledge distillation was made through the attention.

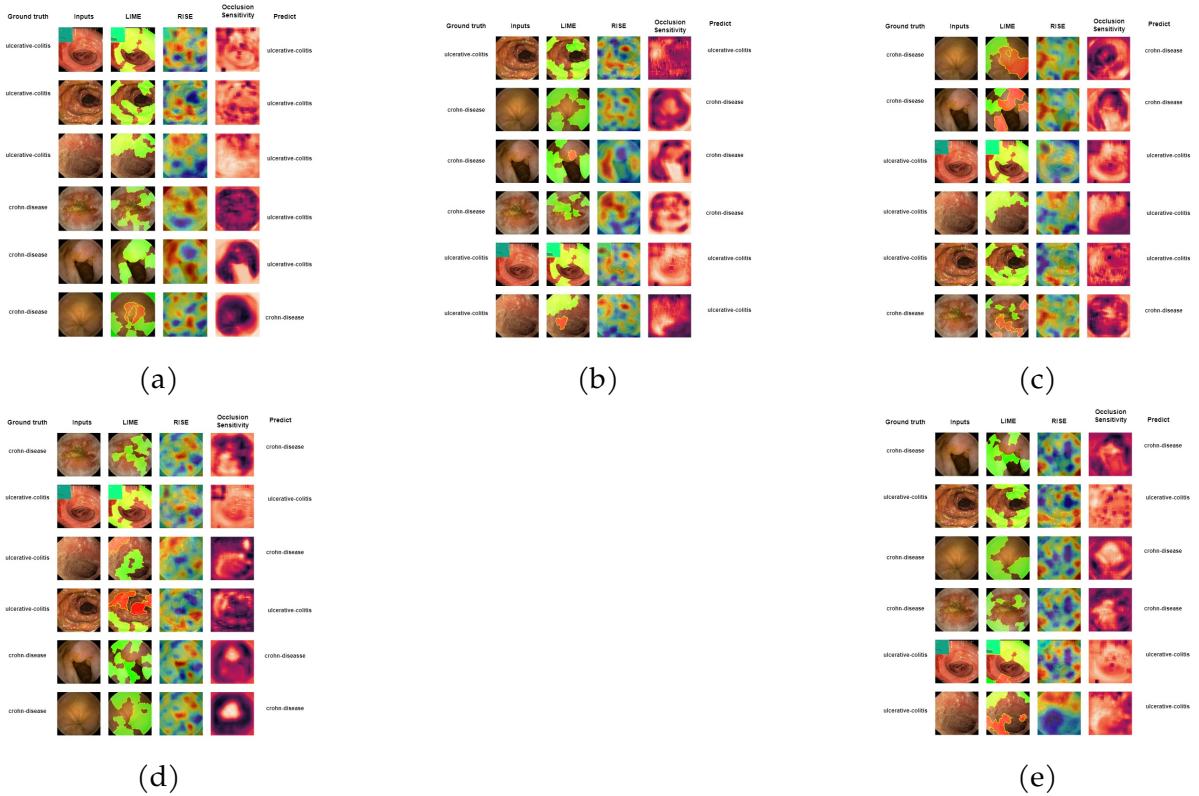


Figure 8.1: Interpretability of the CNNs student models in experiment 1: (a) AlexNet; (b) ResNet50; (c) VGG16; (d) InceptionV3; (e) Ensemble method.

Diagnose of the IBD using Deep Learning Models with Interpretability

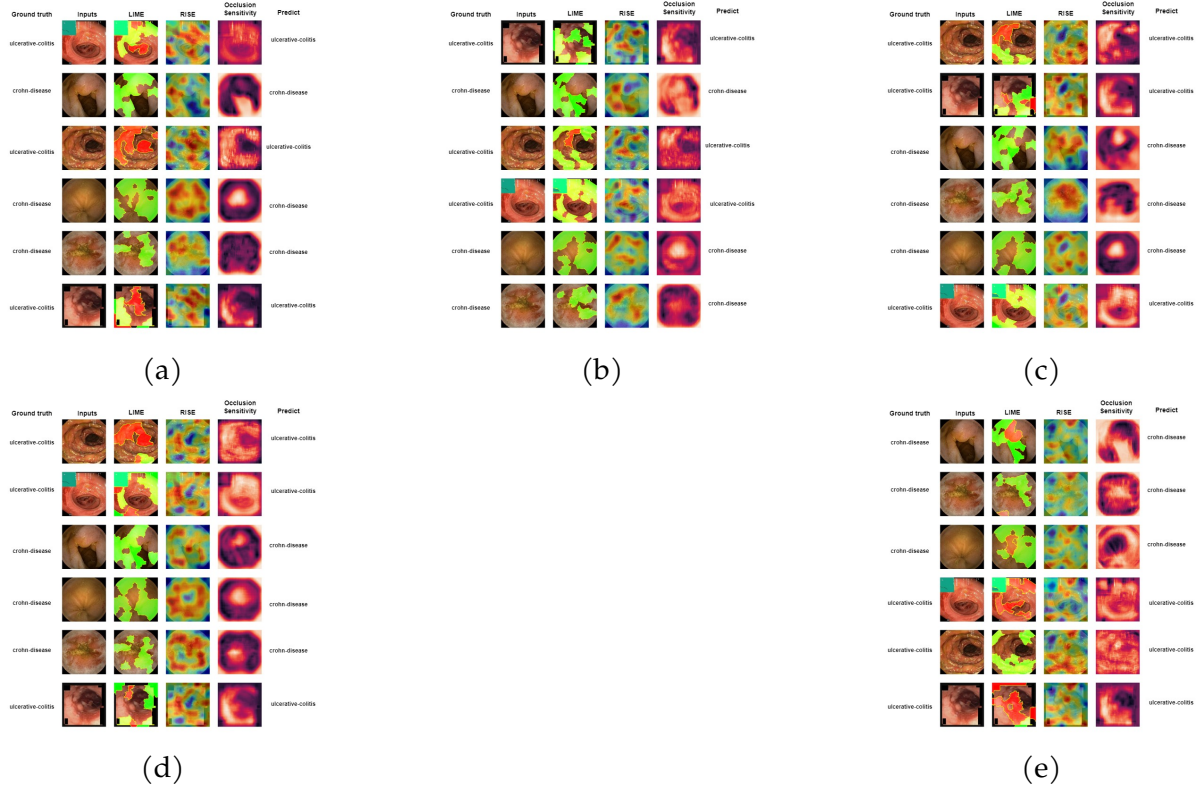


Figure 8.2: Interpretability of the CNNs student models in experiment 2: (a) AlexNet; (b) ResNet50; (c) VGG16; (d) InceptionV3; (e) Ensemble method.

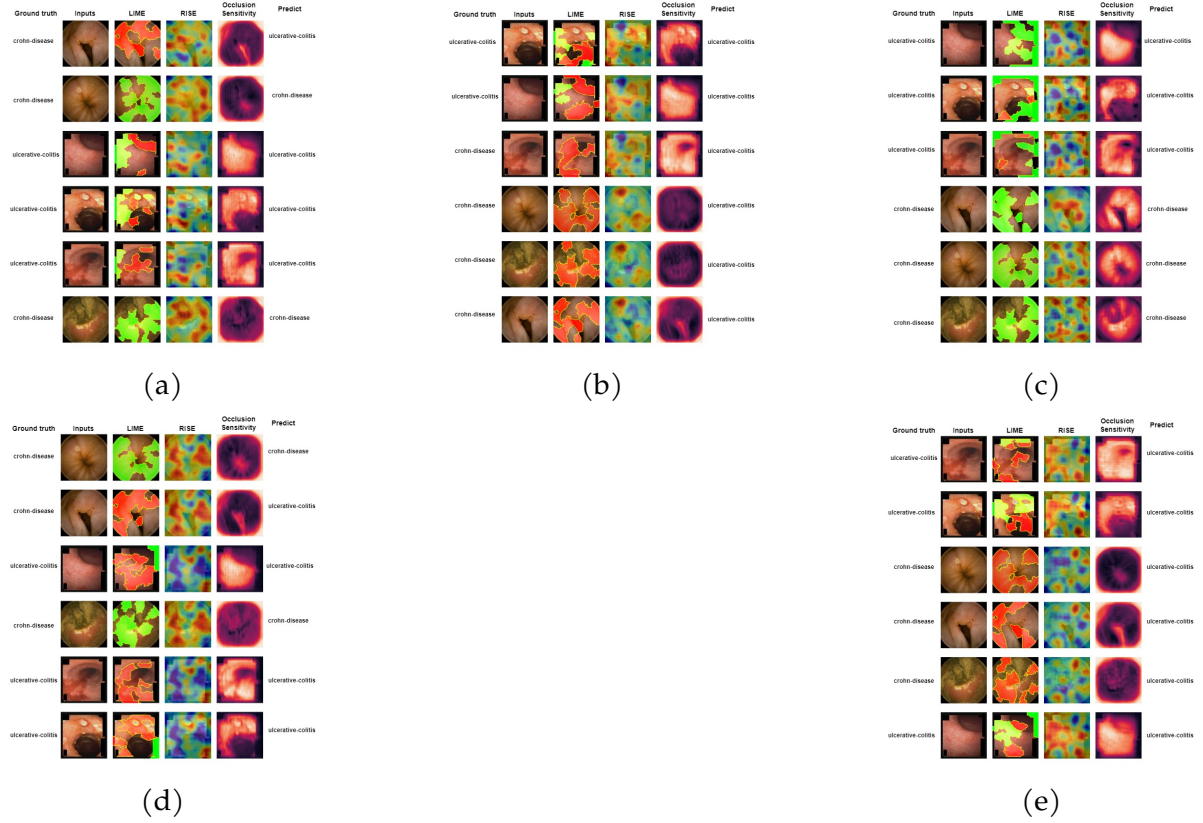


Figure 8.3: Interpretability of the CNNs student models in experiment 3: (a) AlexNet; (b) ResNet50; (c) VGG16; (d) InceptionV3; (e) Ensemble method.

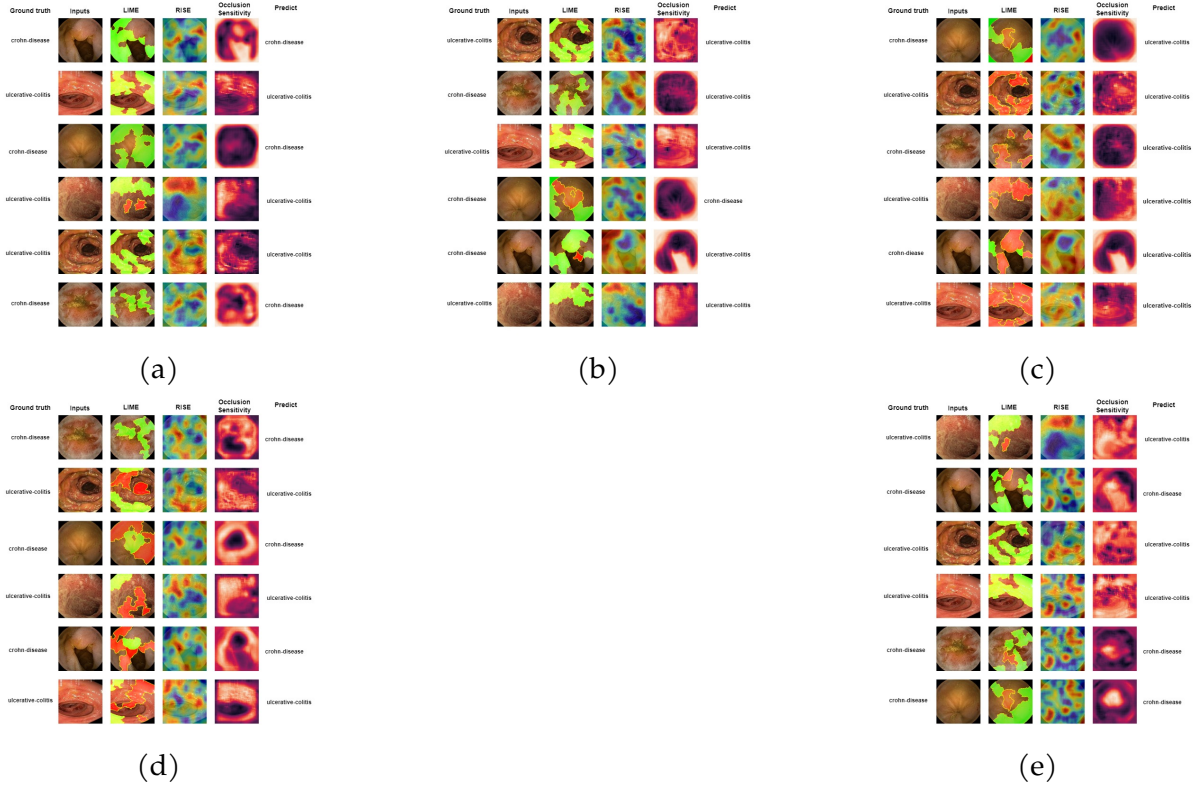


Figure 8.4: Interpretability of the CNNs student models in experiment 4: (a) AlexNet; (b) ResNet50; (c) VGG16; (d) InceptionV3; (e) Ensemble method.

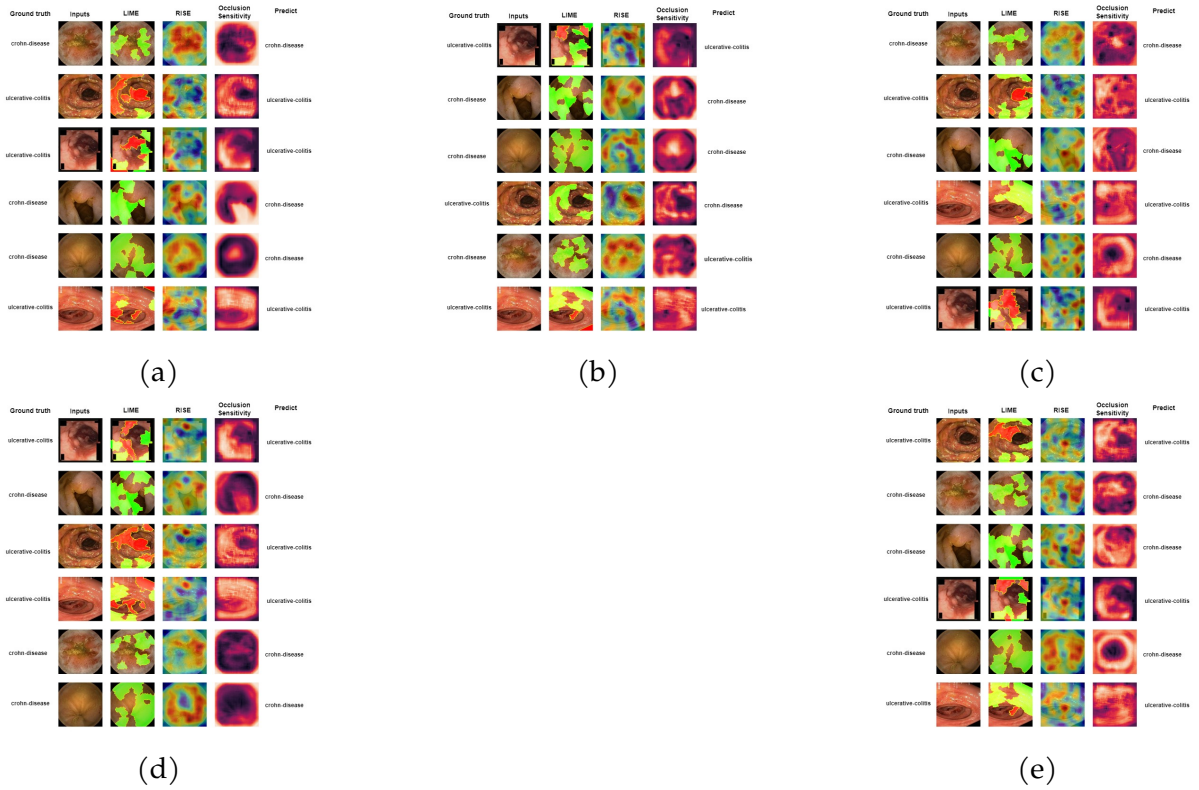


Figure 8.5: Interpretability of the CNNs student models in experiment 5: (a) AlexNet; (b) ResNet50; (c) VGG16; (d) InceptionV3; (e) Ensemble method.

Diagnose of the IBD using Deep Learning Models with Interpretability

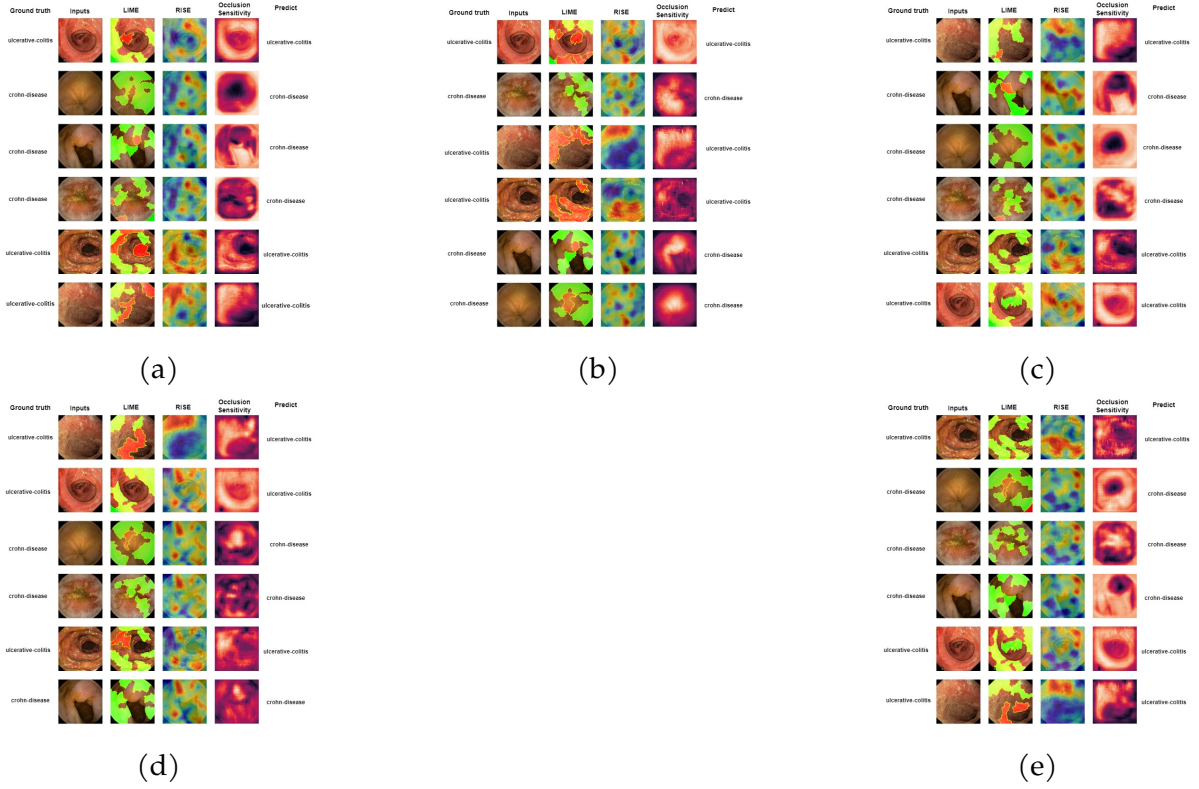


Figure 8.6: Interpretability of the CNNs student models in experiment 6: (a) AlexNet; (b) ResNet50; (c) VGG16; (d) InceptionV3; (e) Ensemble method.

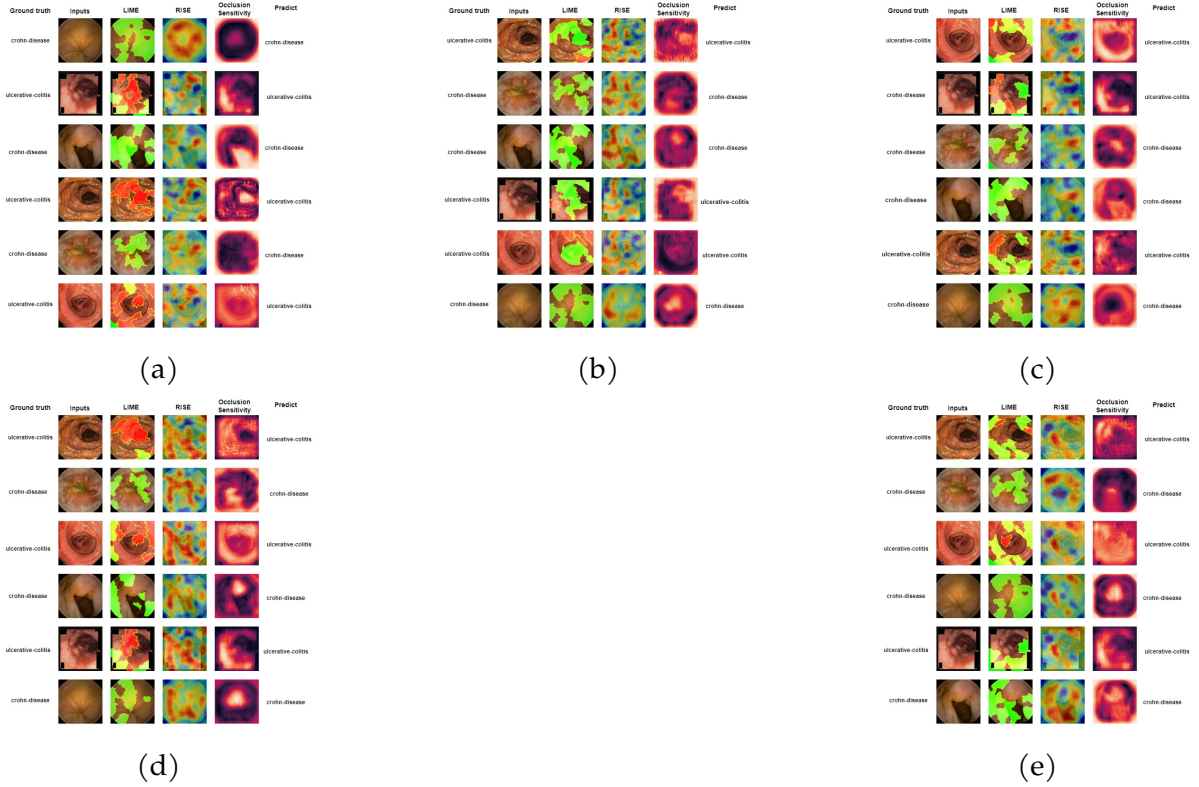


Figure 8.7: Interpretability of the CNNs student models in experiment 7: (a) AlexNet; (b) ResNet50; (c) VGG16; (d) InceptionV3; (e) Ensemble method.

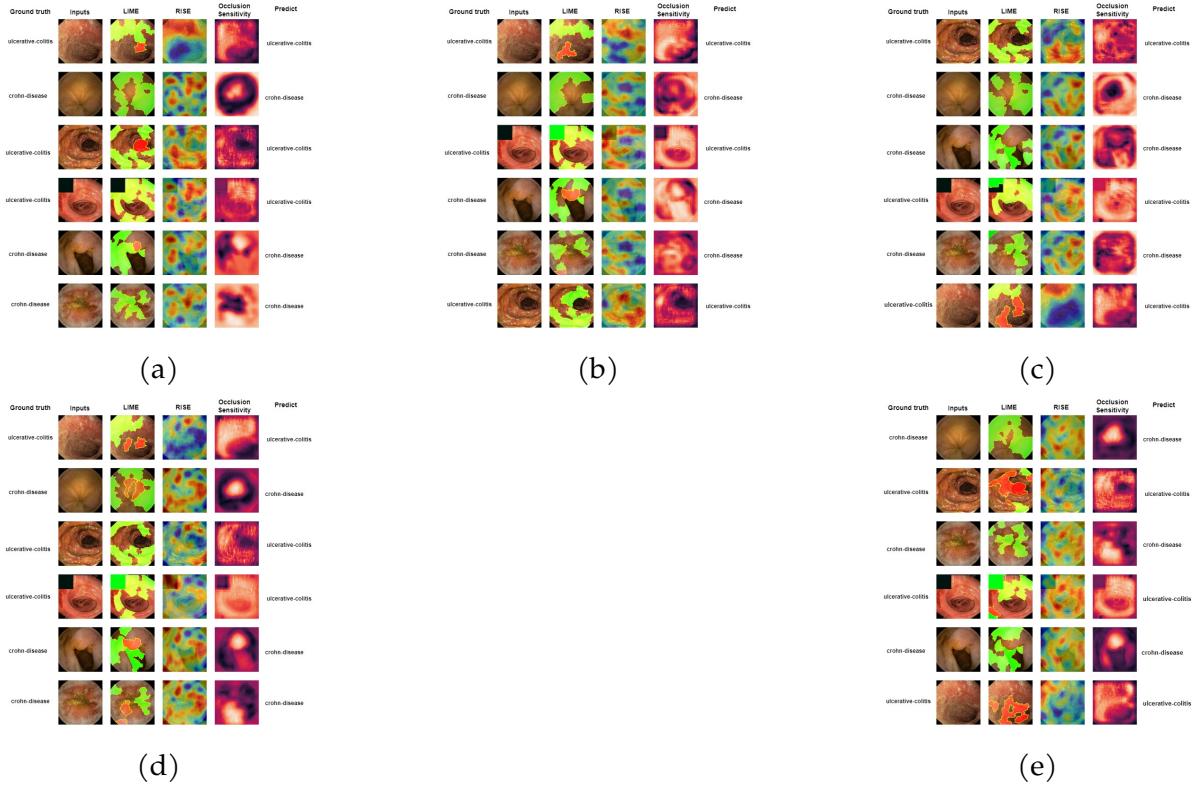


Figure 8.8: Interpretability of the CNNs student models in experiment 8: (a) AlexNet; (b) ResNet50; (c) VGG16; (d) InceptionV3; (e) Ensemble method.

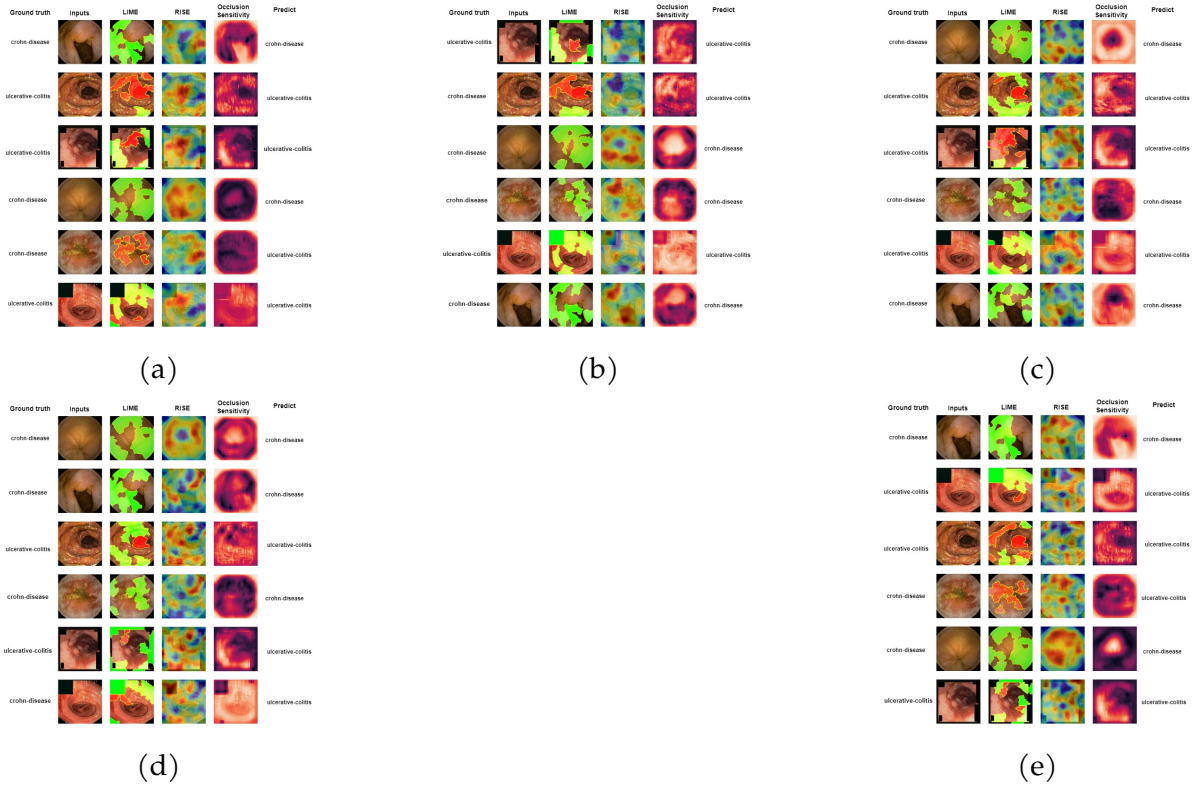


Figure 8.9: Interpretability of the CNNs student models in experiment 9: (a) AlexNet; (b) ResNet50; (c) VGG16; (d) InceptionV3; (e) Ensemble method.

Diagnose of the IBD using Deep Learning Models with Interpretability

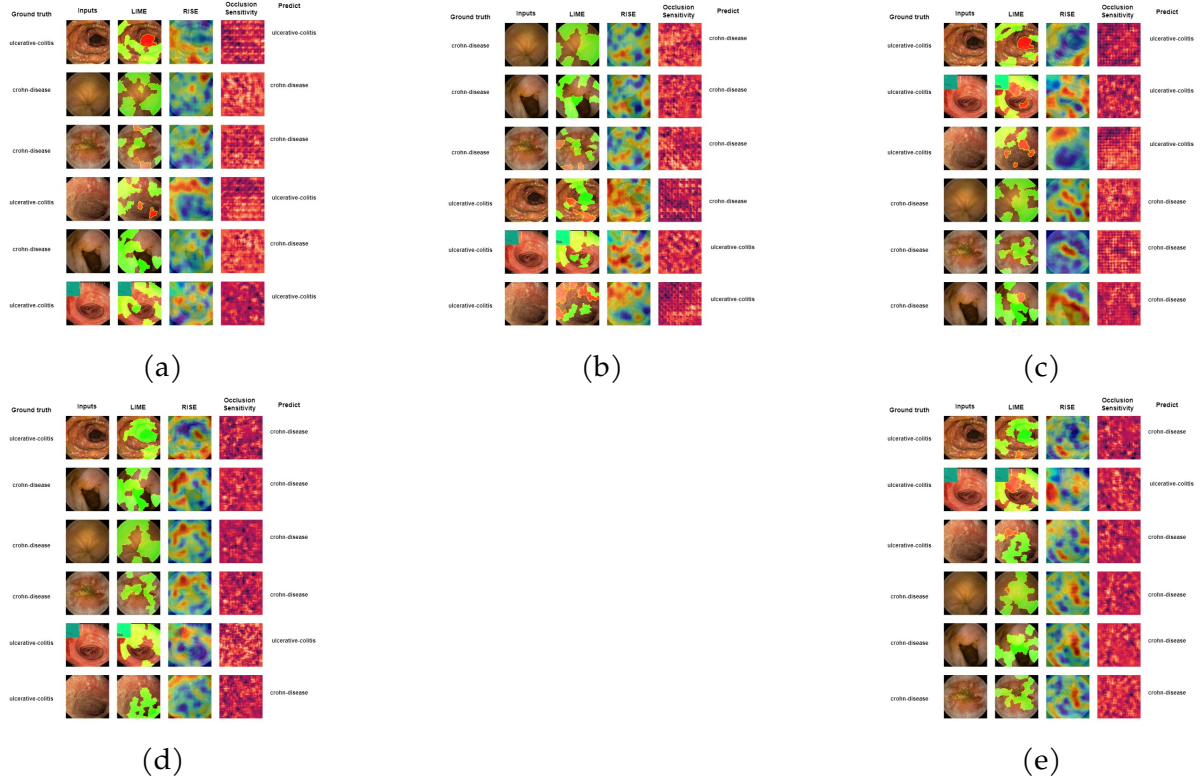


Figure 8.10: Interpretability of the ViTs student models in experiment 1: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.

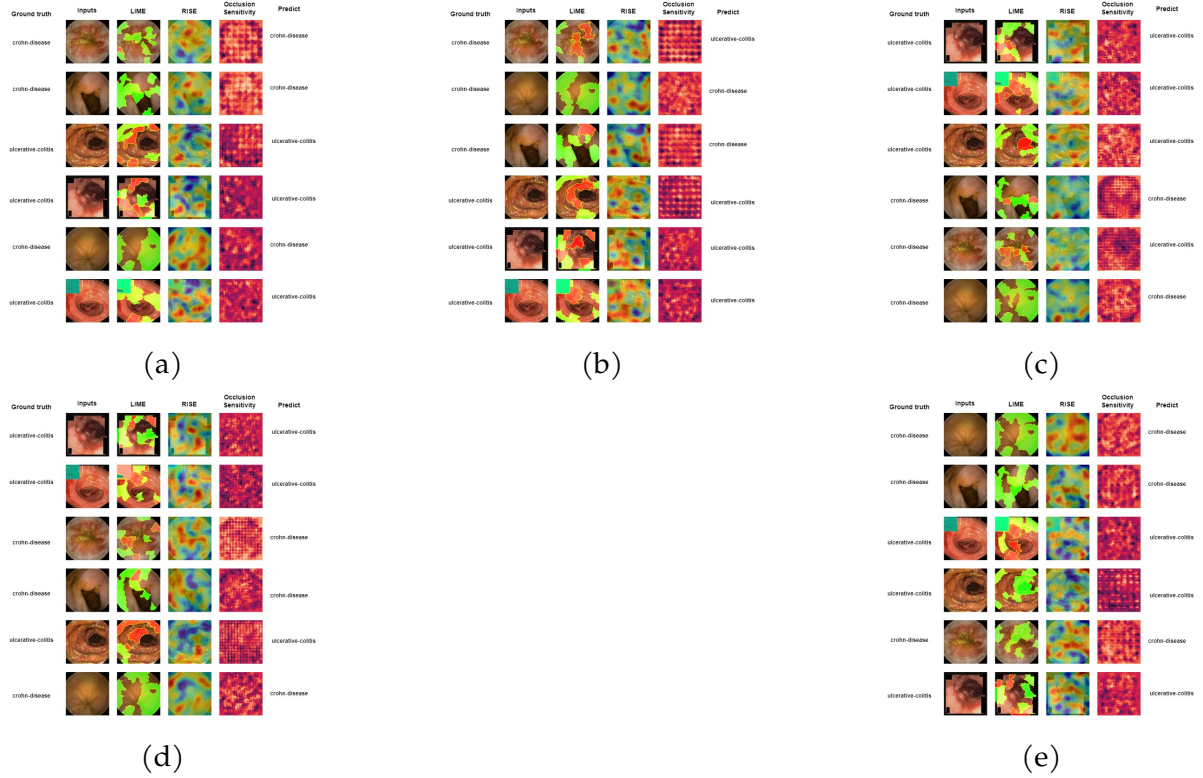


Figure 8.11: Interpretability of the ViTs student models in experiment 2: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.

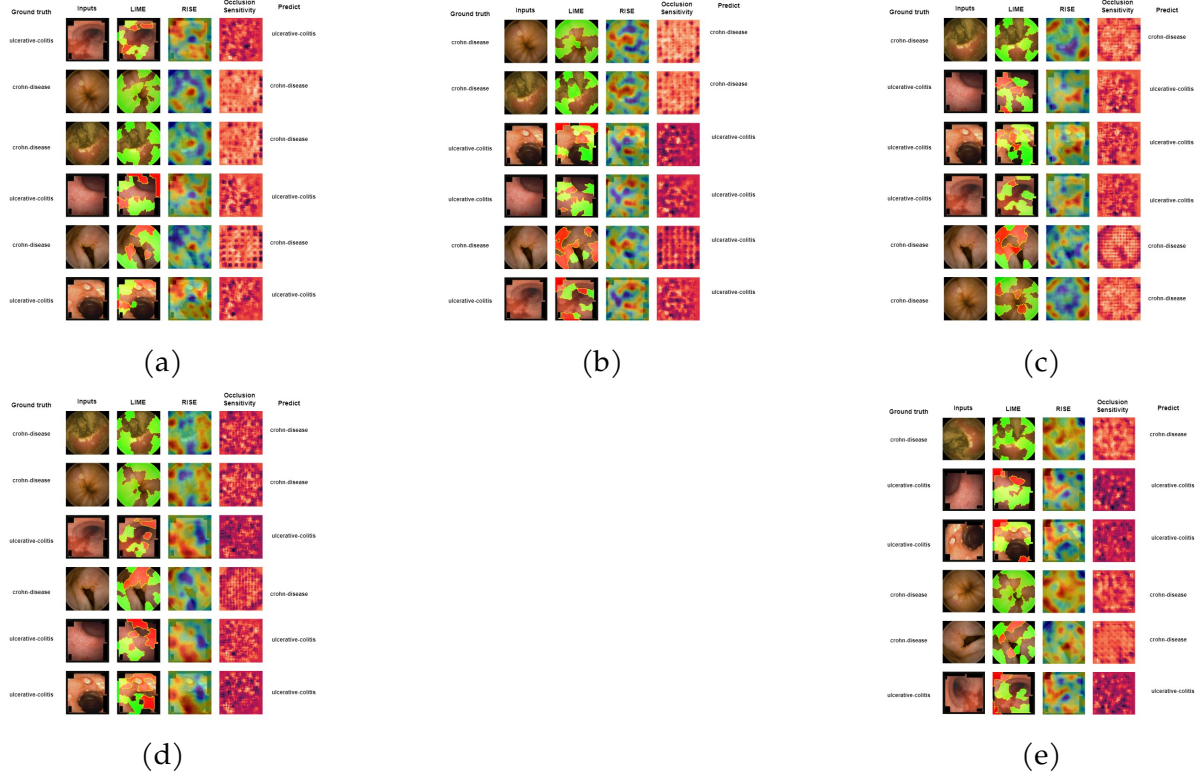


Figure 8.12: Interpretability of the ViTs student models in experiment 3: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.

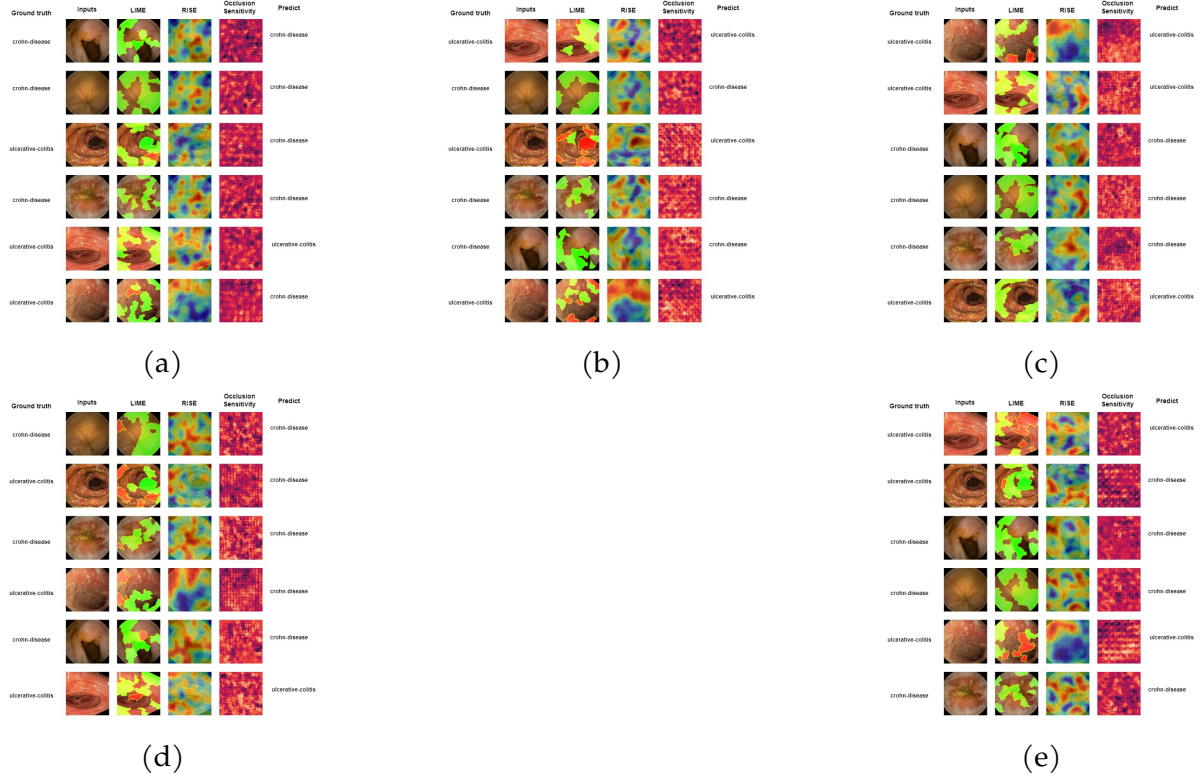


Figure 8.13: Interpretability of the ViTs student models in experiment 4: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.

Diagnose of the IBD using Deep Learning Models with Interpretability

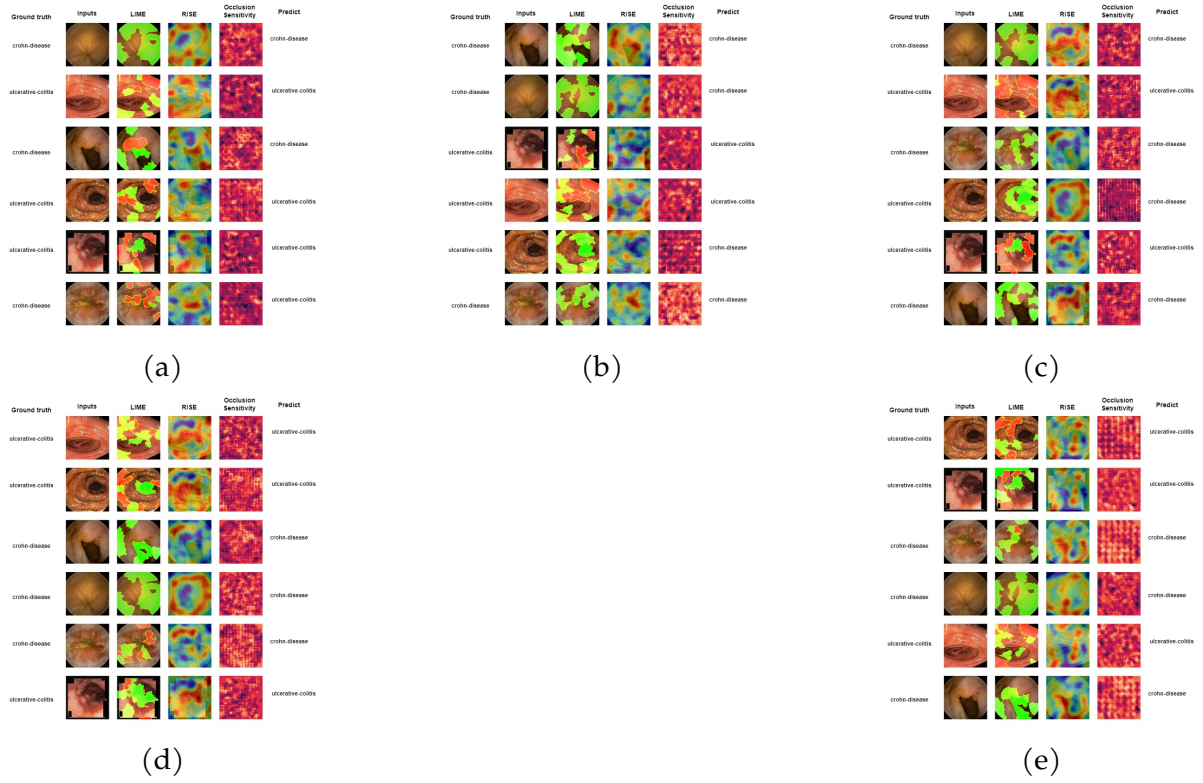


Figure 8.14: Interpretability of the ViTs student models in experiment 5: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.

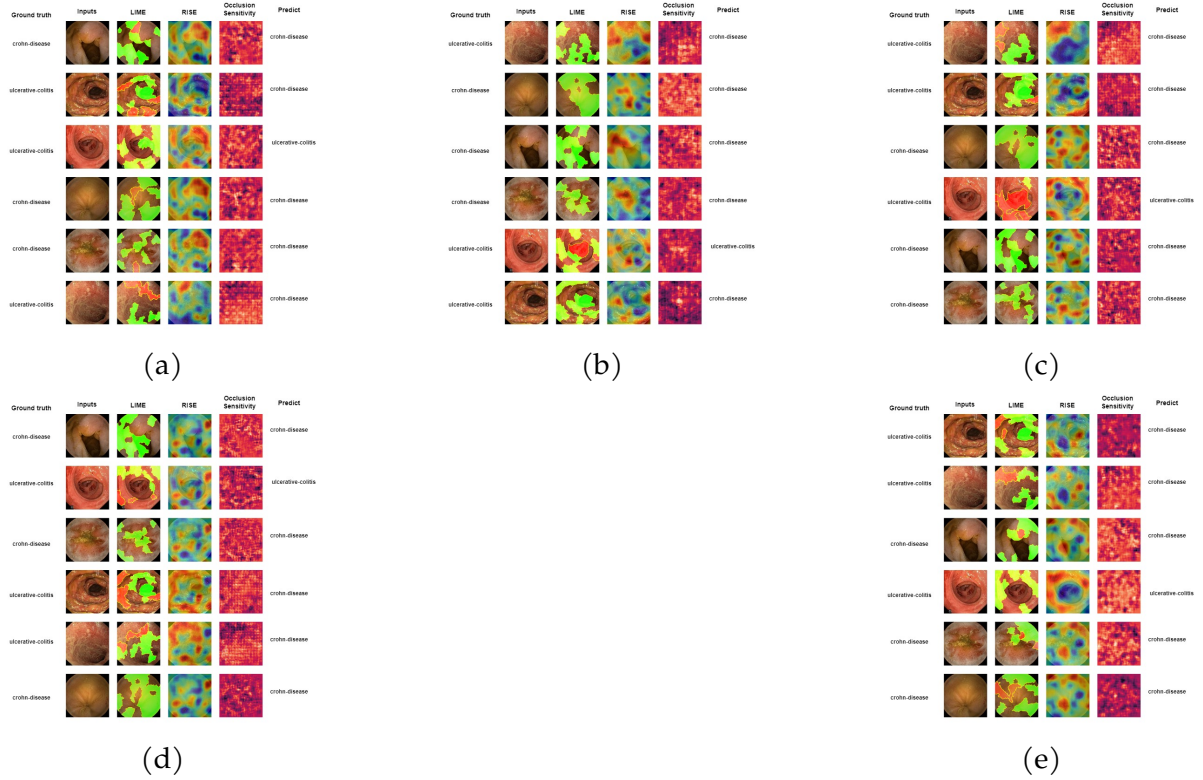


Figure 8.15: Interpretability of the ViTs student models in experiment 6: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.

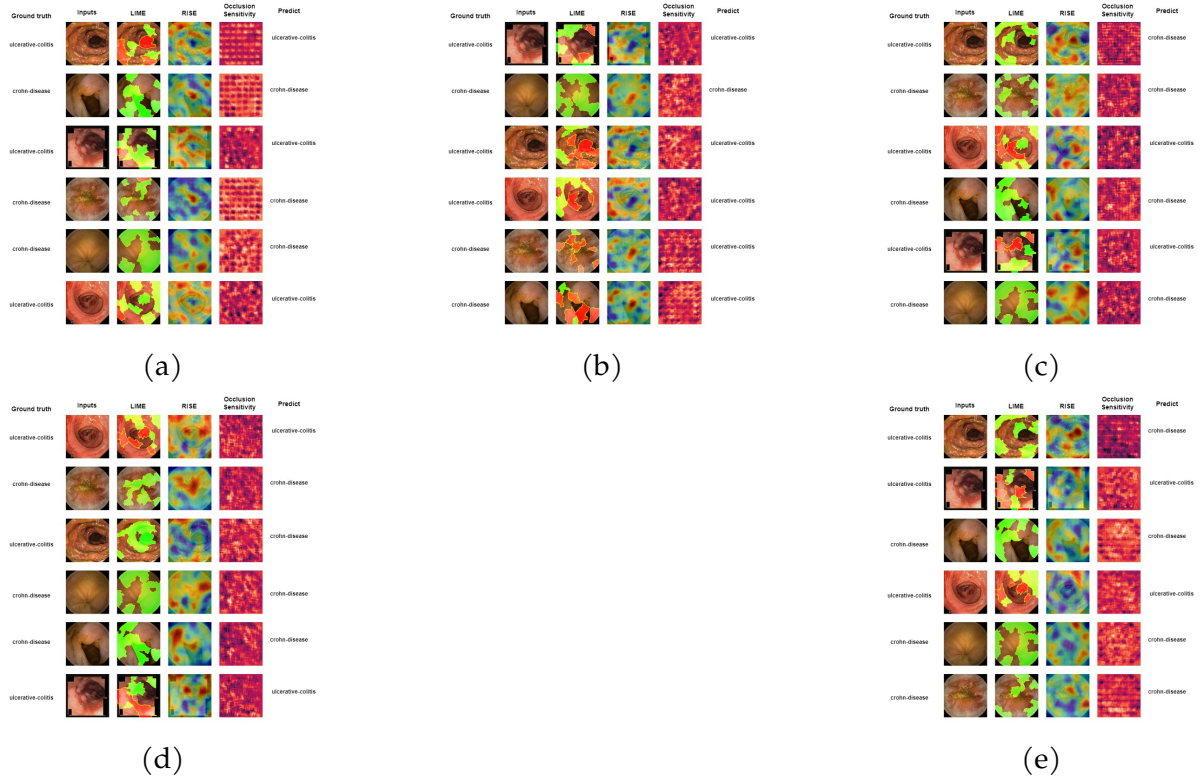


Figure 8.16: Interpretability of the ViTs student models in experiment 7: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.

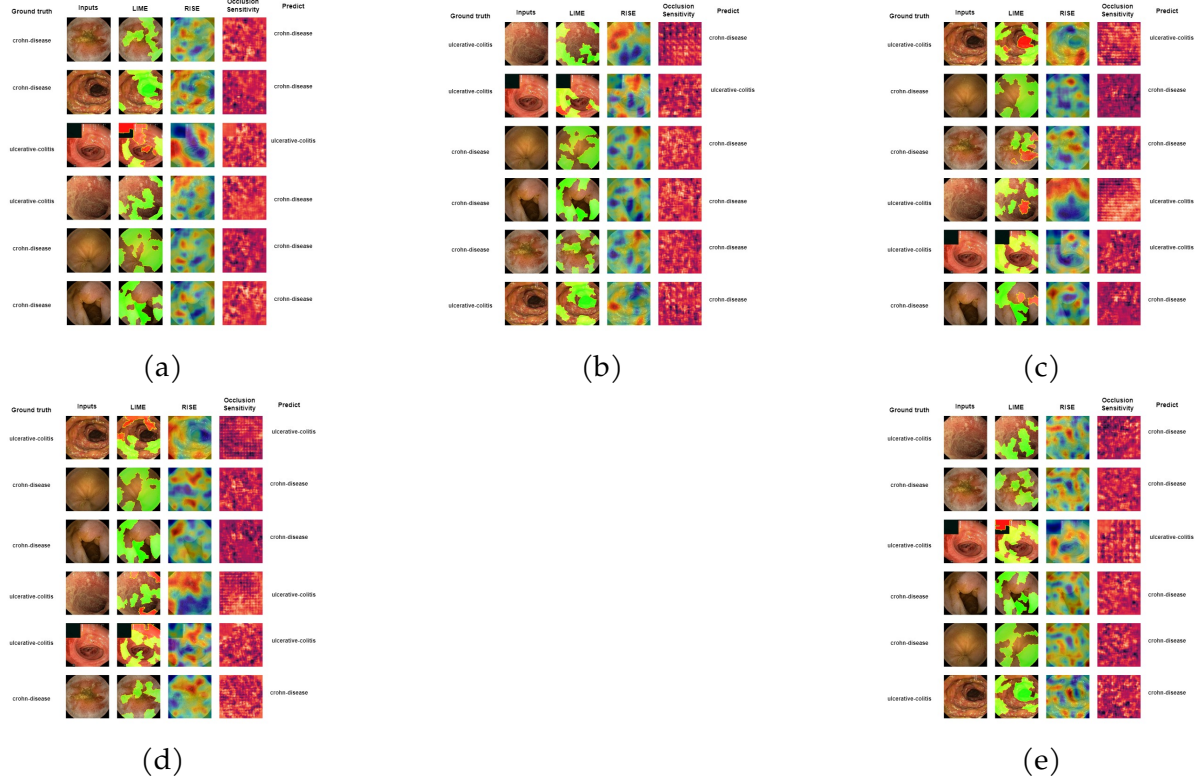


Figure 8.17: Interpretability of the ViTs student models in experiment 8: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.

Diagnose of the IBD using Deep Learning Models with Interpretability

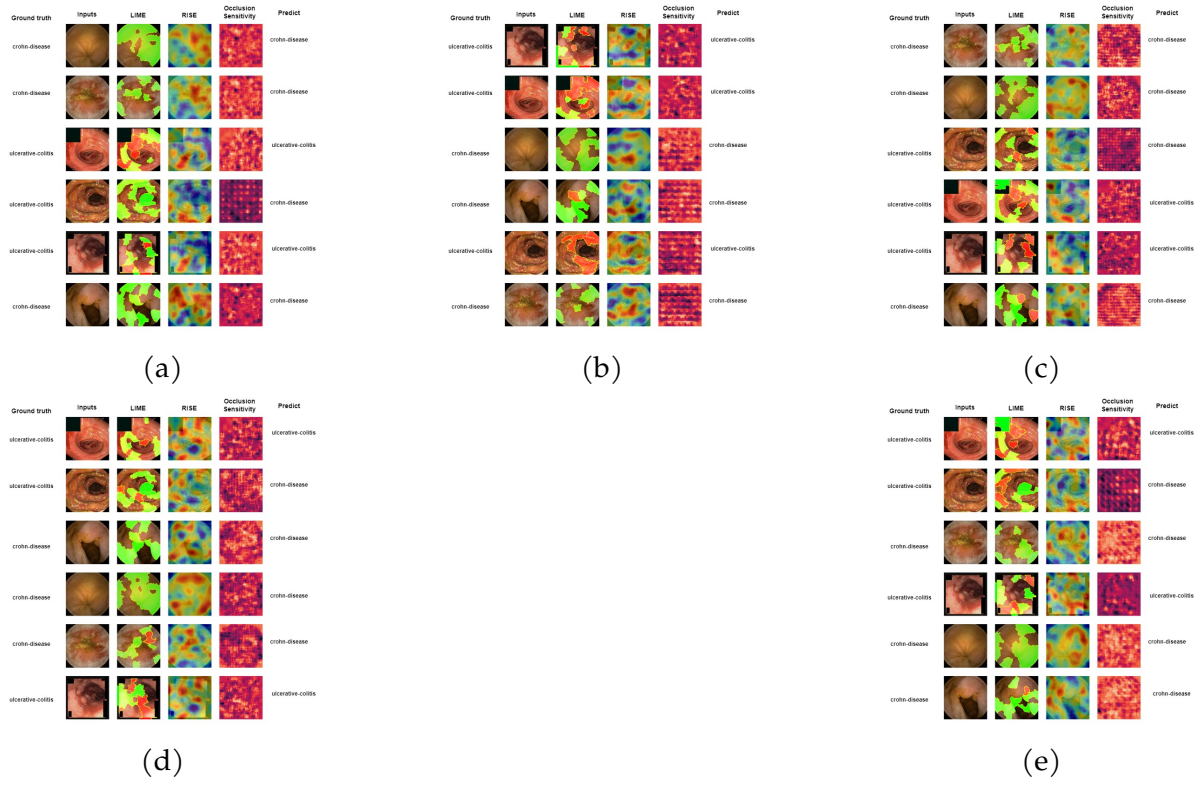


Figure 8.18: Interpretability of the ViTs student models in experiment 9: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.

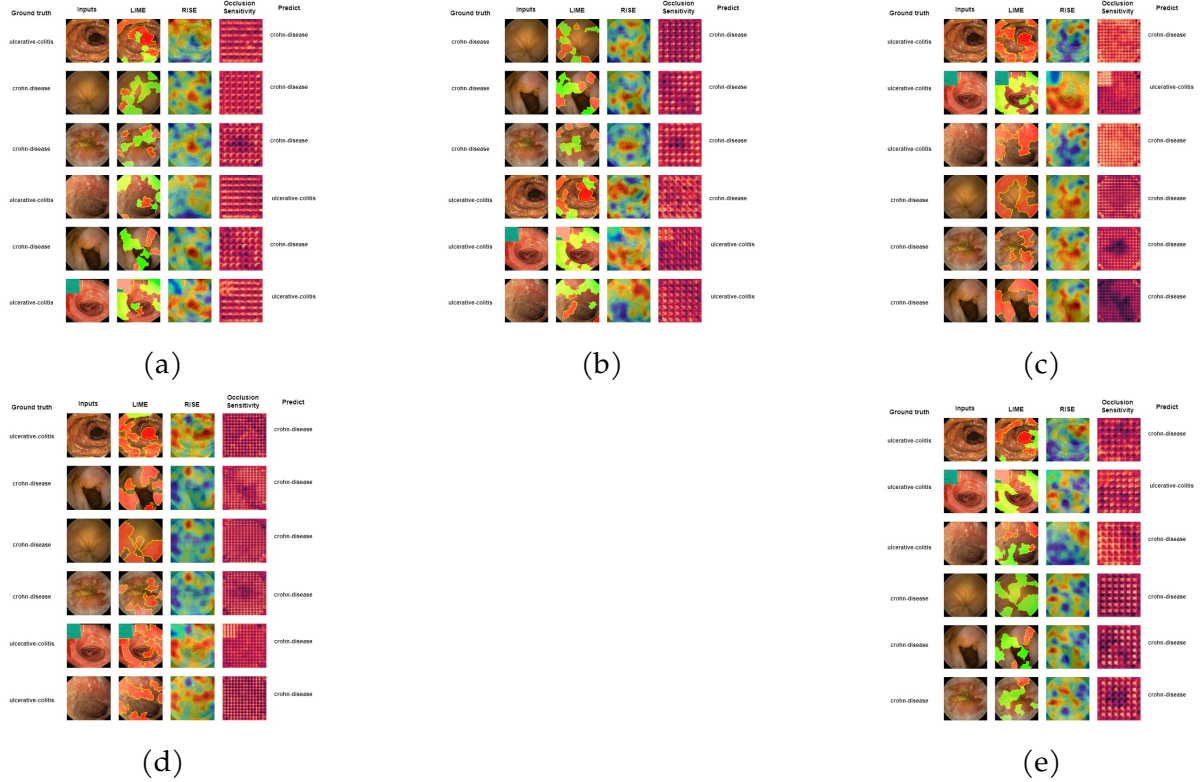


Figure 8.19: Interpretability of the DeiT models in experiment 1: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.

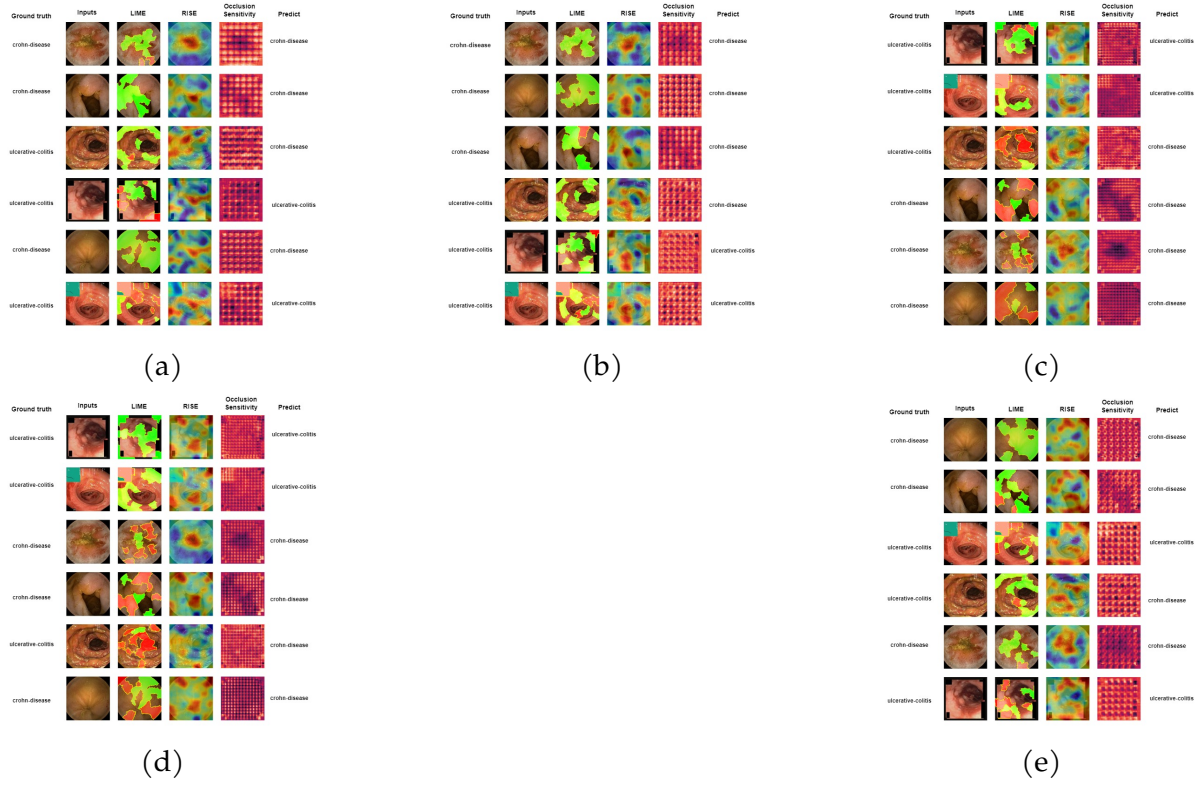


Figure 8.20: Interpretability of the DeiT models in experiment 2: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.

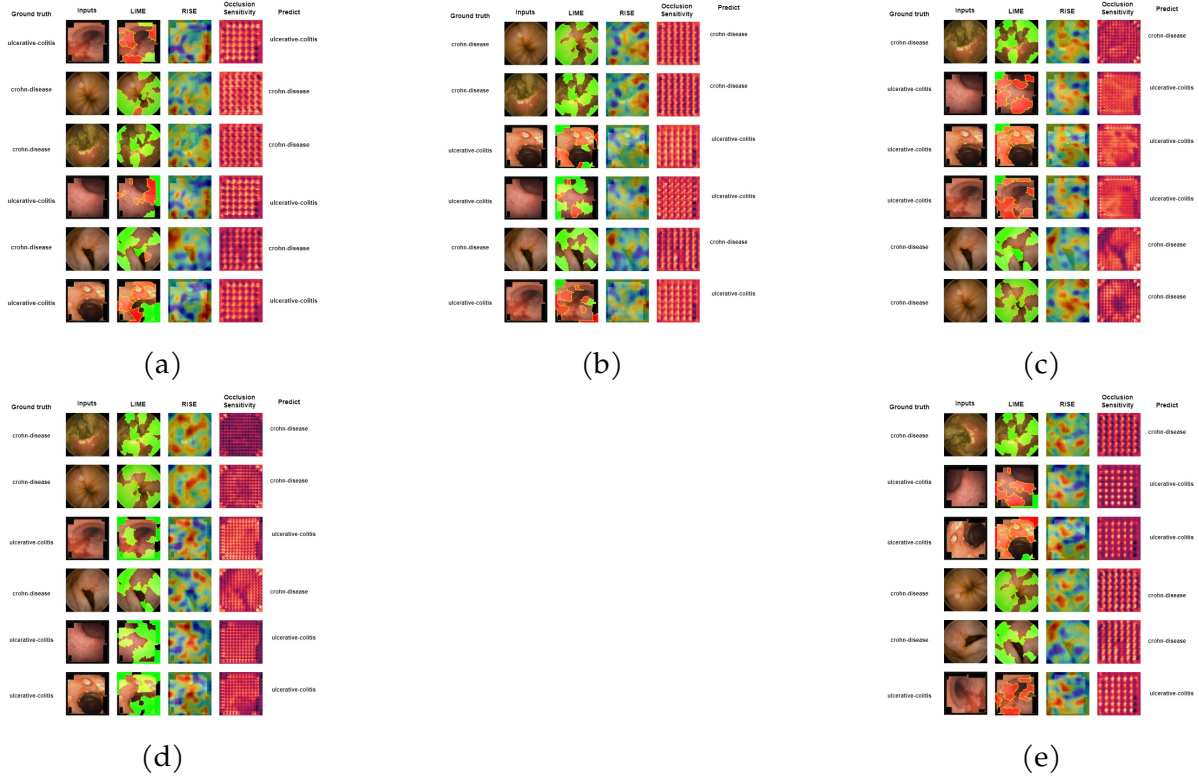


Figure 8.21: Interpretability of the DeiT models in experiment 3: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.

Diagnose of the IBD using Deep Learning Models with Interpretability

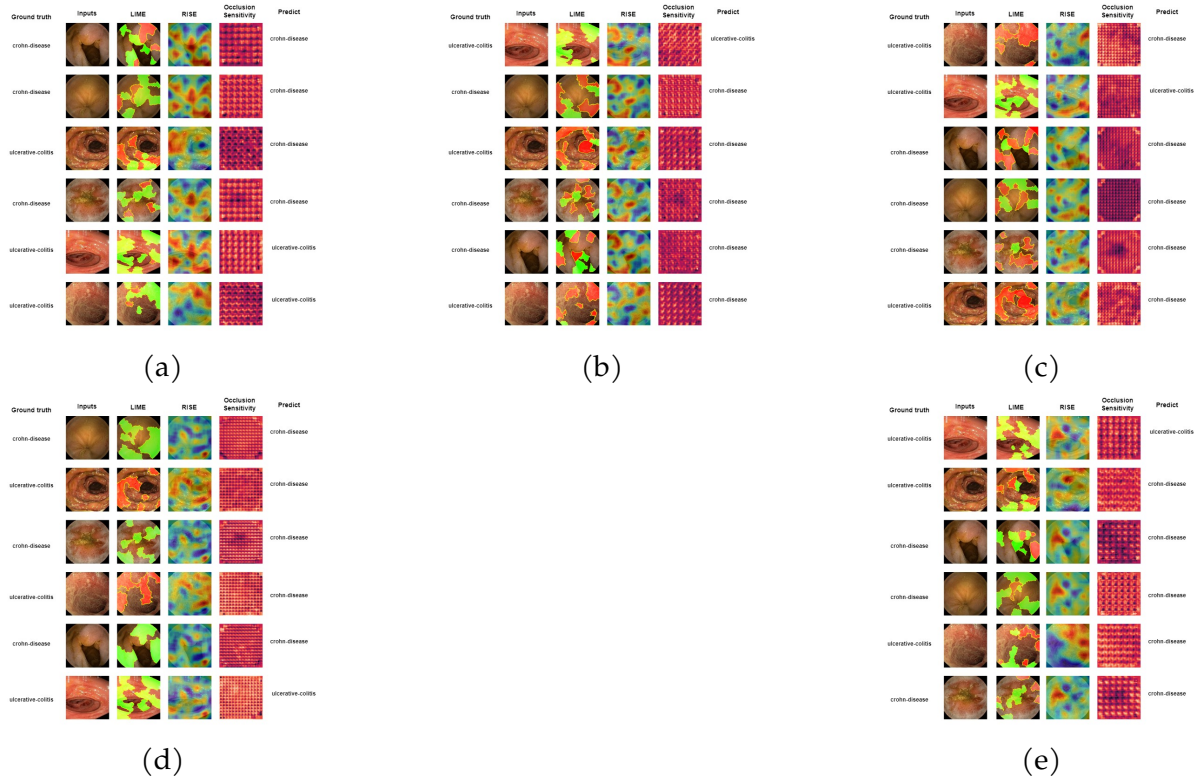


Figure 8.22: Interpretability of the ViT student models in experiment 4: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.

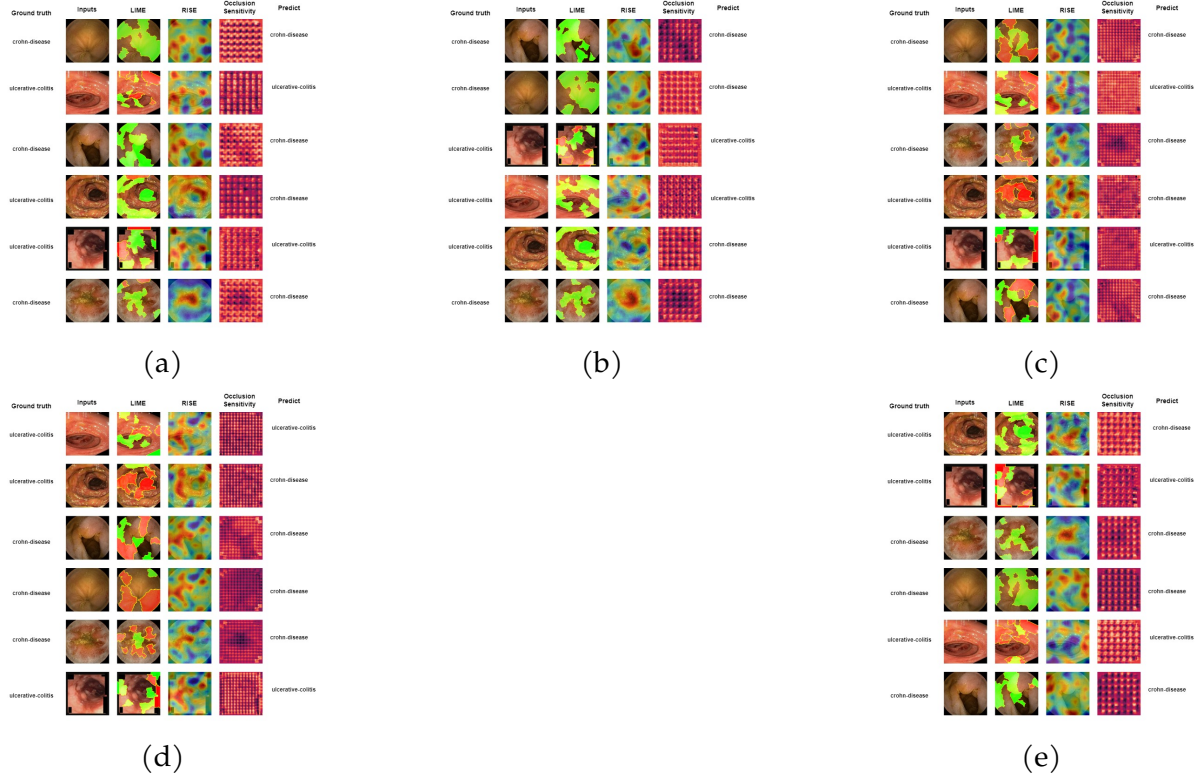


Figure 8.23: Interpretability of the DeiT models in experiment 5: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.

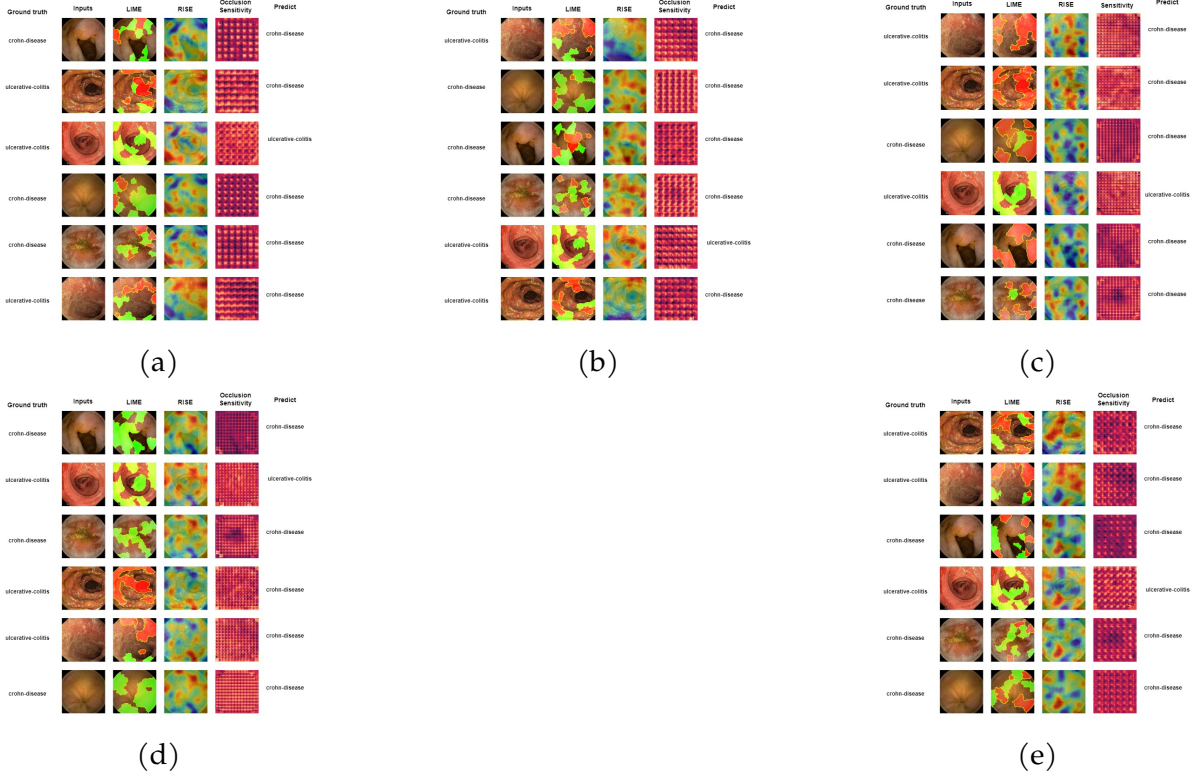


Figure 8.24: Interpretability of the DeiT models in experiment 6: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.

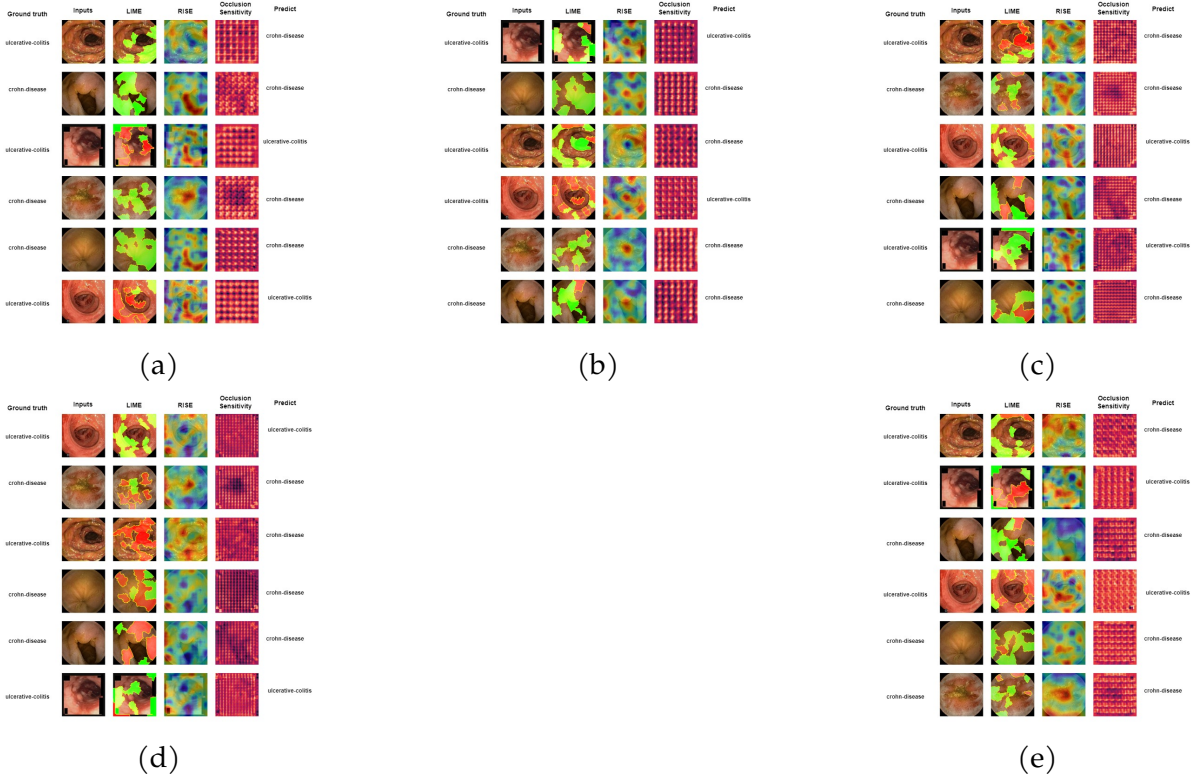


Figure 8.25: Interpretability of the DeiT models in experiment 7: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.

Diagnose of the IBD using Deep Learning Models with Interpretability

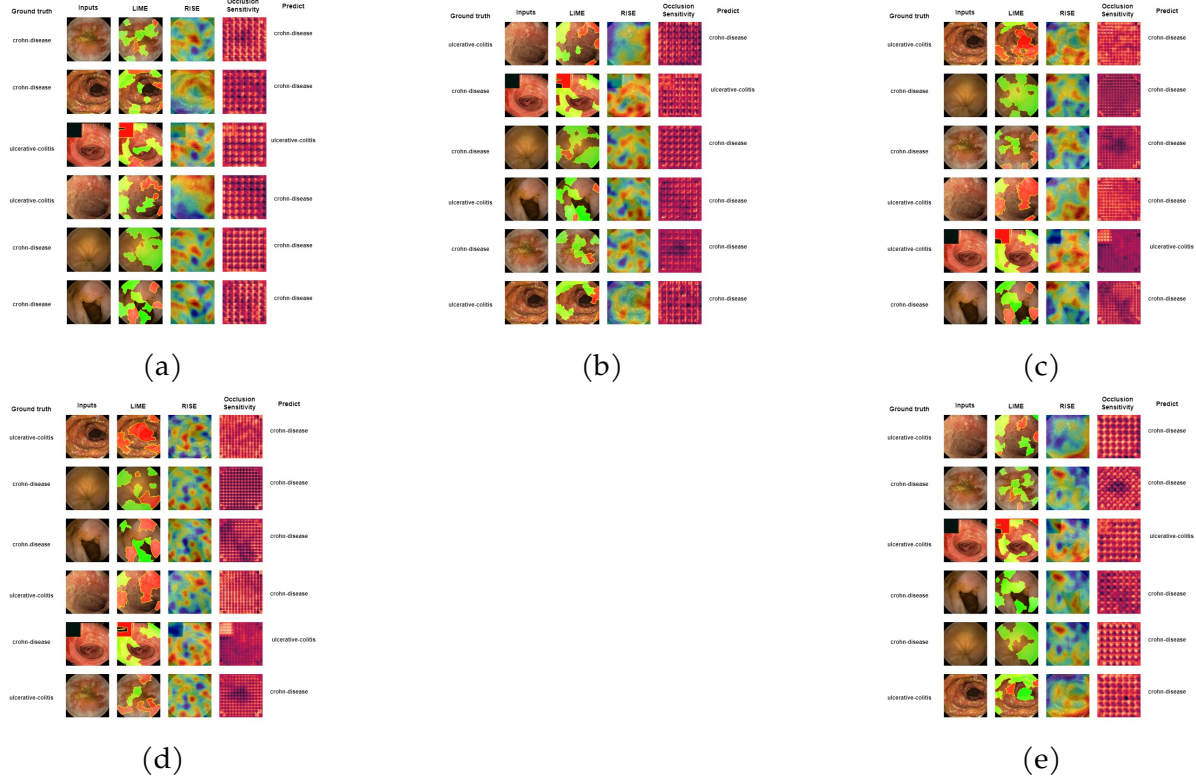


Figure 8.26: Interpretability of the DeiT models in experiment 8: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.

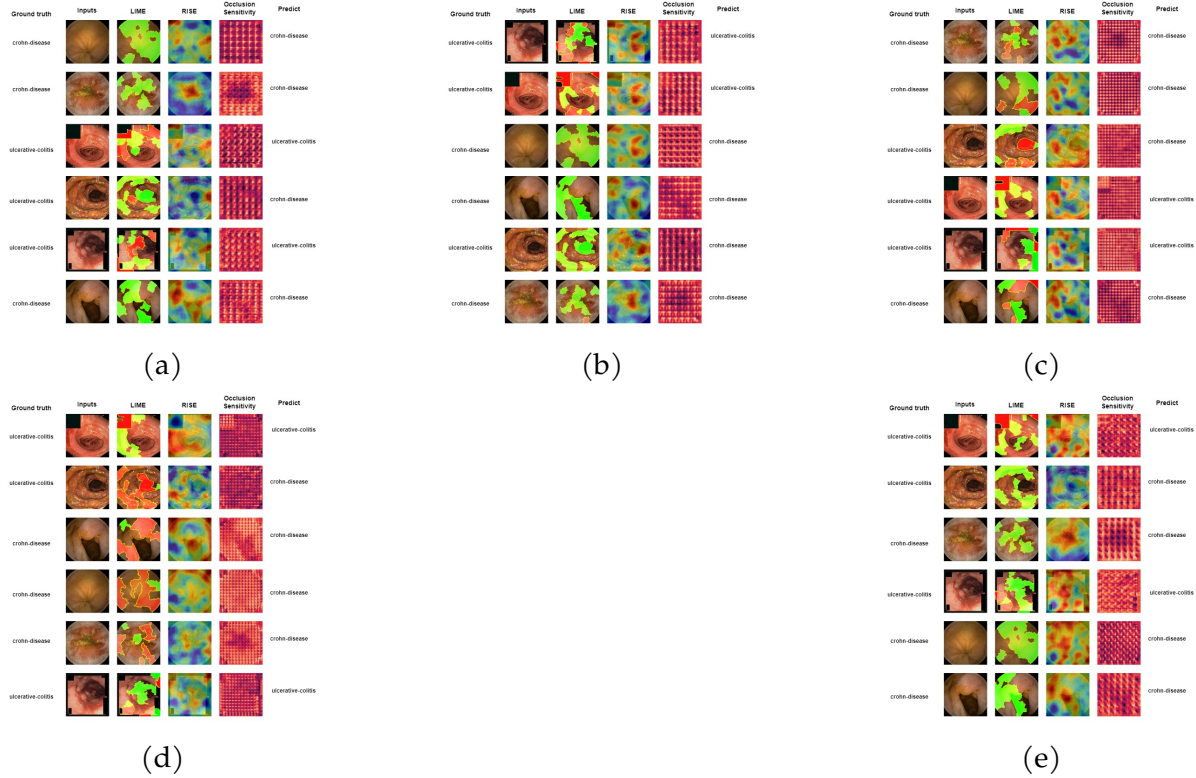


Figure 8.27: Interpretability of the DeiT models in experiment 9: (a) ViT-B/32; (b) ViT-S/32; (c) ViT-B/16; (d) ViT-S/16; (e) R26+S/32.



**Instituto Superior
de Engenharia**

Politécnico de Coimbra