Leveraging Foundation Models for the Classification of Precancerous Gastric Lesions in Endoscopic Images

Tomás Ferreira
Master in Computer Science
Department of Computer Science
Faculty of Sciences of the University of Porto
2025

Leveraging Foundation Models for the Classification of Precancerous Gastric Lesions in Endoscopic Images

Tomás Ferreira

Dissertation carried out as part of the Master in Computer Science

Department of Computer Science

2025

Supervisor

Miguel Coimbra, Full Professor, Faculty of Sciences of the University of Porto

“ Make an oath, then make mistakes
Start a streak you’re bound to break
When darkness rolls on you
Push on through ”

Oldies Station - Twenty One Pilots

Acknowledgements

I would like to thank my supervisor, Miguel Coimbra, for this opportunity to pursue an interesting research on this topic, and most importantly for all the guidance, patience, availability, useful support and valuable insights over this last year of work and research. It is an absolute pleasure to end my academic career with such a remarkably noble work.

I also express my gratitude to the members of the CAGE project for their support, guidance during the weekly meetings, and their availability to help with any doubts. In particular, I thank my colleagues João, David, and Nuno for accompanying me on this journey and for all the moments we shared over the past year.

My gratitude also goes to the Department of Computer Science of the Faculty of Science of the University of Porto for these past years: for all the learning opportunities and precious memories, I can honestly thank for being the ceiling of the 5 best years of my life.

I sincerely wish to thank INESC TEC and IPO-Porto for their innovative contributions. Furthermore, I want to thank the EndoRadiomics project for providing financial support in the scope of the Recovery and Resilience Mechanism of the European Union with reference 2024.07584.IACDC/2024.

My absolute gratitude goes to all the friends who accompanied me during this eventfully chaotic journey that is university, most concretely the people of our Meetings group: you all made my life so memorably fun these past few years. A special thanks to Emídio, Inês, Tiago and Alex for the proximal support you offered me over this beautiful past year.

To my girlfriend Fernanda, I cannot thank you enough for the positive presence you have had over this hard past year, it all would honestly be so much harder if you weren't by my side giving me more support than I could ever ask for, and, for that and so much more, I am thankful for life. Muchas gracias corazón <3

Saving the best for last, and I'll write this paragraph in portuguese, devo todas estas emoções à minha família: obrigado mãe e pai, por me permitirem todos estes anos de educação e por me apoiarem e amarem sempre em cada passo do meu caminho; obrigado maninho Duarte, por me dares forças em tentar ser o melhor exemplo possível para ti e por tudo que tenho o prazer de viver contigo; obrigado avó Marta, por todo o amor incondicional que me deste e dás. A todos vocês e o resto da minha família, só tenho a agradecer profundamente por todas as forças que me deram: amo-vos eternamente.

Resumo

O cancro gástrico continua a ser das principais causas de mortalidade relacionada com o cancro em todo o mundo. A deteção precoce de lesões pré-cancerosas por meio de endoscopia oferece oportunidades críticas de intervenção, mas a classificação precisa continua sendo um desafio devido à sutileza das características visuais e à alta variabilidade entre os observadores. Este trabalho tem como objetivo avaliar se os Modelos Fundacionais (FM), arquiteturas de grande escala pré-treinadas em milhões de imagens diversas, são capazes de classificar imagens endoscópicas de lesões gástricas pré-cancerosas.

Três versões base de paradigmas arquitetónicos foram avaliadas em configurações pré-treinadas e fine-tuned, aplicadas a um dataset de lesões gástricas pré-cancerosas (1280 imagens, 3 classes): ResNet-50 (CNN tradicional), ConvNeXt (CNN modernizada com ViT) e DINOv2 (ViT auto-supervisionado). As metodologias de treino abordaram especificamente as limitações do dataset através do unfreezing gradual, índices de aprendizagem distintos, augmentações avançadas e funções de perda focais. O ConvNeXt fine-tuned alcançou a maior precisão geral (66,80%), comparado com a baseline ResNet-50 pré-treinada (57,81%). No entanto, todas as configurações exibiram dificuldades críticas na classe minoritária (36,84% a 51,22% de recall), resultando em taxas de falsos negativos superiores a 60%. O DINOv2 Fine-Tuned demonstrou a melhor classificação geral das três classes (macro-F1 de 58,58%), mostrando melhor feature clustering e treino mais estável comparado aos modelos supervisionados.

Este trabalho fornece uma avaliação sólida dos FM para classificação de imagens endoscópicas. Embora promissores, persistem limitações críticas: dificuldades na deteção de classes minoritárias, custo computacional 3 vezes maior e desempenho insuficiente para implantação clínica autónoma. Estas descobertas estabelecem expectativas básicas realistas e identificam requisitos para avanços futuros, incluindo exploração de variantes maiores, pré-treino específico de domínio e arquiteturas especializadas para classes desbalanceadas.

Palavras-Chave: Ciência de Computadores, Modelos Fundacionais, Imagens Endoscópicas, Aprendizagem Profunda, Cancro Gástrico

Abstract

Gastric cancer remains one of the leading causes of cancer-related mortality worldwide. Early detection of precancerous lesions through endoscopy provides critical intervention opportunities, yet accurate classification remains challenging due to subtle visual features and high inter-observer variability. This work aims to investigate whether Foundation Models (FM), large-scale architectures pre-trained on millions of diverse images, can effectively classify endoscopic images of precancerous gastric lesions based on the Endoscopic Grading of Gastric Intestinal Metaplasia (EGGIM) protocol.

Three base versions of architectural paradigms were evaluated in pre-trained and fine-tuned configurations, using base model checkpoints on the EGGIM dataset (1280 images, 3 classes): ResNet-50, traditional Convolutional Neural Network (CNN); ConvNeXt, modernized CNN with Vision Transformers (ViT); DINOv2, self-supervised ViT (FM representative). Training methodologies specifically addressed small, unbalanced dataset constraints through progressive unfreezing, discriminative learning rates, advanced augmentation, and focal loss techniques. The results showed that ConvNeXt Fine-tuned achieved the highest overall accuracy (66.80%), compared to the ResNet-50 Pre-trained baseline (57.81%). However, all configurations exhibited critical Class 1 recall rates of 36.84%-51.22%, resulting in false negative rates exceeding 60% for the clinically crucial intermediate lesion category. DINOv2 Fine-Tuned demonstrated the best overall classification of the three classes (highest macro-F1 of 58.58%), showing that the foundation models have a more improved feature clustering and more stable training than their supervised counterparts.

This work provides a solid assessment of Foundation Models for endoscopic image classification. Although foundation models show promise, critical limitations persist: persistent minority class detection failures, 3x increased computational cost, and insufficient performance for autonomous clinical deployment. These findings establish realistic baseline expectations and identify requirements for future advancement, including exploration of larger model variants, domain-specific pre-training, and specialized architectures for addressing class imbalance inherent to endoscopic medical imaging.

Keywords: Computer Science, Foundation Models, Endoscopic Imaging, Deep Learning, Gastric Cancer

Table of Contents

List of Figures

List of Abbreviations

AG Atrophic Gastritis. 14

AI Artificial Intelligence. 1, 2, 33, 81, 84, 86, 88– 90

AMP Mixed Precision Arithmetic. 65, 70, 74

AUC Area Under the Curve. 15

CNNs Convolutional Neural Networks. 2– 4, 19– 21, 24, 44, 49, 50, 52, 57, 81, 86

DL Deep Learning. ix, 2, 17– 19, 23, 24, 26, 32, 43, 48, 49

EGC Early Gastric Cancer. 14, 37, 87

EGD Upper Endoscopy or Esophagogastroduodenoscopy. 10

EGGIM Endoscopic Grading of Gastric Intestinal Metaplasia. ix, 1– 5, 15– 17, 31, 43– 46, 49, 51, 52, 54, 55, 57, 60, 66, 67, 80, 81, 84– 88, 90

EMA Exponential Moving Average. 64, 70, 74

EMR Endoscopic Mucosal Resection. 8, 14

ESD Endoscopic Submucosal Dissection. 8, 14

EUS Endoscopic Ultrasound. 11

FFN Feed-Forward Network. 55, 57

FMs Foundation Models. x, 1– 4, 17, 21, 24, 26– 32, 41– 46, 49, 54, 57, 81– 90

GANs Generative Adversarial Networks. 25, 26

GC Gastric Cancer. ix, 1, 4– 11, 15, 16, 24, 43

GELU Gaussian Error Linear Unit. 51– 54, 57

GI Gastrointestinal. 4, 12, 15, 19, 20, 23, 24, 31, 33– 43, 45, 86, 88

GIM Gastric Intestinal Metaplasia. 3, 15, 16, 83, 87

IM Intestinal Metaplasia. ix, 14– 16, 46, 81, 82, 88

LR Learning Rate. 67, 70, 72, 74, 76, 78, 89

MHSA Multi-Head Self-Attention. 55– 57

ML Machine Learning. 18, 27

MLP Multi-Layer Perceptron. 51, 52, 54, 57

NBI Narrow-Band Imaging. ix, 10, 13– 15, 45

NNs Neural Networks. 22

OLGA Operative Link on Gastritis Assessment. 14

OLGIM Operative Link on Gastric Intestinal Metaplasia. 14, 15

ReLU Rectified Linear Unit. 50, 52, 54, 57

SGD Stochastic Gradient Descent. 70

t-SNE t-distributed Stochastic Neighbor Embedding. 66

TTA Test-Time Augmentation. 62, 63, 85

ViTs Vision Transformers. 2– 4, 20, 21, 24, 43, 44, 49, 52, 54, 81, 86

YOLO You Only Look Once. 22

1. Introduction

This chapter presents the research problem at the intersection of Gastric Cancer prevention and Artificial Intelligence, focusing on the potential of Foundation Models to accurately classify endoscopic images of precancerous gastric lesions. It starts by explaining the significance of this question, contextualizing the investigation in clinical realities and technical challenges posed by small and imbalanced medical datasets, and progresses to the specific objectives that guide this work along with the specific contributions it intends to deliver. Finally, the chapter concludes by detailing the layout of the rest of the dissertation, illustrating how each following section builds toward answering whether these models provide significant advantages over conventional methods in this clinical field.

1.1. Motivation and Context

Gastric cancer constitutes a major global health challenge as the fifth most prevalent cancer worldwide and the fourth leading cause of cancer-related deaths, with particularly high rates in East Asia, Eastern Europe and South America [1], and the 5-year survival rate in areas without effective screening programs remains under 30%, highlighting the crucial necessity of early detection during precancerous phases. Typically, the disease follows the Correa cascade, evolving through chronic gastritis, atrophic gastritis, intestinal metaplasia, dysplasia, and eventually adenocarcinoma over many years or even decades [2], with intestinal metaplasia marking a pivotal point where precise detection directly influences surveillance intensity and clinical management. However, diagnosis is challenging due to the subtle nature of endoscopic features and significant variability between observers, even among experienced gastroenterologists [3].

The EGGIM protocol establishes an organized framework for the endoscopic assessment of gastric intestinal metaplasia by examining its extent and distribution patterns: normal mucosa (Class 0), growing metaplasia affecting up to 30% of visible mucosa (Class 1), and advanced metaplasia exceeding 30% coverage (Class 2) [2], which is crucial for guiding surveillance strategies and evaluating the risk of cancer. However, Class 1 lesions, representing the most crucial period for intervention, exhibit varied and subtle visual characteristics that complicate the straightforward classification, affecting the real-world implementation of the protocol [3].

Artificial Intelligence (AI), particularly Deep Learning based on Convolutional Neural Networks and Vision Transformers, has shown outstanding potential in the field of medical image analysis [4]. Recent Foundation Models, large-scale Deep Learning (DL) architectures pre-trained on millions of diverse images using self-supervised learning techniques, offer the promise of enhanced transfer learning capabilities and improved generalization [5], as evidenced by models such as DINOv2, which utilizes contrastive learning and is trained on 142 million images to achieve cutting-edge results on computer vision benchmarks, and by hybrid architectures like ConvNeXt that efficiently blend the strengths of convolutional operations with the representational advantages of transformers [6]. These developments have generated optimism concerning AI's ability to enhance diagnostic accuracy and standardize evaluations in clinical practice.

However, translating general-domain Foundation Models (FMs) into real-world clinical scenarios encounters substantial obstacles when dealing with small and highly imbalanced datasets: the research investigates whether FMs can effectively classify endoscopic images of precancerous gastric lesions using the Endoscopic Grading of Gastric Intestinal Metaplasia (EGGIM) dataset, which exemplifies these challenges: approximately 1,000 training images with only 18% classified as Class 1, significant visual overlap between classes, and high intra-class variability [2], creating stress test conditions that determine whether the benefits of Foundation Models hold under such difficult circumstances. Recent work has explored FMs for gastric inflammation [7] and domain-specific models like EndoDINO for polyp segmentation [8], yet systematic comparative evaluation of Foundation Models specifically for endoscopic image classification remains limited, leaving essential questions about optimal model selection, training strategies, and realistic performance expectations unaddressed. These gaps are addressed through rigorous empirical investigation, leveraging the EGGIM dataset to evaluate whether Foundation Models can provide practical advantages to classify precancerous gastric lesions.

1.2. Objectives

This research pursues three primary objectives that focus on understanding whether Foundation Models can effectively support endoscopic image classification:

- **Systematic Architectural Comparison:** Evaluate three Foundation Models paradigms: ResNet-50 (traditional CNN), ConvNeXt (hybrid CNN-transformer), and

DINOv2 (self-supervised ViTs; FMs), in both their pre-trained and fine-tuned configurations for classifying endoscopic images of precancerous gastric lesions using the EGGIM dataset, establishing baseline performance metrics and identifying which architectures and pre-training strategies deliver practical benefits for this medical imaging task.

- **Training Methodology Assessment:** Analyze and evaluate training methodologies focusing on the challenges posed by small, highly imbalanced medical datasets, including progressive unfreezing schedules, discriminative learning rates, advanced enhancement protocols and class imbalance mitigation techniques, with an emphasis on determining which strategies are effective for adapting FMs to endoscopic image classification.
- **Limitation Analysis:** Provide an assessment of the existing strengths and weaknesses of FMs in the classification of endoscopic images of precancerous lesions, including predictions of achievable performance, the necessary computational resources, and fundamental data constraints, establishing practical expectations for the use of these systems in clinical practice.

1.3. Contributions

This thesis makes several concrete contributions to understanding FMs applicability for endoscopic image classification:

- **Systematic Benchmark for Endoscopic Image Classification:** As far as can be determined from the literature, this is the first comprehensive comparative evaluation of CNNs, hybrid CNN-transformer, and self-supervised FMs specifically for three-class Gastric Intestinal Metaplasia grading of endoscopic images of precancerous gastric lesions, providing reference performance metrics, detailed methodology and reproducible experimental protocols that establish baselines for possible future research in this domain.
- **Empirical Insights on Foundation Models Fine-Tuning:** Document the varying effectiveness of established training techniques in supervised and self-supervised pre-training paradigms, including the experimental observation that self-supervised models can endure significantly more aggressive unfreezing than their supervised counterparts, allowing for faster and more stable adaptation during fine-tuning in the classification of endoscopic images.

- **Assessment of Foundation Models Performance:** Provide a clear assessment of whether FMs can reliably classify endoscopic images, ensuring documentation of any failures in detecting minority class and computational constraints that persist even in the most advanced models, characterizing the gap between current capabilities and demands for clinical deployment.

1.4. Thesis Structure

The remainder of this dissertation is organized as follows:

Chapter 2: Reviews the epidemiology of Gastric Cancer (GC), the biological basis of precancerous progression, current clinical management approaches, including the EGGIM framework, and the role of endoscopy and biopsis in the diagnosis and management of GC.

Chapter 3: Introduces the fundamental families of deep learning computer vision algorithms, including Convolutional Neural Networks and Vision Transformers (ViTs), traces the evolution of FMs from supervised to self-supervised pre-training approaches, and discusses their specific relevance and challenges for Gastrointestinal (GI) medical imaging application.

Chapter 4: Provides a comprehensive systematic review of segmentation techniques applied to endoscopic gastrointestinal imaging, examining dataset characteristics, annotation practices, and model architectures reported in the recent literature, establishing the current state of the art while identifying gaps in standardization and reproducibility, and, in the end, discussing the potential of FMs for the medical imaging field using the conclusions taken from the systematic review.

Chapter 5: Presents a complete experimental framework to evaluate whether FMs can effectively classify endoscopic images, compared to two other relevant architectures, including detailed descriptions of model pipelines, training methodologies, evaluation protocols, comprehensive results in all configurations, and a detailed discussion of the findings on the effectiveness of FMs for endoscopic image classification.

Chapter 6: Summarizes key contributions and conclusions on the use of FMs for endoscopic image classification, while also discussing possible alternatives to follow as future work.

2. Background

Understanding Gastric Cancer requires an understanding of the biological journey that precedes it: a complex progression rooted in infection, persistent inflammation, and cellular transformation. This chapter lays that foundation by first situating GC within the global health landscape and following the molecular chain of events that transforms normal tissue into a cancerous one, before shifting to a practical clinical perspective by detailing how endoscopy has evolved into a vital tool for identifying early warning signs, particularly highlighting the development of Endoscopic Grading of Gastric Intestinal Metaplasia (EGGIM) as a standardized protocol for risk stratification. Together, these sections provide the essential background for understanding the significance of automated image analysis in endoscopic evaluations, in order to overcome the limitations of subjective visual interpretations and to facilitate earlier, more reliable detection of precancerous conditions when intervention is still an option.

2.1. Gastric Cancer

Gastric Cancer remains a major global health issue. It was the fifth most common cancer worldwide and the fifth leading cause of cancer-related death, accounting for approximately 4.9% of new cases and approximately 6.8% of cancer deaths, where 1 in 5000 people have GC [9]. During the first half of the twentieth century and especially in the 1930s, GC was one of the leading causes of cancer death in western countries [10]. Its incidence strongly declined after that period of time, though in the majority of the Western world, mainly due to improvements in food preservation (a higher tendency to refrigeration, reducing the reliance on salted/pickled foods) and better sanitation, which led to lower bacterial infection rates [10]. These trends sublimate the influence of environmental and infectious factors, leading to a closer examination of the biological processes that underlie gastric carcinogenesis.

2.1.1 Biological Process Overview

Gastric adenocarcinomas are not all the same and therefore can be categorized into different cancers by their histology. According to the *Laurén Classification*, the cancers can be divided into two main histological subtypes, intestinal and diffuse types, and a minority classified as mixed or indeterminate (Figure 2.1) [11].

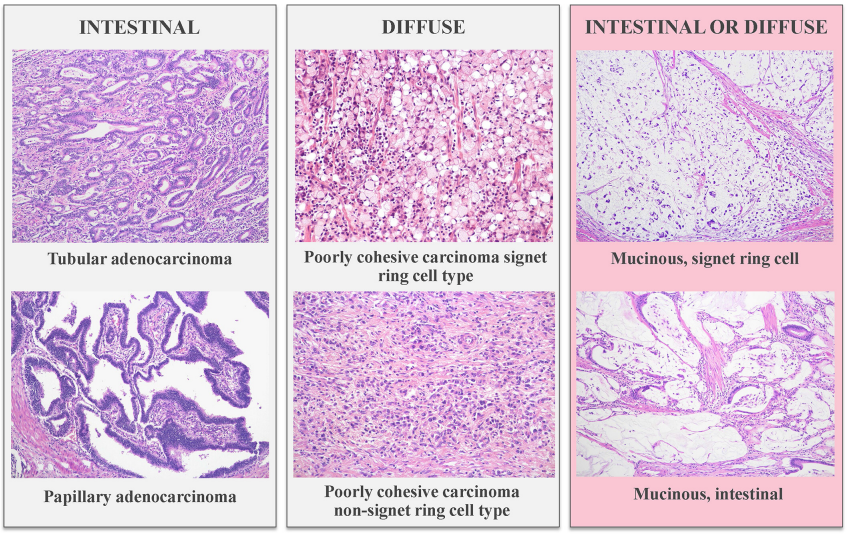


Figure 2.1: Illustration of intestinal, diffuse and mixed/indeterminate (intestinal or diffuse) types of gastric adenocarcinoma. Adapted from [12].

Gastric carcinomas are composed of cohesive gland-forming tumor cells similar to colonic adenocarcinoma, and tend to appear more in older male patients and high-risk populations, usually with its contraction related to its surrounding environment [11]. Its evolution usually follows a stepwise progression, denominated as *Correa Cascade*, driven by chronic inflammation, caused by a normal long-standing action of the *Helicobacter pylori* bacteria. This infection follows a cascade of inflammation - metaplasia - dysplasia - carcinoma that ultimately culminates in the emergence of irreversible GC. [11, 13].

In terms of diffuse-type carcinomas, they are formed by more discohesive cells that lack glandular formation. These types of carcinoma often present with signet rings, malignant cells with mucin pushing the nucleus to the periphery, and tend to occur in younger patients of both sexes of a more hereditary nature [11]. The precancerous sequence of this variant is more difficult to deal with, since these malignancies are used to increase the normal gastric epithelium without significant lesions, which can progress to clonal expansions of poorly cohesive malignant cells throughout the stomach wall through sudden molecular events such as the mutation of the CDH1 tumor suppression protein [11]. In consequence of this, early diffuse cancer can be missed endoscopically, and cancer is often discovered only at a more advanced stage, having more difficulty in achieving a positive prognosis.

The mixed/indeterminate type is characterized as a more non-homogeneous mixture of intestinal and diffuse types, with overlapped pathogenic characteristics of the two former adenocarcinoma variants [14].

Gastric cancers are also classified by location as noncardia/distally (body, antrum, pylorus) or cardia/proximal (near the gastroesophageal junction) [15]. They are also different on the way the epidemic propagates: the distal cancers are more associated with *H. pylori* infection and dietary/conservation factors, being historically more frequent and over the last years having its incidence declining in many countries due to better sanitation, reduced *H. pylori* prevalence and better food preservation [11, 15]; in contrast, the proximal GC is more difficult to deal with because of their late-stage prognosis (usually more complex surgical approaches) and have been actually increasing in some Western populations, sharing some risk factors with even esophageal adenocarcinoma such as obesity and gastroesophageal reflux disease [11].

2.1.2 Current Management of Gastric Cancer

As stated previously, Gastric Cancer remains a major global health concern, ranking among the top causes of cancer mortality worldwide [9]. Incidence rates have been declining in many western countries over the past decades, yet low-income regions and East Asia still face a high burden of GC [11]. Observing Figure 2.2, we can explain the distinct management of GC throughout the world:

- East Asia (Japan, Korea, Mongolia, China) has the highest incidence rates worldwide, often above 20 cases per 100,000;
- Eastern Europe, Russia, and parts of South America (Chile, Colombia) are other notable regions with high incidence of GC;
- North America, Northern/Western Europe, India, and much of Africa present low incidence rates (<5 per 100,000).

This global pattern mirrors differences in *H. pylori* prevalence, dietary exposures, socioeconomic conditions, and the existence of screening programs [11]. In Japan and South Korea, for example, they combine high incidence with relatively favorable survival outcomes, thanks to early detection through systematic screening [17], opposite to regions of low incidence that are not all used with GC and, because of that lack of preparation, often diagnose cases at later stages, contributing to poor outcomes despite a smaller overall number of cases [11].

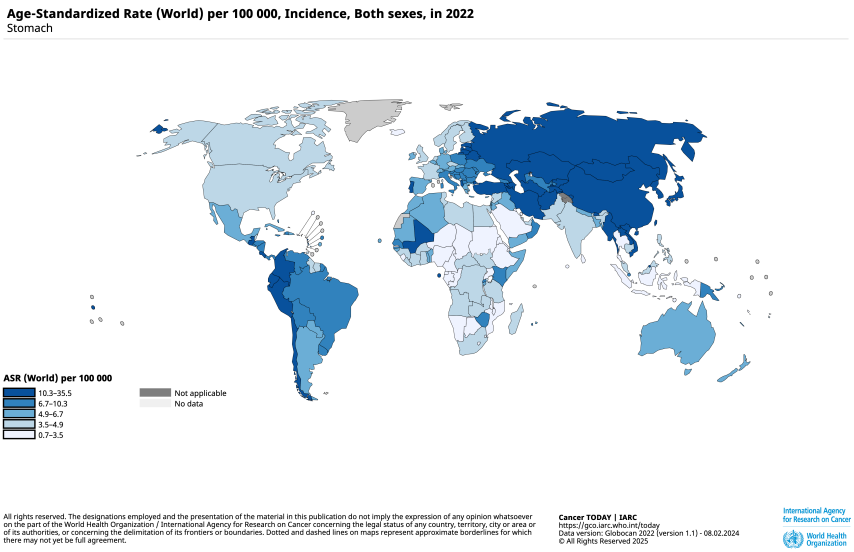


Figure 2.2: Age-standardized incidence rates of stomach/gastric cancer per 100,000 population (both sexes) in 2022. Adapted from [16].

This difference in outcomes matches how GC is treated distinctively in different countries in the world. For more localized diseases, the treatment is usually surgical resection: a subtotal or total gastrectomy (surgery to remove part or whole of the stomach) with D2 lymphadenectomy (more specific surgery to remove regional lymph nodes), to ensure complete clearance [15]. However, very early cases can sometimes be treated with minimally invasive endoscopic proceedings like Endoscopic Mucosal Resection (EMR) or Endoscopic Submucosal Dissection (ESD), even though it isn't that common since most patients are diagnosed too late [11].

One way that prospers the outcomes is when surgery is combined with systemic therapy: a drug-induced treatment that goes throughout the body via the bloodstream to damage or kill cancer cells. In western countries, perioperative chemotherapy is preferred (before and after surgery), while in East Asia it is preferred to undergo only adjuvant chemotherapy (after surgery) [13]. In case preoperative therapy is not given, chemoradiotherapy (combination of chemotherapy, the use of drugs, and radiotherapy, the use of high energy rays to destroy cancer cells) can also be considered as an adjuvant option [13, 15].

For more advanced or metastatic diseases, treatment is based on systemic chemotherapy [15]. A major step forward has been the integration of immunotherapy. Checkpoint inhibitors boost the immune response and are effective in certain tumors or those with high microsatellite instability status (condition in which the patient's cancer cells have a high number of mutations (changes) within short and repeated DNA sequences) [13].

In general, the trend in GC management is moving toward a personalized and multimodal approach, where surgery remains central, but the results are maximized by integrating chemotherapy, targeted therapy and immunotherapy, tailored to the stage and biology of the tumor.

2.1.3 Risk and Prevention in Gastric Cancer

Beyond treatment, understanding the risk factors and available prevention strategies of Gastric Cancer is critical to reduce the global burden of GC. GCs are typically sporadic, but 5% to 10% of the cases have a familial or genetic association [11]. The key risk factors for GC are:

- **Helicobacter pylori Infection:** the most significant risk factor, especially for non-cardia / distal tumors. *H. pylori* causes chronic gastritis that can progress to cancer [11, 18];
- **Dietary Habits:** high intake of salt-preserved foods (like salted or pickled foods) and processed meats increases the risk of GC, specially for non-cardia/distal GCs [11, 18]. In contrast, diets richer in fresh fruits and vegetables are thought to be preventive to this front;
- **Epstein-Barr Virus:** common and highly contagious human herpes-virus, that spreads through bodily fluids, especially saliva, and that in the more pessimist scenarios has been noted in association more and more with GC, usually affecting the cardia/proximal section of the stomach [11];
- **Lifestyle Factors:** tobacco smoking and heavy alcohol consumption are risks well-established for GC in both cardia and non-cardia sites. Obesity is also linked to a higher incidence of cardia cancers, mainly due in part to obesity-related conditions like gastroesophageal reflux [11, 18];
- **Family History and Genetics:** the sole factor of having a first-degree relative with GC significantly elevates one's risk of catching it too, even though truly hereditary cases make less than 3% of total GCs [18]

However, sufficient prevention can be taken to avoid obtaining GC. Some of the most remarkable strategies are:

- **Reduce Key Risk Factors:** intuitively, prevent what is known to cause cancer. Detecting and eradicating *H. pylori* and adopting healthy lifestyle changes are effective primary efforts to mitigate the chances of GC [11];
- **Screening and Early Detection:** as previously stated, population screening programs (typically periodic endoscopic examination) are common on countries with high GC incidence like Japan and South Korea, in order to detect early-stage and more treatable tumors [17];
- **High-risk Prophylactic Measures:** more aggressive preventive interventions, like prophylactic total gastrectomy (surgical removal of the stomach), are sometimes opted for in younger high-risk individuals, particularly those in their twenties carrying hereditary GC syndromes. This is an effective treatment, though it requires careful counseling about its impact on quality of life [15].

2.1.4 Role of endoscopy in Gastric Cancer Management

As observed, effective early detection of Gastric Cancer is heavily based on endoscopic evaluation. Endoscopy not only enables the identification of early or precancerous malignant changes in high-risk populations, but also plays a pivotal role in structured screening programs such as those from Japan [17].

As primary diagnosis, Upper Endoscopy or Esophagogastroduodenoscopy (EGD) is the conventional diagnostic procedure when suspicions of GC arise, allowing direct visualization of the gastric mucosa and immediate biopsy of any suspicious lesions [11], as we can see in Figure 2.3.

In endoscopy, extensive sampling is required, a method in which any suspicious lesion, such as ulcers or nodes, should be sampled with multiple biopsies during endoscopy to increase diagnostic sensitivity as much as possible [11]. Advanced endoscopic techniques (such as Narrow-Band Imaging (NBI), a technique that uses special light to help better detect precancerous conditions), are also common practices during an EGD with the goal of highlighting subtle mucosal abnormalities that might be missed on standard white light examination [11].

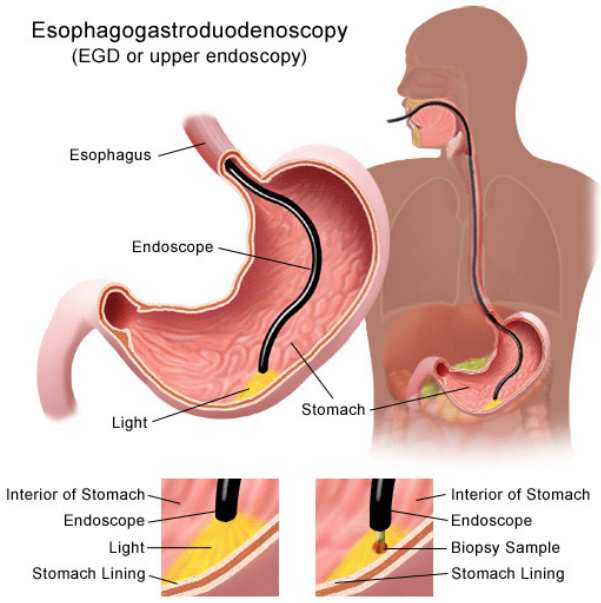


Figure 2.3: Illustration of GC anatomy and diagnostic visualization during upper endoscopy. Adapted from [19].

After this initial endoscopic diagnosis has been performed, Endoscopic Ultrasound (EUS) is typically used to assess the extent of the tumor by examining its depth and evaluating regional lymph nodes (Figure 2.4). EUS can also help the fine-needle aspiration of abnormal nodules [11].

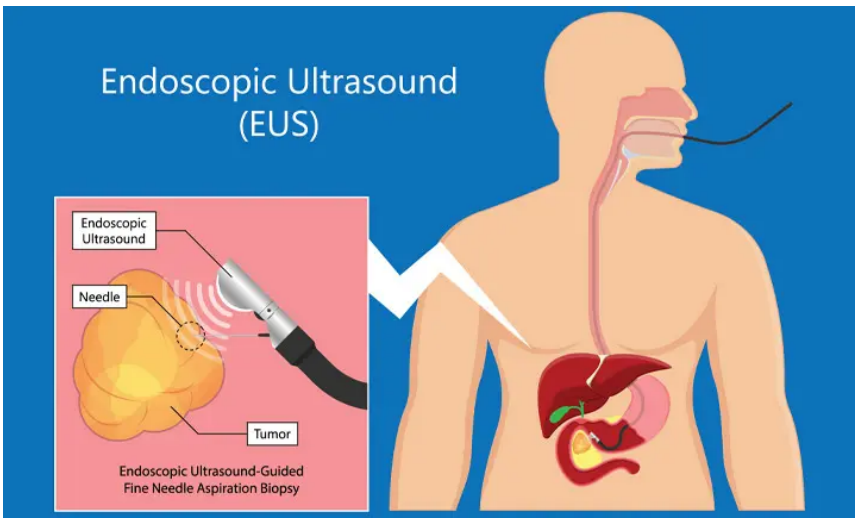


Figure 2.4: Educational diagram showing the role of endoscopic ultrasonography in gastric cancer staging. Adapted from [20].

2.1.5 Role of biopsy in Gastric Cancer Management

Despite the fact that endoscopy provides the visual pathway to suspect malignant lesions, a definitive diagnosis requires tissue confirmation. Biopsy, performed during endoscopic procedures, ensures histological validation and remains indispensable to appropriately confirm the diagnosis of gastric adenocarcinoma [11]. The success of this operation passes by the completion of these requirements:

- **Maximize detection efficiency** (obtaining multiple biopsy samples from a lesion) [11];
- **Molecular profiling** (testing of biopsy specimens for various histological and molecular markers; yield a more targeted response) [13];
- **Metastasis confirmation** (if imaging suggests expansion of the disease, biopsy of suspected metastatic sites is performed to confirm gastric origin and enable additional biomarker testing) [15].

2.2. Endoscopic Imaging and EGGIM

2.2.1 Description of an Endoscopy Exam

As already observed in Figure 2.3, an esophagogastroduodenoscopy is a flexible endoscopic exam of the esophagus, stomach, and proximal duodenum [21]. However, in order to proceed with the exam, some preparations must be made: patients fast (6 to 8 hours without solids and at least 2 hours without liquids) before the exam, should stop taking blood thinners if the patient is diabetic, and are given sedation or light anesthesia to minimize discomfort and induce amnesia [21, 22].

The procedure can now start with the insertion of the endoscope, a long and thin flexible tube with a light and camera (Figure 2.3), through the patient's mouth and down the throat, advancing it through the esophagus to the stomach and duodenum, while pumping air through the endoscope into the patient's stomach and duodenum to make it easier to see the organs [22]. The camera at the tip of the endoscope transmits high-definition real-time video to a monitor in the endoscopy room. Now, the gastroenterologist can directly inspect the mucosal lining of the upper GI tract (esophagus, stomach and duodenum) for abnormalities [22]. Modern endoscopes even use high-definition optics and specialized

modes, such as NBI, to improve mucosal visualization. These enhancements can highlight subtle lesions (such as intestinal metaplasia or dysplasia) that might be missed in the standard white-light view [23].

To further interact safely with the stomach, the endoscope has inside it a working channel for the gastroenterologist to use instruments. Through this channel, the physician can introduce tools (such as biopsy forceps or snares, tissue retrieval devices) with the objective of sampling tissue or treating problems on the spot, such as taking biopsies, removing polyps, or dilating structures, during the current exam [21, 22]. Once the stomach has been fully inspected and the tissues have been sampled (if sampled), the endoscope is then withdrawn: before leaving the stomach, the air should be suctioned, and then the esophagus is again examined after the endoscope is removed [21].

In parallel, a nurse monitors the vital signs of the patient, such as heart rate, blood pressure, and oxygen saturation, throughout the exam: if something is wrong, the gastroenterologist is immediately alerted to check it. An upper endoscopy usually takes about 30-60 minutes, and complications following the endoscopy are rare (less than 2% of the cases) [21, 22].

2.2.2 Lesions and Visual Evolution of Gastric Cancer process viewed by Endoscopy

It is now evident that endoscopy allows for a direct and high-quality visualization of the sequential mucosal changes that underlie gastric carcinogenesis. These changes follow the aforementioned *Correa Cascade*, a stepwise progression that illustrates how a Helicobacter pylori infection can lead to chronic gastritis, atrophic gastritis, intestinal metaplasia, dysplasia, and eventually carcinoma [11, 13] (Figure 2.5).

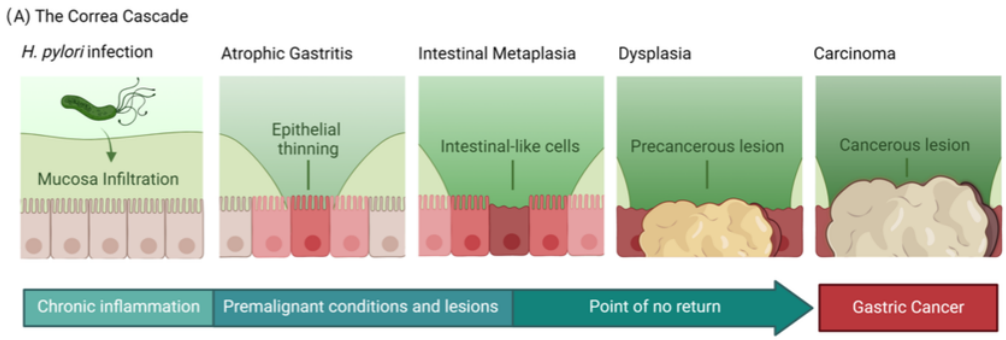


Figure 2.5: Schematic representation of the Correa Cascade, illustrating the sequential steps of gastric carcinogenesis: from chronic gastritis to atrophic gastritis, intestinal metaplasia, dysplasia, and carcinoma. Adapted from [7].

This framework is crucial for interpreting the distinctive characteristic lesions seen during an endoscopy.

1. **Atrophic Gastritis (AG)**: after a chronic gastric inflammation of the mucosa caused by the *H. pylori*, the mucosa starts to thicken itself more, often subtly but endoscopically noticeable as pale or transparent mucosa with prominent vasculature and a clear atrophic border (demarcation between normal pink and pale atrophic areas) [24];

2. **Intestinal Metaplasia (IM)**: phase where there's a growing replacement of gastric mucosa by intestinal-type cells. Endoscopically, IM appears as a patchy gray-white nodular with slightly elevated plaques in the uneven pale mucosa [24]. Narrow-Band Imaging often highlights the "light blue crest" sign, strongly associated with IM, as per Uedo et al. [25];

3. **Dysplasia**: also known as Intraepithelial Neoplasia, it is considered as the true precancerous stage; the metaplastic intestinal tissue becomes now slightly elevated, depressed, and irregular mucosal areas, now clearly distinguishable. Classification frameworks like the Paris Classification help describing the subtle surface changes evolution [26];

4. **Early Gastric Cancer (EGC)/Carcinoma**: defined as gastric adenocarcinoma confined to the mucosa or underlying tissue, regardless of lymph node status [15]. Endoscopically, these lesions may appear as flat, slightly elevated, or depressed areas with subtle irregularities in surface or vascular patterns, often requiring high-quality imaging for detection [17]. The Paris Classification also provides a standardized framework to categorize these more advanced superficial lesions, which also assists in guiding correct treatment decisions such as EMR or ESD [26].

2.2.3 EGGIM Protocol

Together, these lesions form a visual continuum (Figure 2.5), from subtle background changes (AG, IM) to precancerous dysplasia and finally invasive EGC. Although histological systems such as Operative Link on Gastritis Assessment (OLGA) and Operative Link on Gastric Intestinal Metaplasia (OLGIM) focus on evaluating and stratifying the severity and distribution of chronic Atrophic Gastritis with or without Intestinal Metaplasia [27], endoscopic morphology systems such as Paris Classification emphasize more on dysplasia and Early Gastric Cancer [26].

To standardize endoscopic risk assessment, these approaches converge in the EGGIM. This protocol provides a systematic NBI-based staging of intestinal metaplasia, as it parallels the histologic OLGIM system to stratify cancer risk: extensive Gastric Intestinal Metaplasia (GIM) on NBI (high EGGIM score) correlates with a markedly higher risk of GC [1, 2].

The scoring method is the following: five different areas were considered (two areas in the antrum, two in the corpus, and one in the incisura). Each area was scored 0 (no GIM), 1 (focal GIM, ≤ 30% of the area) or 2 points (extensive GIM in that area, > 30% of the area), culminating in graded endoscopic pictures for the amount of GIM in the section of the GI tract, as we can observe in Figure 2.6. The total EGGIM score (0–10) is then the sum between the regions [2].

Figure 2.6: Representative NBI endoscopic images used in the EGGIM protocol. (a) Normal mucosa (EGGIM=0); (b) focal Intestinal Metaplasia (EGGIM=1, ≤30%); (c) extensive Intestinal Metaplasia (EGGIM=2, >30%). Adapted from [3].

Performing a correlating analysis between EGGIM and OLGIM, a stratification of cancer risk can be calculated: in Pimentel-Nunes et al., a multicenter validation study conducted by some of the creators of the EGGIM, an EGGIM cutoff ≥5 predicted high-risk OLGIM III/IV, stage associated with advanced GIM, with an Area Under the Curve (AUC) around 0.97, an extremely high precision value [3]. This and other studies ultimately support the same conclusion of fixing an EGGIM overall cutoff score EGGIM of 4, to better help to decide whether biopsies are needed (>4) or not (0-4) to identify patients at risk of GC [3].

This way of grading GIM is reproducible and reliable even in more advanced endoscopy scenarios, but it could encounter some slight limitations regarding the necessity of the EGGIM graders having a significant NBI experience for consistent results and the use of comparable gastroscopes [3].

By mapping IM in the stomach, EGGIM guides the biopsy strategy. In practice, targeted biopsies of endoscopically identified IM regions achieved >90% diagnostic accuracy [2], which simplifies surveillance by giving more focus to high-risk patients: endoscopists can now immediately grade the GIM extent at the time of examination, optimizing their time.

Together, these features establish EGGIM as a practical endoscopic tool that complements the Paris/OLGA frameworks: it quantifies IM burden at the time of endoscopy and highlights patients at elevated GC risk.

3. Deep Learning and Foundation Models

This chapter lays the foundation for the research discussed in the following chapters by first exploring how Deep Learning tackles major challenges in endoscopic interpretation, such as observer variability and subjective grading inconsistencies, and then elaborating on the four families of DL algorithms: classification, detection, segmentation, and generation, while highlighting their architectural foundations and applications in clinical settings. The focus then moves to Foundation Models, which represent a paradigm distinct from traditional DL by employing large-scale pre-training and efficient task adaptation, and by comparing task-specific training with transfer learning, the chapter explains the advantages of FMs in the context of medical imaging under data scarcity. Finally, it reviews successful applications in gastroenterology and their practical limitations, setting the stage for the subsequent experimental work.

3.1. Deep Learning in Gastrointestinal Imaging

Endoscopic grading protocols, in the same way that revolutionized the area, still failed to achieve certain aspects that compromised their diagnostic reliability, mainly in substantial variability between observers: pathologists often disagree with logical classifications even when using established systems [28]. This subjectivity creates reproducibility challenges that directly impact patient care and treatment decisions.

The Endoscopic Grading of Gastric Intestinal Metaplasia (EGGIM) protocol, while systematic and extremely helpful, is vulnerable to these expert-dependent limitations regarding the nature of interpretations, which inevitably leads to inconsistent grading among different annotators and compromises the trustworthiness of the results.

DL enters this field offering an unbiased lens by providing an objective, reproducible classification that eliminates human subjectivity [28]. Veldhuizen et al. proved that DL models are capable of achieving better performance at faster speeds in gastric cancer stratification by mining subtle image characteristics and applying them consistently in clinical data sets, showing better survival stratification than pathological Laurén classification, even when pathologists did not identify relevant prognostic patterns. The transition from these expert-dependent protocols to DL models proves to be the unavoidable path to standardized and automated assessment systems and promises to reduce the clinical impact of variability between observers on gastric cancer classification.

3.2. Main Tasks of Deep Learning Algorithms

Basically, Deep Learning is a subset of Machine Learning that uses multilayered neural networks (deep neural networks) to mimic the complex decision-making power of the human brain, and what makes these models different from the Machine Learning (ML) ones is the fact that they use more than three computational layers of simple neural networks (typically hundreds or thousands), instead of the one or two layers used on traditional ML models [29]. Based on their learning process on the unsupervised learning thesis, DL models are able to extract the characteristics, features, and relationships they need to produce accurate output from raw, unstructured data and evaluate and refine their outputs if needed for increased precision [29]. With these characteristics, Deep Learning enters the medical imaging fields as a real problem solver, helping to perform tasks such as the detection of abnormalities, segmentation of medical images, classification of diseases, or even the generation of synthetic medical data, all while generally with high response accuracy [4].

Main Tasks of Deep Learning Models: DL in medical imaging generally clusters into four task families: classification, object detection, segmentation, and image generation, each solving distinct clinical questions that map well to endoscopic workflows (Figure 3.1) [4], with encoder backbones that can be adapted between tasks.

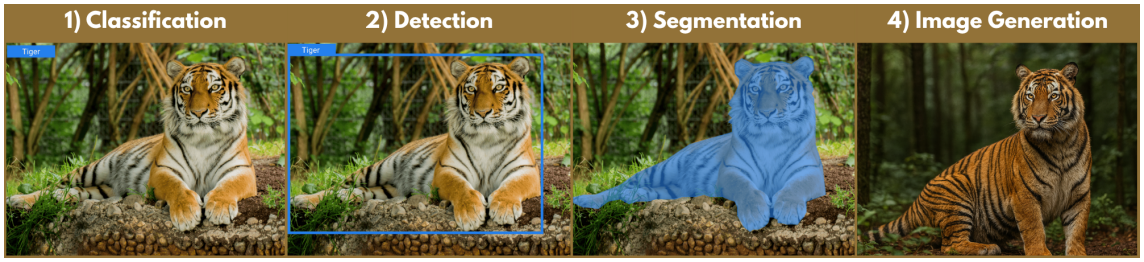


Figure 3.1: Overview of the four main DL task families in computer vision: (1) Classification assigns image-level labels, (2) Detection localizes objects with bounding boxes, (3) Segmentation provides pixel-level classification, and (4) Image Generation creates synthetic visual content. These foundational approaches form the basis for medical imaging applications across diverse clinical workflows. Adapted from [30] with image generation of the fourth image via ChatGPT.

3.2.1 Image Classification

The DL models with classification as its task have the goal of categorizing images into user-defined classes for various applications [30]: in the medical field, which could be determining the presence of a certain disease or even grading the severity itself.

Deep learning has significantly improved classification accuracy in Gastrointestinal imaging, reducing the incidence of missed diagnoses and diminishing the variability between observers discussed earlier, by learning subtle features (color, texture, patterns) to distinguish patterns with an accuracy that often matches expert performance. In the GI field, they face a huge obstacle referring to the limited number of large annotated GIendoscopic datasets: as of that, the common solution is to practice transfer learning from general image datasets. The models go through a pre-training process in complex datasets such as ImageNet [31] to be fine-tuned in the endoscopic images, taking advantage of the learned characteristics and increasing the accuracy of the model predictions on the small GI datasets [4, 7, 28].

Some of the main classification architectures include:

- **Convolutional Neural Networks (CNNs):** type of neural network, which is composed of node layers, containing:
  - an input layer;
  - one or more hidden layers (the more layers, the more complex the CNNs get);
  - an output layer.

  Every node is linked to another, having an associated weight and threshold. A node becomes activated and transmits data to the next network layer if its output exceeds the given threshold; otherwise, the data doesn't pass through the next layer of the network [29]. CNNs can also be deconstructed into three types of layers (Figure 3.2):
  - convolutional layers (extract features through learnable filters identifying features such as edges, textures, and patterns);
  - pooling layers (downsample feature maps, decreasing spatial dimensions while retaining important information);
  - fully connected layers (map extracted features with class probabilities to make final classification decisions) [4, 29].

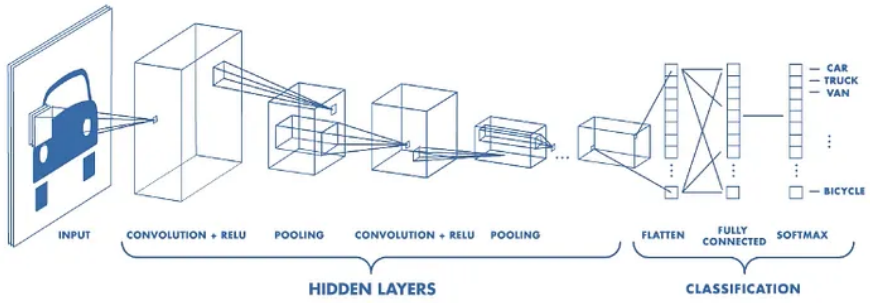


Figure 3.2: Convolutional Neural Network (CNN) architecture showing the hierarchical feature extraction process. The network progressively transforms input images through convolutional layers, pooling operations, and fully connected layers, extracting increasingly complex features from low-level edges to high-level semantic representations suitable for classification tasks. Adapted from [32].

In medical image classification, CNNs are positioned as the cornerstone of the field: architectures like ResNet, VGG, and DenseNet extract hierarchical features (edges, textures, shapes) from endoscopic images and are achieving high accuracies in GI tasks [33], and as it is going to be seen later, CNN-based classifiers often form the backbone of computer-aided diagnosis systems in endoscopy, surpassing the performance of earlier established approaches.

- **ViTs:** innovative neural network architecture introduced by Dosovitskiy et al. that applies transformer mechanisms originally created for natural language processing to computer vision tasks, mainly applying self-attention mechanisms to model global context across the image. ViTs are comprised of various key components:
  - an input preprocessing phase that segments images into fixed-size patches (typically 16x16 pixels, with each patch as a token similar to the words in natural language processing);
  - an embedding layer that projects these flattened patches into high-dimensional vectors;
  - plenty of transformer encoder blocks incorporate multi-head self-attention mechanisms, enabling every patch to consider all other patches;
  - a classification head with a learnable CLS token that generates final predictions [35]

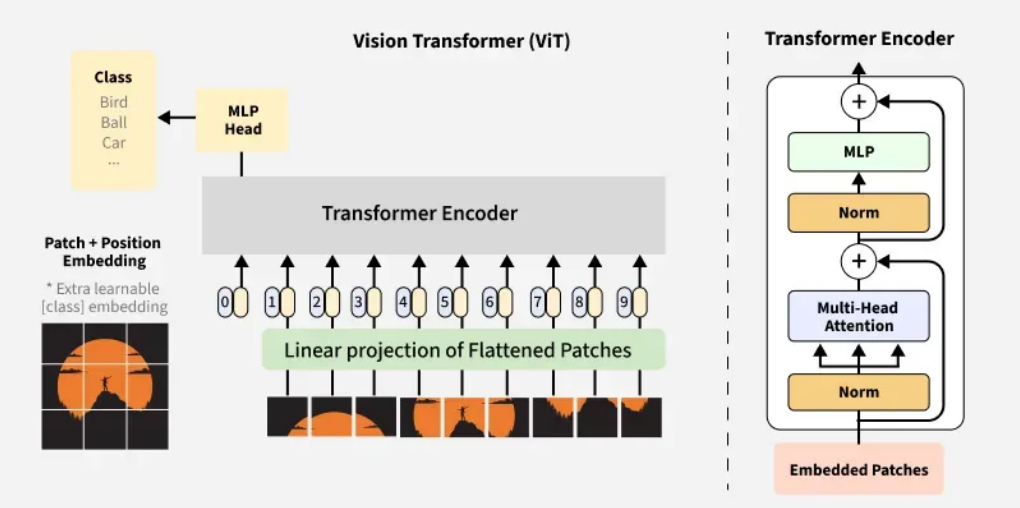


Figure 3.3: Vision Transformer (ViT) architecture demonstrating the patch-based approach to image processing. Input images are divided into fixed-size patches, linearly embedded, and processed through transformer encoder blocks with multi-head self-attention mechanisms, enabling global context modeling across the entire image simultaneously. Adapted from [36].

In contrast to conventional CNNs, ViTs handle image patches as sequences (Figure 3.3), allowing each patch to pay attention to all other patches using self-attention mechanisms, enabling the model to simultaneously capture long-range dependencies throughout the whole image, instead of iteratively developing features through local receptive fields. In the field of medical image classification, ViTs represent an emerging approach that enhances conventional CNNs methods, with architectures like ViT-Base and DeiT demonstrating strong performance on various medical tasks [34]. For example, Ayana et al. introduced ViTCol, a pure ViT-based model for classifying endoscopic pathological findings for the early identification of colorectal cancer, which delivered exceptional results and surpassed CNN-based approaches by a considerable margin [37].

- **FMs:** Large-scale pre-trained models leveraging self-supervised learning on vast datasets to build universal visual representations. In classification tasks, FMs like DINOv2 or CLIP can be fine-tuned or employed as feature extractors for classifying endoscopic image categorization, providing robust images. They deliver strong performance with minimal labeled medical data by capturing complex anatomical patterns and pathological characteristics under various imaging conditions [7]. These type of models will be dealt with more depth in a future chapter.

3.2.2 Object Detection

Object detection task-driven algorithms combine localization with classification abilities to detect and precisely pinpoint pathological structures within medical images using bounding boxes, providing crucial spatial information and diagnostic classification in a single inference step [30]. Object detection models have transformed medical imaging processes by offering automated secondary observation capabilities that complement human expertise [4]. In particular, in the field of gastroenterology, these systems enable the real-time identification of lesions during live procedures, and clinical studies have found significant improvements in detection rates [33, 38, 39].

Deep learning-based object detection can be composed of 2 types (Figure 3.4):

- **Two-Stage Detectors**: separate object detection into two sequential phases: region proposal generation followed by classification and refinement [4]. Basically, this approach first identifies candidate regions, then applies specialized Neural Networks (NNs) to classify and refine these generated proposals, significantly improving speed and accuracy. A significant framework that follows this paradigm is Faster R-CNN, developed by [40], which merges these networks with additional classification stages: first, generate candidate proposals through learned anchor boxes, then apply second-stage classifiers for final detection predictions. Although computationally demanding, these architectures are particularly effective in medical settings demanding high precision, such as detecting subtle lesions, where accuracy is more than prioritized over speed [4, 39].

- **Single-Stage Detectors**: object detection with regression techniques as single-stage networks, directly identifying bounding box coordinates and class probabilities from image pixels in whole images in a single forward pass, prioritizing computational efficiency over exhaustive region examination [4]. The You Only Look Once (YOLO) [41] framework exemplifies this paradigm by treating detection as a direct regression task, which allows ideal real-time predictions for live clinical applications, and some of the architecture variants outperform the primary models of the main models [42]. Generally, in single-stage detectors, some accuracy compared to two-stage approaches is sacrificed in favor of speed, the main requirement for real-time medical application cases that require instant clinical feedback, such as live endoscopic procedures [4].

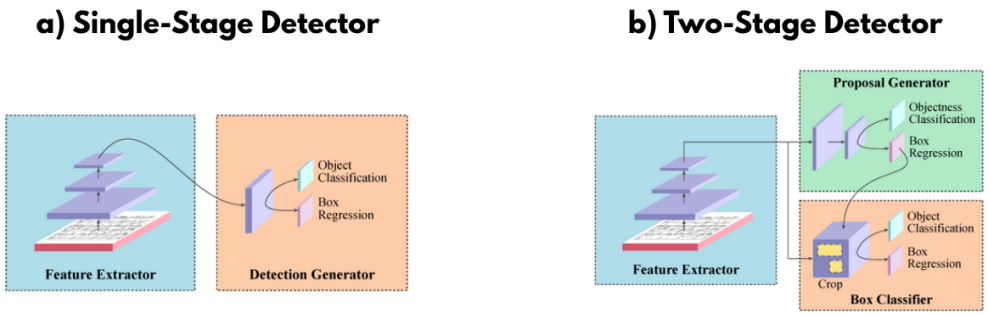


Figure 3.4: Comparison between single-stage and two-stage object detection architectures. (a) Single-stage detectors perform object classification and bounding box regression simultaneously in one forward pass, prioritizing speed. (b) Two-stage detectors separate region proposal generation from classification and refinement, achieving higher accuracy through sequential processing. Adapted from [43].

3.2.3 Image Segmentation

Image segmentation is a highly effective technique for examining medical images, providing detailed outlines of anatomical structures and pathological areas at the pixel level, which is crucial for accurately measuring the boundaries, areas, and shapes of the lesions [4, 30]. Unlike image classification, which assigns image-level labels or detection, which provides boundary boxes, segmentation assigns specific class labels to individual pixels, enabling comprehensive mapping of tissue distribution and accurate assessment of pathological changes. Medical image segmentation has progressed from traditional methods to advanced DL models, achieving clinical grade accuracy and enabling precise disease quantification, treatment response monitoring, and surgical planning [4]. More specifically in GI imaging, segmentation is crucial to measure the extent of the lesion, which aids in therapeutic and research applications [44, 45], as the switch to automated segmentation is part of an ongoing significant change in workflows, delivering performance comparable to that of human experts while providing reliable results in the most diverse range of imaging scenarios.

The primary frameworks associated with this DL methodology are:

- **U-Nets**: encoder-decoder architecture with skip connections for biomedical segmentation, transforming the field by proving that networks could be trained end-to-end using minimal images and surpassing previous sliding-window approaches. This architecture consists of a contracting path to capture context and an oppositely

symmetric expanding path that ensures accurate localization. In addition, the existence of skip connections between the encoder and decoder layers (Figure 3.5) allows the model to maintain spatial information lost in downsampling, further allowing for accurate boundary delineation essential in these medical applications [30, 46].

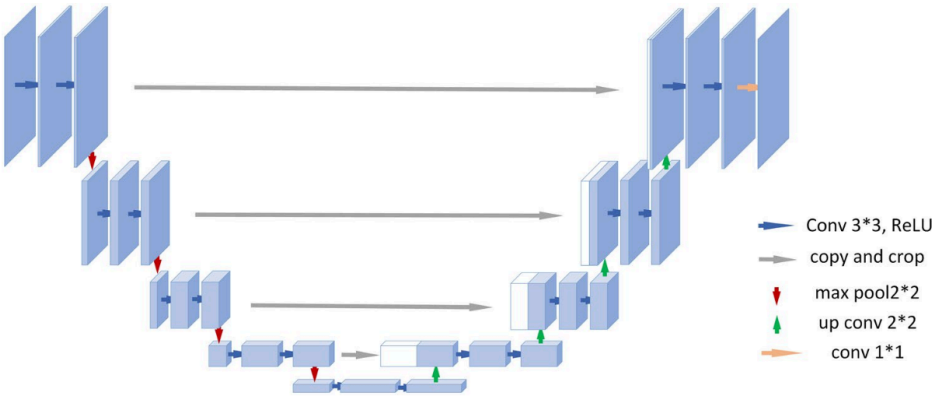


Figure 3.5: U-Net encoder-decoder architecture with skip connections for biomedical image segmentation. The contracting path captures context through progressive downsampling, while the expansive path enables precise localization through upsampling. Skip connections preserve spatial information lost during downsampling, essential for accurate boundary delineation in medical applications. Adapted from [46].

Variants of this architecture, such as UNet++ or Attention U-Net, enhance their performance by introducing dense connections, attention mechanisms, and volumetric data processing (respectively), while maintaining the core encoder-decoder structure [46]. Some applications of these u-net-based variants culminated in huge successes in various GI tasks like GC lesion segmentation or lesion region segmentation [44, 45].

- Architectures like ViTs are a viable option for segmentation tasks, with its self-attention mechanisms being adapted to capture long-range dependencies, but these architectures are way more hybrid with CNNs ideas: pure ViTs lack localization ability for fine spatial details and require extensive computational resources [47]. To address these limitations, variant ViTs strategically combine CNN's local feature extraction with Transformer's global context modeling, culminating in a plethora of capable models for image segmentation [47].

- FMs are also an emerging DL modality in medical image segmentation, but will be explained in more depth in the next subchapter.

3.2.4 Image Generation

As previously observed, CNNs have been widely applied in the medical imaging field, but acquiring high-quality, well-balanced labeled datasets poses challenges due to privacy issues and the considerable time required for labeling. To mitigate these challenges, image generation has evolved from simple data augmentation to sophisticated generative algorithms that produce high-quality synthetic medical images [4]. These images serve to anonymize the data, tackle the shortage of datasets for certain medical conditions, and allow controlled experimental studies. In particular, recent developments in this field have led to synthetic images that exhibit such a high degree of realism that they can even deceive trained annotators and experts, thus improving AI diagnostic training with more balanced datasets [48, 49].

Some of the more noteworthy architectures to follow these paradigms include:

- **Generative Adversarial Networks (GANs)**: the synthesis of data in these models occurs by employing two competing neural networks in adversarial training: a generator network that creates synthetic data and a discriminator network that distinguishes between real and generated content, with the adversarial process driving both networks to iteratively improve through competitive dynamics, where the generator aims to maximize the probability of fooling the discriminator, while the discriminator enhances its detection capabilities [4, 29, 30]. Innovative solutions like Medigan, created by Osuala et al., offer a wide range of pre-trained GANs models in the most diverse medical imaging modalities, empowering users to generate synthetic medical images effortlessly with a single line of code, overcoming the challenges associated with complex training. It makes accessible the availability of high-quality synthetic datasets that are indistinguishable from expert-validated output, allowing the controllable generation of images with specific attributes, such as lesion textures and anatomical shapes, with the focus of customized specialized research applications [48].

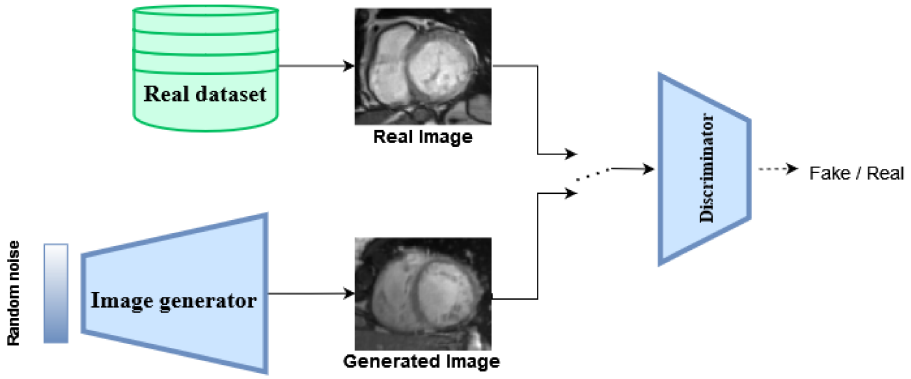


Figure 3.6: Generative Adversarial Network (GAN) architecture showing the adversarial training process between generator and discriminator networks. Real medical images and synthetic generated images are evaluated by the discriminator, which provides feedback to improve the generator's ability to create clinically realistic synthetic medical data for data augmentation and privacy preservation. Adapted from [50].

- **Diffusion Models**: generate high-quality synthetic data through a dual-phase mechanism of forward diffusion (noise addition) and reverse diffusion (denoising) processes, where forward diffusion gradually adds Gaussian noise to clean data until it becomes pure noise, while reverse diffusion learns to progressively remove noise to recover the original data [29, 49]. In medical imaging, diffusion models match GANs efficacy and training efficiency: fine-tuned with minimal data, they produce quality outputs on par with GANs yet require significantly less time [49].

3.3. Foundation Models

Foundation Models are part of an ongoing shift in medical imaging by learning universal visual representations from extensive and varied datasets, separating them from traditional DL: these models undergo large-scale pretraining to develop adaptable and general-purpose capabilities, which can then be fine-tuned for specific clinical needs with minimal additional training [7].

Key Principles and Characteristics: Foundation Models differ from traditional DL in three main ways: scale, generalizability, and adaptability. As conventional medical imaging techniques need a lot of task-specific training on labeled datasets, often struggling with limited data and transferability across clinical situations, FMs in the other hand are pre-trained on large and varied imaging datasets, learning complex visual representations through methods that don't rely much on manual labeling [7]. The core aspects of FMs (Figure 3.7) include:

1. **Pre-training**: FMs use vast datasets with millions of images from different modalities, such as ImageNet [31], to learn universal visual features without needing many labeled samples, addressing data scarcity in medical imaging by making good use of unlabeled data [7].

2. **Transfer learning**: models that have been pre-trained will have their learned representations adjusted for new clinical tasks, reducing significantly the computational load and time needed for training [7].

3. **Fine-tuning**: every FMs are precisely tailored to unique medical tasks using smaller datasets, while still benefiting from foreign knowledge gathered during pre-training [7].

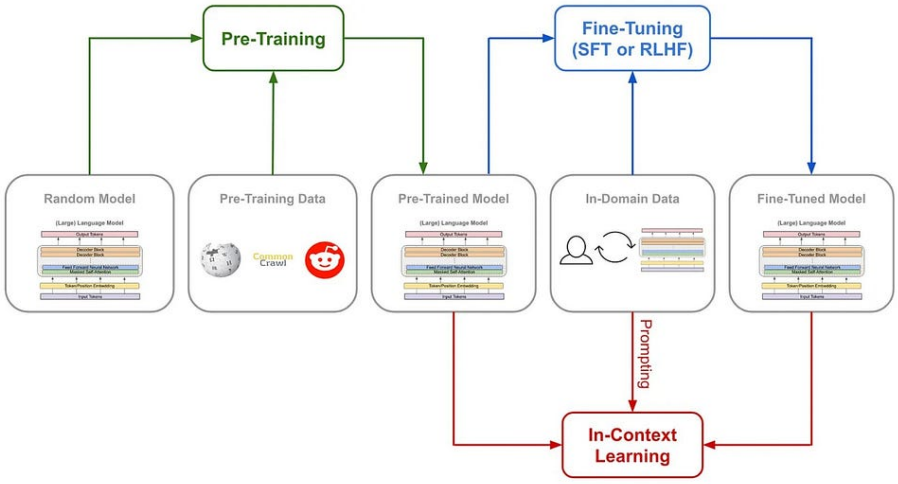


Figure 3.7: Foundation model development pipeline illustrating the paradigm shift from traditional task-specific training to large-scale pre-training followed by fine-tuning. The process involves pre-training on massive diverse datasets, creating versatile FMs, and subsequent fine-tuning with domain-specific data for specialized imaging applications. Adapted from [51].

Foundation Models also exhibit several distinctive characteristics, architecturally or functionally speaking, that distinguish them from other traditional ML approaches [52], which include:

- **Scale**: fundamental characteristic for enabling FMs full capabilities, consisting of three critical components [52]:

  - hardware improvements (increased GPU throughput and memory capacity);
  - transformer architectures (efficient parallel processing of massive datasets);
  - unprecedented data availability (in the houses of the millions of diverse samples).

- **Emergence**: model behaviors emerge naturally instead of being explicitly programmed, resulting in abilities that cannot be directly traceable to deliberate training mechanisms [52].
- **Homogeneization**: diverse applications can use single generic learning algorithms, reducing the need for specialized architectures in different tasks [52].
- **(Multi)modality**: allows FMs to process and integrate multiple data types simultaneously (such as image data with metadata or image context), overcoming the constraints imposed by focusing on a single modality as usual in traditional methods [52].
- **Incompleteness**: FMs act as blank canvas that need customization for specific tasks: they aren't complete solutions on their own, thus being meant to serve as adaptable bases for specific applications [52].

Despite their broad potential, FMs pose many challenges, including the following:

- **Computational requirements** require substantial hardware resources for both training and deployment, limiting accessibility in resource-constrained clinical environments [5, 52];
- **Data bias and quality** might present some concerns, as models trained on imbalanced datasets could be skewed by existing healthcare disparities or perform poorly in diverse patient populations [5];
- **Interpretability and explainability** still remain fundamental barriers to clinical adoption, as the black-box nature of these models conflicts with the medical decision-making requirements for transparency and accountability [52].
- **Regulatory compliance and validation processes** for medical FMs is still evolving, creating uncertainty for clinical implementation [5].

In summary, even though these models aren't in no way perfect or polished and present a number of limitations, they're still eclipsed by the various significant benefits, like superior generalizability across different imaging modalities, reduced dependence on labeled clinical data, improved performance in rare diseases where data are naturally scarce, and the capability to function as universal feature extractors for a variety of downstream tasks. These features make FMs especially useful in the field of medical imaging, where data acquisition is costly, time-consuming, and frequently restricted by privacy issues [7].

3.4. Foundation Models Main Architectures

Foundation Models in medical imaging can be divided into three primary architectural approaches, each designed to address specific requirements and data characteristics:

3.4.1 Vision Foundation Models

Vision FMs are mainly dedicated to developing visual representation knowledge. Models like DINOv2 follow this methodology by learning universal visual representations through self-supervised pre-training on diverse medical imaging datasets. These models leverage transformer-based frameworks with attention mechanisms in order to capture long-range dependencies and complex visual patterns across different anatomical structures (Figure 3.8). In particular, they are specially designed to detect subtle visual indicators that are not easily seen by conventional methods, improving their effectiveness in identifying early pathological changes in medical images, for example [5].

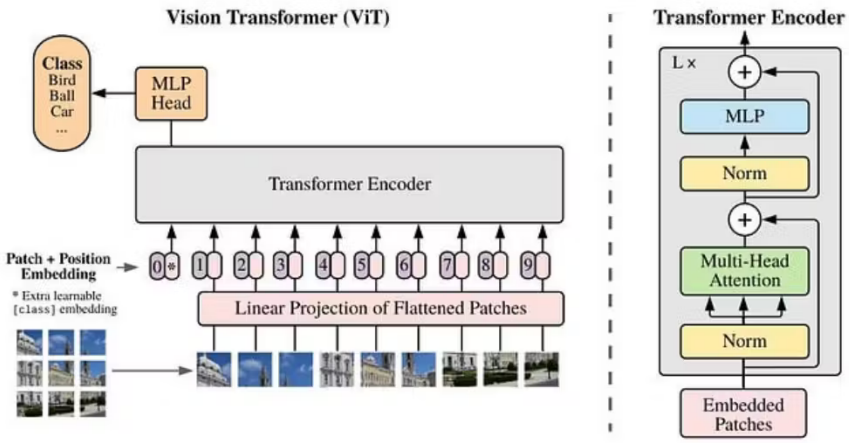


Figure 3.8: Vision Foundation Model architecture exemplified by DINOv2's self-supervised learning framework. The model employs teacher-student distillation with momentum updates during pre-training, learning robust visual representations without requiring labeled data, enabling efficient adaptation to downstream medical imaging tasks through minimal fine-tuning. Adapted from [53].

3.4.2 Vision-Language Foundation Models

Vision-Language FMs merge visual and textual information, creating multimodal representations that combine imaging data with clinical reports and medical expertise: models based on the CLIP architecture (Contractive Language-Image Pretraining) represent this category by learning combined representations of medical images alongside their textual descriptions (Figure 3.9). These models facilitate advanced applications such as automated report generation, visual question answering, and cross-modal retrieval within clinical workflows. The integration of the textual context offers interpretability, vital for clinical implementation, enabling clinicians to understand model decisions through explanations in natural language [5].

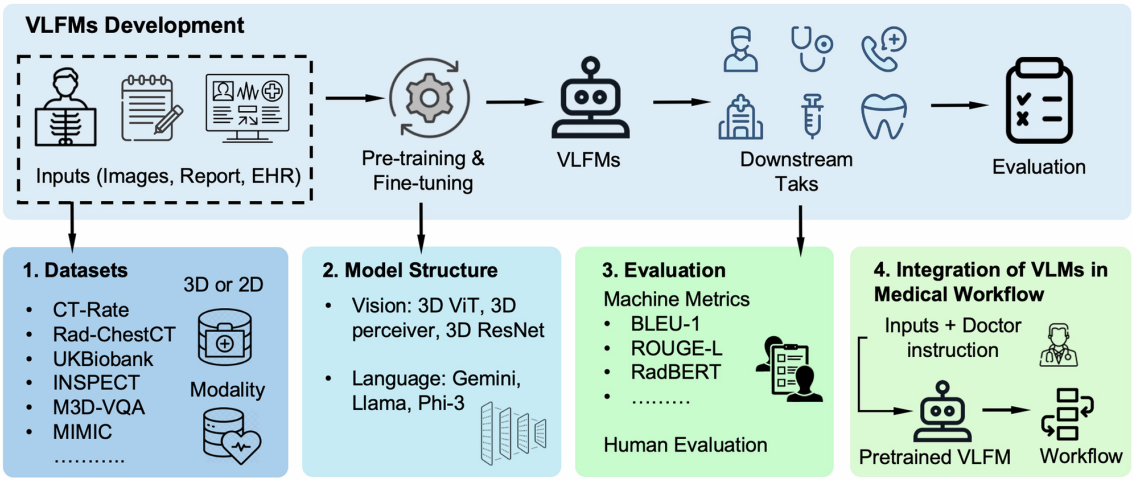


Figure 3.9: Vision-Language Foundation Model (VLFM) development workflow showing the integration of multimodal medical data. The framework combines imaging data with clinical reports and electronic health records, enabling applications such as automated report generation, visual question answering, and cross-modal retrieval in clinical workflows through comprehensive evaluation metrics and workflow integration. Adapted from [54].

3.4.3 Medical-Specific Foundation Models

Medical-Specific FMs represent specialized architectures designed explicitly for medical applications. MedSAM exemplifies this approach, serving as a universal medical image segmentation model. Trained in more than 1.5 million pairs of image masks in 10 imaging modalities, MedSAM outperforms general-purpose segmentation models by markedly improving the segmentation of medical structures that feature unclear boundaries and low-contrast areas typically found in clinical imaging. In addition, SAM-Med variants further extend these capabilities to particular medical domains while preserving their general applicability [55].

3.5. Potential of Foundation Models in the Gastrointestinal Field

Foundation Models continues to increase in addressing key issues within the field of GI endoscopy, such as the scarcity of annotated data sets, the high variability between observers in interpreting the injury and the need for real-time analysis during procedures. The complex visual patterns present in endoscopic images, which range from normal anatomical differences to subtle pathological alterations related to precancerous conditions, make GI imaging a prime growth area for the application of FMs [7].

These models have multiple clinical applications, especially in the realm of medical imaging [5, 7]:

- In polyp detection and segmentation, FMs excels in identifying subtle morphological characteristics that differentiate polyps from normal tissue;
- Automated navigation support during endoscopic procedures is aided by the recognition of anatomical landmarks detected by FMs;
- Assessing the severity of inflammatory bowel disease is enhanced by FMs pattern recognition, which helps in the standardization of scoring systems for various conditions;
- The generation of automated reports is accomplished by integrating vision-language FMs with endoscopic observations to create structured clinical documentation.

Successful examples in the application of FMs within real medical imaging are found in files like the work of Kerdegari et al., having explored the effectiveness of FMs in the analysis of gastric inflammation. By fine-tuning DINOv2 for the classification of gastric intestinal metaplasia per the EGGIM protocol, they achieved superior results compared to traditional CNN methods, all of this achieved with minimal task-specific fine-tuning required [7]. Another notable case, the EndoDINO framework from Dermyer et al., stands out as an advanced foundation model specifically for gastrointestinal segmentation tasks. The model was pre-trained on the largest dataset of gastrointestinal endoscopy videos available, with model sizes ranging from 86 million to 1 billion parameters. For tasks such as anatomical landmark classification, polyp segmentation, and Mayo endoscopic scoring in the evaluation of ulcerative colitis, EndoDINO sets new performance benchmarks, serving as a good portrait of the scalability and clinical practicality of FMs [8].

In synthesis, Foundation Models represent a paradigm shift in medical imaging through large-scale pretraining and efficient clinical adaptation, with impactful success stories demonstrating their broad transformative and problem-solving potential [7, 8]. The shift from traditional Deep Learning techniques to FMs represents a major change in medical imaging: DL set the groundwork for automated analysis, while FMs offer broader applicability and require fewer labeled data due to extensive preliminary training. Together, these technologies equip us to build advanced diagnostic systems, aiming for better patient care through more reliable medical image analysis.

4. Segmentation Algorithms in GI Imaging: A Systematic Review

This chapter presents a comprehensive systematic review of segmentation algorithms applied to GI imaging, representing a state-of-the-art literature research focused specifically on current trends in AI for 2D gastric endoscopy image segmentation. The work presented here was initially developed as a literature research as part of this dissertation, with the intention of getting a deeper understanding of the current landscape of AI use in the field of Gastrointestinal medical imaging and the methodological approaches in the area, but given the quality of the conclusions collected during this study, it ended up being published as a contribution for the 8th IEEE Portuguese Meeting on Bioengineering (ENBENG 2025) conference:

Ferreira, Tomás Marques, David Dinis-Ribeiro, Mário Coimbra, Miguel. (2025). A Review of Segmentation Algorithms for Gastrointestinal Images: Datasets, Annotations and Architectures. 5-8. 10.1109/ENBENG67130.2025.11199537. [56]

As observed in previous chapters, Gastrointestinal image segmentation plays a vital role in improving diagnostic accuracy, particularly for early detection and treatment planning of diseases such as colorectal and gastric cancers. With the growing interest in computer-aided diagnostic tools, accurate and reliable segmentation algorithms are essential for clinical integration, especially in endoscopic workflows. Despite rapid advancements in deep learning, there remains a pressing need to systematically consolidate and evaluate the breadth of segmentation techniques applied to GI imaging.

4.1. Review Methodology and Results

4.1.1 Literature Search Strategy

The process of constructing this systematic review began with the formulation of a comprehensive search query designed to retrieve the recent and relevant scientific literature on segmentation algorithms applied to medical imaging of GI. The objective was to capture publications aligned with the domains of medical image analysis, particularly in relation to cancers of the digestive tract, while ensuring coverage of modern techniques from computer science and engineering disciplines.

Three widely recognized scholarly databases were selected: PubMed, Scopus, and IEEE Xplore. The final query was iteratively optimized to balance sensitivity and specificity and was structured as follows.

> *Segmentation AND medical AND image AND cancer AND (colon OR intestine OR gastric OR endoscopy OR 2D)*
> *(Publication Date: 2022–2024 | Filters: Computer Science, Engineering, Medicine)*

This search resulted in a total of 537 PubMed records, 353 Scopus records, and 180 IEEE Xplore records, which yielded 919 unique documents after deduplication. The deduplication and review process was facilitated using Rayyan, an AI-assisted platform tailored for systematic review screening and conflict resolution.

4.1.2 Inclusion and Exclusion Criteria

To refine the dataset, inclusion and exclusion criteria were defined a priori and applied in two filtering stages: title-based and abstract-based screening.

Inclusion Criteria:

- Keywords present: Segmentation, Medical Image, Gastric, Intestinal, Cancer, 2D, Endoscopy
- Focus on algorithmic or architectural contributions to image segmentation
- Studies involving human GI imaging modalities

Exclusion Criteria:

- Topics focused on Detection, Prediction, Comparison, Rectal (non- GI scope), Tumor biology without image analysis, 3D modalities, Review articles, or non-research documents (ex: practice guidelines)
- Papers focused on non-GI cancers or non-relevant imaging technologies (ex: White Light Interferometry)
- Non-English publications or inaccessible full texts

4.1.3 Screening Procedure

The review used a dual-stage filtering protocol, each independently conducted by two authors, with conflict resolution through consensus meetings.

Title-Based Screening

The documents were assessed on the basis of the relevance of their titles.

- Approved by Author 1: 287 (31.2%)
- Approved by Author 2: 196 (21.3%)
- Overlap (approved by both): 180 (19.6%)
- Conflicting decisions: 123 (13.4%)

After resolving conflicts, 27 additional documents were accepted, resulting in 207 titles advancing to the next stage.

Abstract-Based Screening

A total of 194 documents were eligible after removing 13 non-English or unavailable entries. The abstracts were then screened using the same inclusion/exclusion criteria:

- Approved by Author 1: 53 (27.3%)
- Approved by Author 2: 33 (17.0%)
- Overlap: 20 (10.3%)
- Conflicts: 46 (23.7%)

Eight additional documents were accepted after conflict resolution, bringing the total to 28 documents.

Final Eligibility Check

A full text review ensured alignment with the scope of the review. Eight studies were excluded at this stage due to the following:

- Use of incompatible imaging modalities
- Lack of focus on segmentation or on relevant GI cancers

The final corpus included 20 peer-reviewed publications [33, 38, 39, 44, 45, 48, 57–70], which form the evidence base for the subsequent review and comparative analysis. A summary of the whole process can be inspected in the PRISMA flow diagram in Figure 4.1.

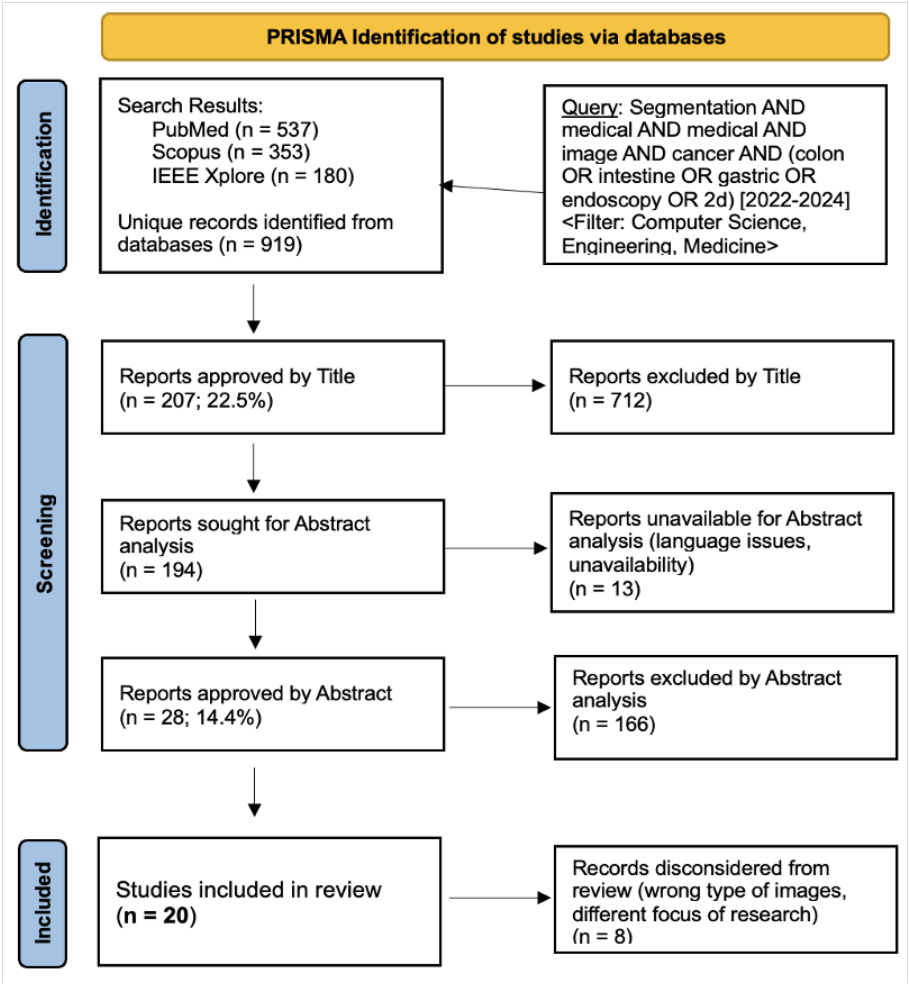


Figure 4.1: PRISMA flow diagram detailing the identification, screening, and inclusion of studies for this review.

4.2. Dataset and Annotation Protocols

4.2.1 Dataset Characteristics

The analysis of the size of the dataset in the reviewed papers reveals a significant concentration of studies utilizing relatively small image collections. As shown in Figure 4.2, most of the data sets contained fewer than 3000 images, the most common range being $\leq$ 1,000 images. This highlights a prevalent limitation in current GI segmentation research, where access to large-scale annotated image repositories remains scarce. Despite this, some outliers reported datasets with more than 20,000 images, pushing the maximum to

28,939 images, although the median dataset size was just 1,252 images and the mean was approximately 3,877, indicating a skewed distribution.

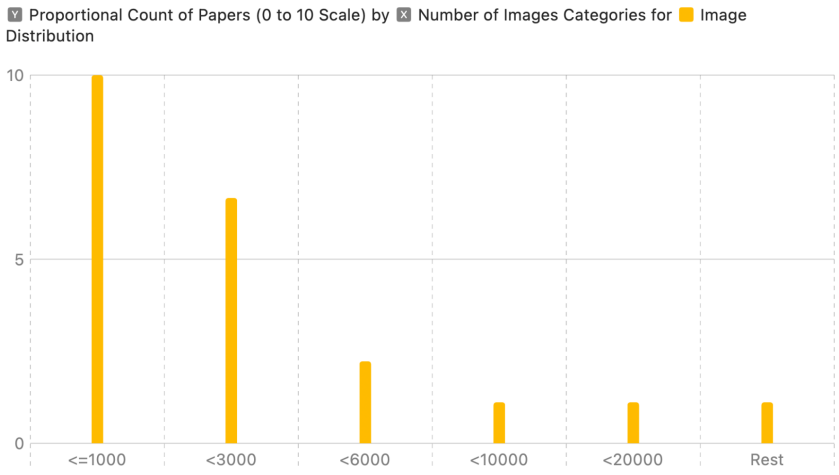


Figure 4.2: Distribution of dataset sizes used across the reviewed studies.

In terms of clinical segmentation targets (Figure 4.3), lesion region segmentation was the most common focus (n=7) [38, 39, 44, 45, 59, 67, 68], followed by Early Gastric Cancer segmentation (n=5) [33, 38, 39, 57, 60], and polyp segmentation (n=4) [48, 63–65]. A smaller subset addressed the boundaries of the GI tract (n=2) [63, 66], metaplasia (n=2) [61, 69], or other types of cancer (n=3) [58, 62, 63]. These results suggest a growing attention to high-impact lesion detection tasks, especially for early-stage disease, but also reflect the heterogeneity in segmentation objectives, posing challenges for standardizing datasets or benchmarking algorithms.

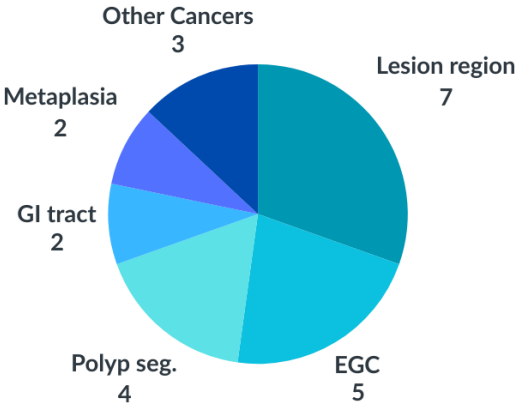


Figure 4.3: Primary segmentation targets in the selected studies.

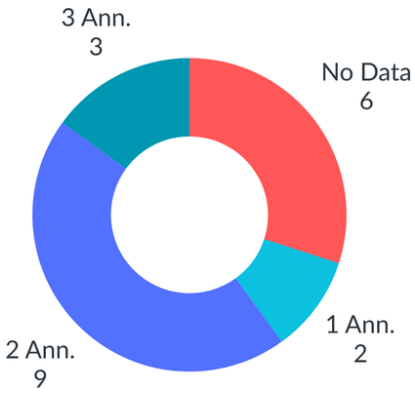


Figure 4.4: Reported number of annotators per study.

4.2.2 Annotation Practices

Regarding dataset annotation practices, a substantial proportion of studies (14 of 20) explicitly reported involving human annotators in the labeling process (Figure 4.4). However, transparency on the number of annotators remained inconsistent. Among those who provided information, the most common configuration was two annotators (9 articles) [38, 57, 58, 61, 63–65, 67, 68], followed by three annotators (3 papers) [39, 60, 62], and only two papers employed a single annotator [33, 44].

In particular, six studies [45, 48, 59, 66, 69, 70] did not provide information on annotation personnel, raising concerns about the quality of the label, the variability between observers, and the reliability of the data set. These findings highlight the pressing need for clearer reporting standards in annotation protocols, especially in clinical image segmentation tasks where expert consensus is crucial.

4.2.3 Dataset Availability and Transparency

The availability and transparency of datasets emerged as a critical factor in the reviewed studies. As shown in Figure 4.5, there was an even split between public and private datasets, with 10 papers relying on openly accessible datasets and 10 utilizing private or institution-specific collections. Among the public datasets, Kvasir-SEG [71] was the most widely used (9 articles) [44, 48, 57, 58, 62–65, 67], followed by CVC-ClinicDB [72] (5 papers) [44, 62–65] and ETIS-Larib [73] (2 papers) [62, 63], indicating a strong dependence on a small subset of well-established GI imaging repositories. This limited diversity underscores a broader challenge in the field: the need for expanded public datasets to support generalizability and reproducibility in segmentation research.

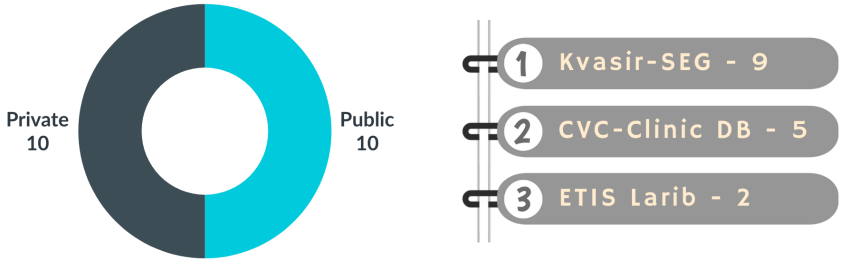


Figure 4.5: Proportion of studies using public versus private datasets. An even split was observed, with Kvasir-SEG [71] and CVC-ClinicDB [72] among the most frequently used public repositories.

4.3. Segmentation Architectures and Trends

An analysis of the architectural frameworks adopted in the reviewed studies reveals a strong trend towards task-specific customization rather than strict adherence to canonical models (Figure 4.6). The most common approach involved the use of custom designed models (n=6) [58, 60, 63, 64, 69, 70], often blending elements from multiple standard architectures to better suit the unique challenges posed by GI imaging data, namely subtle lesion boundaries, varied lighting, or limited dataset sizes.

Among established models, U-Net and its derivatives remained the most popular backbone architecture (n=4) [44, 45, 59, 66], continuing its prominence in medical image segmentation due to its efficient encoder-decoder structure and strong performance in small datasets. The mask variants R-CNN followed (n=3) [38, 39, 57], often enhanced with modules such as bidirectional feature fusion or region discrimination to improve detection in gastroscopic imaging.

More novel methods such as Transformers (n=2) [62, 65] have begun to emerge, signaling a growing interest in the use of generative or attention-based mechanisms, although their adoption remains limited, probably due to data requirements and training complexity. CNN-based models (n=2) [33, 61] and FCN-based approaches (n=1) [68] played more foundational roles, either in baseline tasks or early-stage feature extraction, while still contributing meaningfully to detection and classification tasks.

Ensemble methods [67] and GANs [48] were the least utilized (n=1 each), suggesting that while theoretically promising, either their complexity may hinder widespread adoption as a current research trend, or there could still exist training instability or domain specificity associated with architecture implementation. As a final comment, only two of the 20 articles inspected provided a easily available and reproducible source code [44, 48].

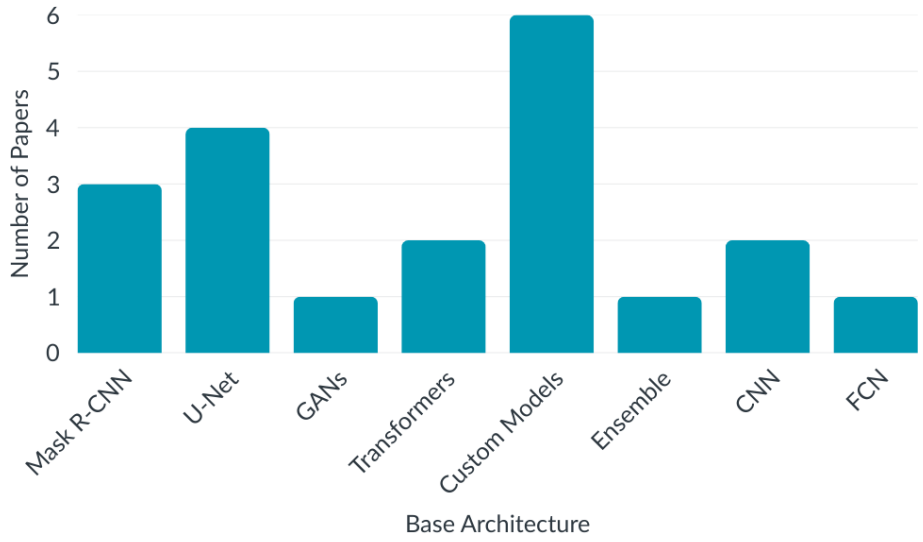


Figure 4.6: Types of deep learning architectures employed across the reviewed papers.

4.4. Discussion

4.4.1 Discussion of the Results

This systematic review reveals that while the segmentation of GI images is gaining traction, the field remains constrained by several challenges. Most studies rely on relatively small datasets, often less than 3,000 images, with a strong dependency on a handful of public datasets such as Kvasir-SEG [71] and CVC-ClinicDB [72]. This concentration of research around limited datasets raises concerns about the generalizability of proposed methods and the potential for overfitting to specific dataset characteristics.

Annotation practices lack consistency, with limited reporting on the number and expertise of annotators raising concerns about label quality and reproducibility. The finding that six studies provided no information on annotation personnel is particularly concerning, as inter-observer variability is a well-known challenge in medical image analysis, and the quality of ground-truth annotations directly impacts model performance and clinical applicability.

Methodologically, there is a trend towards task-specific, custom-built architectures, reflecting innovation but limiting standardization and comparability. Although this approach may deliver optimal performance for specific datasets or clinical contexts, it obstructs the development of standardized benchmarking protocols and makes it difficult to assess the true progress of the field, making it even worse with the fact that only two out of the 20 articles inspected conducted an available path to look at their architecture's source code.

The limited adoption of newer paradigms, such as transformers and ensembles, suggests that complexity and data demands remain barriers to adoption.

The predominance of U-Net architectures, while understandable given their proven effectiveness in medical image segmentation, also suggests that the field may benefit from a more systematic exploration of alternative approaches. Similarly, the low adoption of transformer-based methods could indicate either insufficient adaptation of these architectures to medical imaging specifics or the need for larger datasets to realize their potential. Overall, these findings underscore the need for larger, more diverse datasets, transparent annotation protocols, and more unified benchmarking efforts to advance reproducible and clinically relevant GI segmentation research. This is further reinforced by methodological proposals such as QUAIDE by Antonelli et al., which aim to create standards for algorithmic results of preclinical trials in this field [74].

4.4.2 Potential of Foundation Models in Medical Imaging

The findings of this systematic review highlight several areas where Foundation Models could address current limitations and advance the state of GI image segmentation: the challenges identified (small datasets, inconsistent annotation practices, and limited architectural diversity) align precisely with the strengths that FMs can bring to medical imaging applications.

Tackling Data Scarcity: The predominance of small datasets (median size of 1,252 images) in current GI segmentation research represents a significant limitation that FMs are uniquely positioned to address. Through large-scale pre-training on diverse medical imaging datasets, FMs are able to learn robust visual representations that transfer effectively to GI-specific tasks with minimal fine-tuning of the data. This capability could substantially reduce the dependency on large institution-specific datasets that are often difficult to obtain due to privacy concerns and annotation costs.

Standardization and Reproducibility: The trend towards custom built architectures, while innovative, creates challenges for standardization and comparison between studies. FMs offer a pathway toward more standardized approaches, providing a common pre-trained backbone that can be fine-tuned for specific GI segmentation tasks, further improving reproducibility and enabling more meaningful comparisons between different research efforts.

Cross-Modal Learning Opportunities: Vision-language FMs can possibly enhance the transparency in current methodologies by integrating endoscopic images with clinical reports and metadata. This multimodal approach has the potential to increase interpretability, a critical requirement for clinical adoption, and improve segmentation performance by incorporating contextual clinical information.

Advancing Context-Specific Cases: Focusing on common segmentation targets (lesion regions, polyps) identified in this review suggests limited exploration of specific conditions datasets: the strong generalization capabilities of FMs could enable more robust performance in these uncommon cases.

Added to this, the emerging success of medical-specific Foundation Models like MedSAM [55] or EndoDINO [8] provides promising evidence for this potential. These models demonstrate that FMs can achieve state-of-the-art performance in medical segmentation tasks while requiring minimal task-specific fine-tuning, suggesting a way forward to address many of the limitations identified in current GI segmentation research.

However, the adoption of FMs in GI imaging also presents challenges that must be addressed, including computational requirements, the need for diverse large-scale training data and the assurance of robust performance in different clinical settings and patient populations. Future research should focus on developing efficient fine-tuning strategies, creating comprehensive benchmarking frameworks, and establishing validation protocols that ensure clinical safety and reliability.

5. Foundation Models for Endoscopy Imaging

This chapter encapsulates a comprehensive exploration of the use of Deep Learning structures for the automated classification of Gastric Cancer within the Endoscopic Grading of Gastric Intestinal Metaplasia (EGGIM) framework. The research aims to explore the feasibility of deploying Foundation Models in endoscopic medical imaging applications by systematically assessing three different architectural paradigms, utilizing both pre-trained and fine-tuned models, including:

- **ResNet-50** (traditional Convolutional Neural Network (CNN));
- **ConvNeXt** (modernized CNN with Vision Transformers (ViTs));
- **DINOv2** (Foundation Model (FM) with ViTs base).

5.1. Motivation and Objectives

5.1.1 Clinical Motivation and the limits of a Foundation Model

Building on the detailed analysis of gastric cancer biology and endoscopic imaging from previous chapters, this research tackles key limitations in the existing EGGIM protocol that directly affect patient outcomes and workflow efficiency. As highlighted in the systematic review of GI segmentation algorithms, this field struggles with issues such as small dataset sizes, varying annotation standards, and overreliance on specialized architectures that limit generalizability in clinical settings.

The Foundation Model Paradigm: Recent advances in computer vision FMs mark a shift from specialized DL methods to universal feature extraction that can be adaptable to medical applications with minimal training data. Unlike traditional methods that require large labeled datasets, Foundation Models use extensive pre-training to develop robust visual representations transferable across various imaging modalities. However, the exact definition and boundaries and what defines exactly ” Foundation Models” is still far from a consensus [7], especially in medical imaging where standard pre-trained models may function like FMs in specific applications, which requires careful analysis of their architecture and training methods to assess their impact on gastric cancer detection.

A nuanced perspective: Regarding this paradigm, in this research, I opted to take a subtle view of model categorization, recognizing the spectrum of pretraining paradigms rather than rigid classifications. Considering this perspective, DINOv2 is a 'true' representation of FMs, using self-supervised learning on 142 million images, while ResNet-50 and ConvNeXt use supervised pre-training on ImageNet datasets, providing transferable representations for medical imaging. This perspective aligns with the broader view that FMs are not discrete types but lie on a continuum of scale, training methods, and adaptability, which is clinically important for understanding how different pre-training approaches, such as ImageNet training and extensive self-supervised learning, impact diagnostic performance in medical applications.

5.1.2 Research Objectives and Technical Contributions

The goal is to evaluate the possibility of use of FMs on the medical imaging field by comparing the base models of a traditional CNN, a modernized CNN with ViTs and an actual FMs with a ViTs base, across pre-trained and fine-tuned learning configurations in a three-class unbalanced small precancerous gastric lesions dataset for EGGIM classification.

This research focuses on advancing endoscopic imaging technology by evaluating and comparing different architectural paradigms, such as traditional CNNs (ResNet-50), transformer-inspired CNNs (ConvNeXt), and the Foundation Models counter part: self-supervised ViTs (DINOv2). Through this evaluation, one of the main goals of this study pass by identifying the model's benefits in handling severe class imbalance prevalent in medical applications, with innovative training methods (like progressive unfreezing and discriminative learning rates) being explored to enhance model performance under these medical imaging constraints. In addition, a robust evaluation framework is established, incorporating techniques such as test-time augmentation and ensemble methods, to ensure clinical deployment readiness, with this research also offering a drag path into the usual computational needs and common challenges of deployment of these heavy models in clinical environments as such with limited resources.

Regarding a point of view more oriented towards possible contributions in this field of computer vision applied to endoscopic imaging, from my research, this appears to be the first systematic comparison of CNNs versus FMs/ ViTs for EGGIM classification, as well as the subsequent development of optimized training protocols, comprehensive evaluations

of FMs for gastric lesions detection, and the establishment of reproducible benchmarks that highlight ongoing challenges which need further investigation.

In genesis, both efforts address the significant gaps identified in GI segmentation studies, emphasizing the need for standardized evaluation methods and exploring Foundation Models potentials where traditional methods still tend to struggle due to limitations such as data and nature of the models themselves.

5.2. Materials

5.2.1 EGGIM Dataset Characteristics and Clinical Context

Dataset Composition and Clinical Acquisition: The “2025-01_EGGIM_Dataset3” dataset represents a comprehensive collection of high-resolution NBI endoscopic frames acquired under rigorous clinical protocols at Instituto Português de Oncologia - Porto (IPO-Porto), following established EGGIM guidelines for gastric cancer risk stratification. The dataset encompasses two complementary acquisition phases designed to maximize clinical representativeness while ensuring technical consistency:

- Dataset-A (n=414) was retrospectively collected from routine clinical practice between December 2019 and September 2020, representing real-world clinical diversity with patients included irrespective of clinical indications provided they met stringent image quality requirements;
- Dataset-B (n=866) was prospectively collected in 2024 under controlled high-quality acquisition protocols, representing 80 consecutive patients undergoing upper GI endoscopy.

In terms of clinical expertise, all endoscopic procedures were performed by gastroenterologists with extensive experience (>5 years, >500 endoscopies annually) using standard Olympus high-definition endoscopes (185° or 190° series) equipped with NBI technology. The images were acquired in JPG format at 640x480 pixel resolution with 24-bit color depth, ensuring consistent technical parameters throughout the dataset. All data were subjected to a rigorous anonymization process to ensure the protection of patient confidentiality in accordance with ethical standards.

Three-Class Classification Framework: The dataset implements a clinically relevant three-class classification system that directly maps to EGGIM risk stratification requirements while addressing the class imbalance challenges prevalent in medical imaging applications:

- **Class 0 (Normal Mucosa)**: 0% of visible IM; n=359 in training set, distribution of approximately 43.7%;
- **Class 1 (Growing Metaplasia)**: up to 30% of visible IM; n=158 in training set, distribution of approximately 19.2%;
- **Class 2 (Advanced Metaplasia)**: more than 30% of visible IM; n=303 in training set, distribution of approximately 36.9%.

This class distribution reflects natural epidemiological patterns of gastric lesion progression while presenting significant machine learning challenges that mirror real-world clinical deployment scenarios. In particular, the severe underrepresentation of cases of Class 1 (growing metaplasia), which is the diagnostic scenario with the greatest clinical impact, creates an optimal testing environment to evaluate the advantages of FMs in the detection of these types of rare medical conditions.

In this work, the dataset is divided into 64%, 16%, and 20% for training, validation, and testing, respectively.

5.2.2 Computational Infrastructure and Development Environments

1. Local Development Environment (macOS Apple Silicon): The development was carried out on a MacBook Pro M4 Pro equipped with 24-32 GB of memory, facilitating rapid prototyping essential for algorithm development and visualization. This setup was crucial for tasks like designing model architectures, investigating hyperparameters, quick debugging, and generating high-quality visualizations such as t-SNE, confusion matrices, and calibration curves. Versioning of the code and reproducible experiments were maintained through a private GitHub repository.

```
fm_class/
| fine-*-training.py  # Architecture-specific (*) training
| fine-*-testing.py   # Architecture-specific (*) testing
| pre-training.py     # Unified pre-training pipeline for all models
| pre-testing.py      # Unified pre-testing pipeline for all models
| models/
|   | dinov2.py           # DINOv2 base implementation (FM)
|   | convnext.py         # Modernized CNN base architecture
|   | resnet50.py         # Pre-training ResNet-50 configuration
|   | resnet50FT.py       # Fine-tuning ResNet-50 configuration
| src/
|   | losses.py           # Advanced loss functions
|   | eggim_class.py      # EGGIM dataset implementations
|   | data_loading.py     # Data preprocessing utilities
|   | model_evaluation.py # Evaluation frameworks
| results/
    | prime_results/      # Best model configurations (separated by
                          # learning setups and models subdirectories)
```

2. High-Performance Computing Environment (kempo2 GPU Server): Training and evaluation were performed on the kempo2 server at FCUP/IPO-Porto, accessed via SSH with university credentials and SSH key authentication. The server ran Ubuntu 22.04 LTS and Python 3.8.16, with CUDA 11.8/12.1 and high-end NVIDIA GPUs (A100/RTX A6000), ultimately serving its role of supporting large batch processing, hyperparameter exploration, model weights training, and ensemble evaluation.

Additional rigorous reproducibility measures were imposed, including deterministic algorithm settings (`torch.use_deterministic_algorithms(True)`), fixed random seeds (`torch.manual_seed(42)`), and consistent evaluation protocols across all architectural paradigms. Mixed precision training was selectively applied only on GPU servers to avoid Moving Particle Semi-implicit (MPS) numerical instability, while backbone architectures remained in evaluation mode during fine-tuning phases per established medical imaging best practices.

5.2.3 Software Dependencies and Implementation Libraries

The computational requirements for this project span both general-purpose DL frameworks and domain-specific medical imaging utilities, all managed through Python 3.8.16 for compatibility with server infrastructure. The key dependencies and their roles are summarized below and reflected in the project's `requirements.txt` file (Figure 5.1), with core libraries standardized in versions that balance performance and reproducibility.

requirements.txt

```
torch==2.3.0
torchvision==0.18.0
transformers==4.41.1
huggingface-hub==0.23.3
opencv-python==4.9.0.80
numpy==1.26.4
scikit-learn==1.4.2
Pillow==10.3.0
tqdm==4.66.4
pandas
matplotlib
```

Figure 5.1: List of specific versions of Python dependencies to be installed

Deep Learning Framework Stack: PyTorch 2.0.0-2.3.0 serves as the primary framework, selected for its robust GPU acceleration and adaptability with the HuggingFace interface. TorchVision 0.15.0-0.18.0 provides standard computer vision operations (transforms, pre-trained models, data loaders). Model checkpoints leverage Hugging Face Transformers 4.41.1, which manages pre-trained weights for ResNet-50, ConvNeXt-base, and DINOv2-base through `AutoModel` APIs, with offline mode enforced via the environment variable `TRANSFORMERS_OFFLINE=1` and `local_files_only=True`, in order to prevent unintended network calls during training.

Model Loading and Pre-trained Checkpoints: All backbone architectures are initialized from local directories (`./models/resnet-50`, `./models/convnext-base-224`, `./models/dinov2-base`) containing HuggingFace-compatible `config.json` and `pytorch_model.bin` files. The `AutoImageProcessor` class is responsible for managing normalization parameters specific to the architecture (ImageNet means and standard deviations), to ensure that input preprocessing aligns with pre-training protocols. Custom model wrappers (`models/resnet50.py`, `models/resnet50FT.py`, `models/convnext.py`,

`models/dinov2.py`) extend base classes with medical imaging-specific modifications, like progressive unfreezing (`unfreeze_last_layers()`) and specialized classifier heads.

Data Processing and Augmentation: Image manipulation relies on OpenCV 4.9.0-4.10.0 for low-level operations (like preprocessing) and Pillow 10.3.0 for high-level PIL-based transformations. Augmentation pipelines in `src/data_loading.py` implement progressive strategies: conservative augmentations for pre-training (random crops, orientation flips, grayscale conversion,...) and aggressive transformations for fine-tuning (MixUp/CutMix, random erasing,...).

Loss Functions and Class Imbalance Handling: Advanced custom loss implementations in `src/losses.py` address severe class imbalance, while Scikit-learn 1.4.2-1.5.1 computes class weights via `compute_class_weight()` for weighted sampling and loss initialization.

Evaluation and Visualization: Performance metrics leverage Scikit-learn's `f1_score`, `accuracy_score`, and confusion matrix utilities, with model evaluation scripts in `src/model_evaluation.py`. Dimensionality reduction for t-SNE embeddings (`src/tsne_embed.py`) uses `sklearn.manifold.TSNE` to visualize feature spaces in 2D. Matplotlib 3.9.2 renders publication-quality plots, while Pandas 2.2.2 structures training logs and per-class metrics for export.

Reproducibility Utilities: Deterministic training sets predefined random seeds for Python, NumPy and PyTorch (42). Progress is monitored using TQDM 4.66.4, offering real-time feedback for multi-epoch training runs, with logs serialized to JSON format for further analysis. The Huggingface-hub 0.23.3 handles model registry operations during checkpoint saving, with offline limitations restricting uploads to local storage only.

5.3. Model Architectures and Implementation

This chapter provides an overview of the DL model structures used for classifying the discussed three-class EGGIM dataset. Three main types of architecture were assessed: traditional Convolutional Neural Networks (**ResNet-50**), modern CNNs architectures influenced by ViTs (ConvNeXt), and Vision Transformers-based Foundation Models (**DINOv2**). Each model was trained and evaluated under two learning setups:

- **pre-training**, where only the classification head is trained and the backbone stays unchanged;
- **fine-tuning**, which involves gradually unfreezing backbone layers to allow task-specific adjustments.

These architectures were chosen to showcase different design approaches in similar base model checkpoints, offering a detailed evaluation of classical CNNs methods, hybrid convolutional frameworks and transformer-based strategies for this type of endoscopic imaging task.

5.3.1 ResNet-50: Legacy Deep Residual Learning Network

ResNet-50 is a 50-layer deep residual convolutional neural network introduced by He et al. for large-scale image recognition tasks. This architecture tackles the issue of vanishing gradients by incorporating residual connections, which are shortcut paths that enable gradients to travel directly through the network during backpropagation, and one of the reasons why this model can train very deep networks while keeping optimization stable. In this study, the `microsoft/resnet-50` pre-trained checkpoint was used, providing weights learned from ImageNet-1K supervised training [31, 75, 76]. The ResNet-50 architecture created can be observed in Figure 5.2.

Backbone Structure: This ResNet-50 implemetation backbone consists of an initial stem layer (Stage 1) followed by four sequential residual stages (Stages 2-5). The stem applies a `7x7` convolution with stride 2, reducing the spatial resolution from `224x224` to `112x112`, followed by batch normalization, Rectified Linear Unit (ReLU) activation, and `3x3` max-pooling with stride 2, further downsampling to `56x56`. Each subsequent residual stage contains multiple bottleneck blocks: each one of them employs three convolutions: a `1x1` convolution reduces channel dimensionality (example: from 256 to 64), a `3x3` convolution processes spatial features at reduced dimensionality, and a final `1x1` convolution expands back to the output channel count (following the same example, returns to 256).

After each convolution, batch normalization and ReLU activation are applied: most specifically, a residual shortcut connection adds the block input directly to its output, allowing identity mapping when needed. In the integral convolutional blocks (the first block of each stage), the shortcut includes a `1x1` convolution with stride adjustment to match dimensions, while in identity blocks the shortcut is simply a direct pass-through.

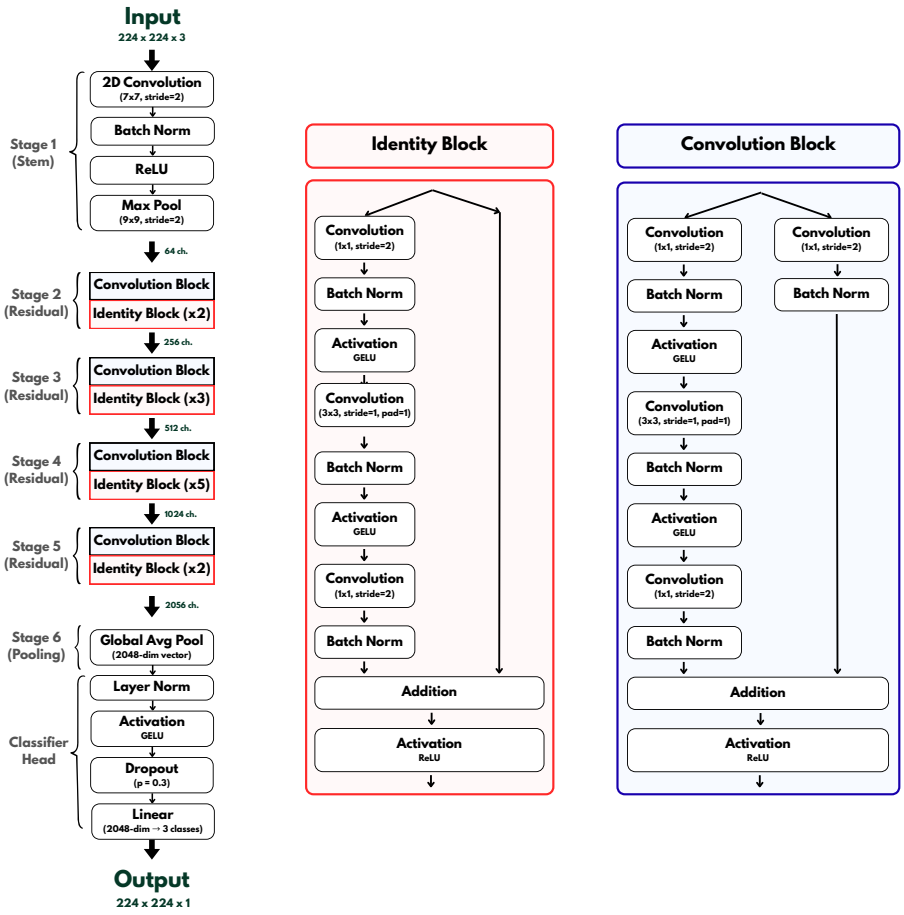


Figure 5.2: Diagram of the ResNet-50 Architecture

Following the residual stages, a progressive channel expansion can be observed:

- **Stage 2**: contains 3 bottleneck blocks outputting 256 channels (64 $\rightarrow$ 256);
- **Stage 3**: contains 4 blocks outputting 512 channels (256 $\rightarrow$ 512);
- **Stage 4**: contains 6 blocks outputting 1024 channels (512 $\rightarrow$ 1024);
- **Stage 5**: contains 3 blocks outputting 2048 channels (1024 $\rightarrow$ 2048).

After the final stage, global average pooling reduces the `(7x7x2048)` feature map to a single 2048-dimensional vector per image.

Classifier Head: The classification head replaces the original ResNet-50 ImageNet classifier with a custom Multi-Layer Perceptron (MLP) tailored for the three-class EGGIM task in hand. This head classifier starts with a layer normalization (normalizes the input across the features, helping to stabilize the learning process), passing next to a Gaussian Error Linear Unit (GELU) activation function (application of a nonlinear transformation that

is smooth and differentiable, in order to model complex patterns). After the GELU, a Dropout is applied with a probability of 0.3 (30% of the nodes are randomly set to zero during training; prevents overfitting and improves generalization), and in the end a final linear transformation layer is implemented, projecting the dimensionality from 2048 channels to 3 output logits corresponding to the 3 EGGIM classes.

5.3.2 ConvNeXt: Modernized Convolutional Network

ConvNeXt represents a modernization of traditional convolutional architectures, incorporating design principles from Vision Transformers while maintaining pure Convolutional Neural Networks's operations. Introduced by Liu et al., ConvNeXt reconsiders the standard ResNet design with systematic modifications: replacing ReLU with GELU activation, substituting batch normalization with layer normalization, employing larger kernel sizes (such as `7x7` depthwise convolutions), and adopting an inverted bottleneck structure where the MLP expansion precedes dimensionality reduction, ultimately implementing changes that bridge the performance gap between CNNs and ViTs without requiring attention mechanisms. The base checkpoint of `facebook/convnext-base-224` was used, providing weights from the supervised pre-training of ImageNet-1K [31, 75, 76], and the ConvNeXt architecture designed for this research purpose can be observed in Figure 5.3.

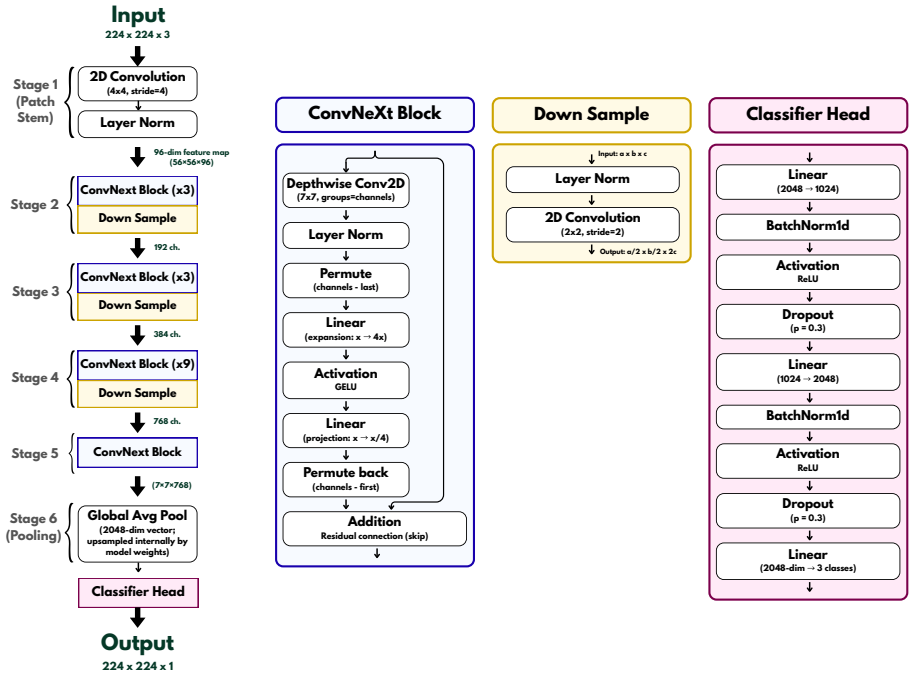


Figure 5.3: Diagram of the ConvNeXt Architecture

Backbone Structure: The ConvNeXt backbone employs a hierarchical stage-based architecture: Stage 1 unfolds with a patchify stem applying a `4×4` convolution with stride 4 to the input, aggressively downsampling the `224×224` input to `56×56` with 96 channels, mimicking the embedding of the vision transformer patch, but using convolutions. Following, there are 3 stages (Stages 2, 3 and 4) of repeated ConvNeXt blocks followed by downsampling, with one final stage (Stage 5) after with a single ConvNeXt block.

A ConvNeXt block implements an inverted bottleneck design: a depthwise `7x7` convolution starts, mixing spatial information by operating on individual channels independently, followed by a layer normalization block. The data is then permuted to a *channels-last* format, rearranging the dimensions into batch, height, width, and channels [B,H,W,C], and a linear expansion step succeeds, increasing the channel dimension from C to 4C to allow more complex feature learning. Afterwards, a GELU activation is applied, with a linear projection projected on the channel dimension from 4C to C. The data is then permuted back to the original *channels-first* format. In the end, a residual connection is added, helping to preserve the original input signal and mitigate the problem of vanishing gradients mentioned before [6]. The permutation operations enable efficient application of linear layers to the channel dimension, analogously to transformer MLP blocks. Analogously to the ResNet-50 residual phase, a progressive channel expansion can also be observed during this ConvNeXt/Downsampling blocks sequence:

- **Stage 2**: contains 3 ConvNeXt blocks at 96 channels (56×56 resolution). Downsampling via sucessive LayerNorm and `2×2` convolution (of stride 2) operations then reduce the resolution to `28×28` and expands channels to 192;
- **Stage 3**: processes 3 ConvNeXt blocks at 192 channels, with another downsampling process in the end reducing the resolution to *14×14* and expanding to 384 channels;
- **Stage 4**: includes 9 ConvNeXt blocks at 384 channels, suceeded by a final downsampling that reduces to `7×7` with 768 channels;
- **Stage 5**: encompasses 3 blocks at 768 channels, with no downsampling operation after.

Then, the global average pooling phase collapses the `7×7` spatial dimensions, and the HuggingFace implementation internally projects the 768-dimensional pooled features to 2048 dimensions, ensuring compatibility with the other model classifier dimensions.

Classifier Head: The custom classification head for EGGIM takes the 2048-dimensional output of the backbone and applies a three-layer MLP: begins with a linear layer reducing dimensions from 2048 to 1024, followed by Batch Normalization (BatchNorm1d) normalizing inputs to enhance stability. A ReLU activation function introduces non-linearity, and a Dropout layer at 0.3 probability (30%) helps prevent overfitting. The pattern repeats with another linear layer that expands dimensions from 1024 to 2048, followed by similar batch normalization, ReLU activation, and 30% chance of Dropout. Ultimately, a final linear transformation layer is implemented, projecting the dimensionality from 2048 channels to 3 output logits corresponding to the 3 EGGIM classes. This classification design expands and contracts the feature space, allowing the head to learn complex decision boundaries, the difference compared to the ResNet-50 head being that ConvNeXt uses batch normalization (BatchNorm1d) instead of layer normalization and ReLU rather than GELU. Also, the fact of it being a three-layer MLP structure (compared to the effective two-layer head of ResNet-50) provides additional representational capacity for the model.

5.3.3 DINOv2: Self-Supervised Vision Transformer

DINOv2 is a self-supervised vision transformer foundation model introduced by Oquab et al., pre-trained on 142 million curated images using self-distillation without labels. Unlike ResNet-50 and ConvNeXt that rely on supervised ImageNet pre-training, DINOv2 learns visual representations through contrastive and distillation objectives, enabling it to capture richer semantic structure without human annotations: a true representation of Foundation Models. The `facebook/dinov2-base` checkpoint was used, providing a Vision Transformers architecture with 12 transformer encoder layers and hidden dimension of 768 [77, 78], with the architecture perfected for this study classification task available to visualize in Figure 5.4.

Self-Supervised Pre-Training Methodology: DINOv2 is pre-trained within an unsupervised self-distillation framework, where a teacher and a student network each process differently augmented versions of the same image and are optimized to produce similar high-level representations without the use of manual annotations [77]. The teacher's parameters are updated as an exponential moving average of the student's weights, resulting in a slowly changing target distribution, while the student is updated by gradient descent to align its output distribution with that of the teacher with strong data augmentation [77]. Applied on a scale to a curated corpus of roughly 142 million heterogeneous

images, this approach drives the model to learn robust and semantically rich visual representations that transfer effectively to a plenitude of downstream tasks, such as EGGIM-based classification of precancerous gastric lesions in this research's use case [7, 77].

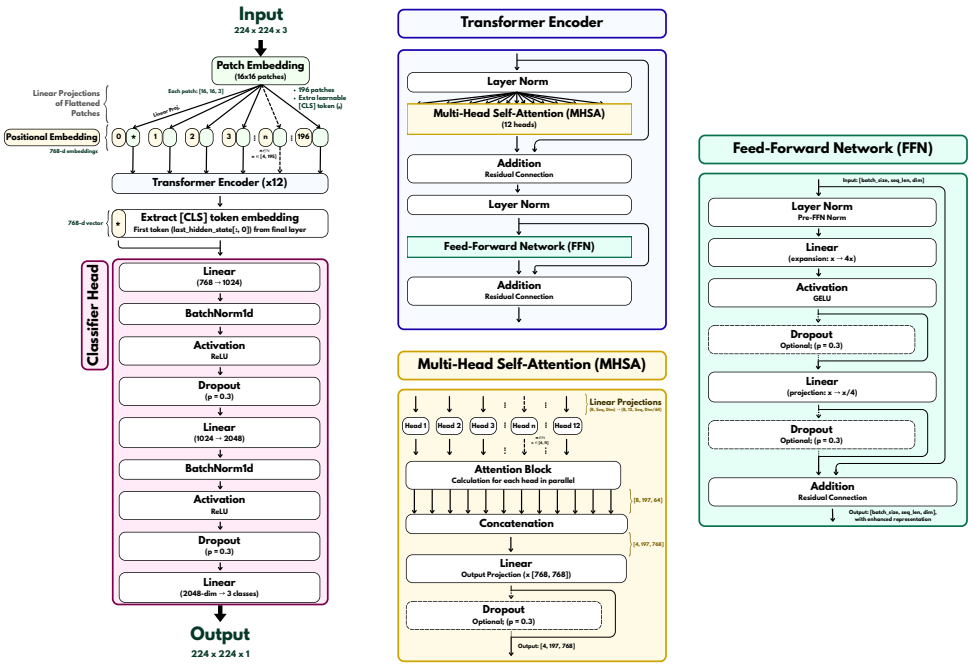


Figure 5.4: Diagram of the DINOv2 Architecture

Input Processing and Patch Embedding: DINOv2 processes images by treating them as patches sequences instead of using hierarchical convolutional layers. The `224×224×3` input image, is divided into `16×16` non-overlapping patches, forming a `14×14` grid of 196 patches in total. Each patch is then flattened to 768 values (`16×16×3` = 768) and linearly projected into a 768-dimensional patch embedding. A learnable CLS (classification) token, which is a special 768-dimensional embedding, is added at the beginning of the patch sequence, making a total of 197 tokens (1 CLS + 196 patches). Along with these, learnable positional embeddings are included, which are unique 768-dimensional vectors assigned to each of the 197 positions, providing an element for every token embedding. These positional embeddings preserve spatial information despite the permutation invariance of the transformers, resulting in the sequence `[B, 197, 768]` that is fed into the next phase: the transformer encoder.

Transformer Encoder: The transformer encoder consists of 12 identical successive layers, each containing essentially two highly influential subblocks: Multi-Head Self-Attention (MHSA) and Feed-Forward Network (FFN), both with residual connections and layer normalization between them sequentially.

- **Multi-Head Self-Attention**: mechanism that allows each token to aggregate contextual information from all other tokens, allowing the model to capture long-range dependencies within the image. Given an input sequence $\mathbf{X} \in \mathbb{R}^{N \times d}$, where $N = 197$ tokens (196 patches plus the CLS token) and $d = 768$ feature dimensions, MHSA first projects $\mathbf{X}$ into query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ matrices for each of the $H = 12$ attention heads:

$$\mathbf{Q}_h = \mathbf{X}\mathbf{W}_h^Q, \quad \mathbf{K}_h = \mathbf{X}\mathbf{W}_h^K, \quad \mathbf{V}_h = \mathbf{X}\mathbf{W}_h^V$$

where $\mathbf{W}_h^{Q,K,V} \in \mathbb{R}^{d \times d_h}$, and $d_h = \frac{d}{H} = 64$.

For each head h, the attention scores $\mathbf{S}_h \in \mathbb{R}^{N \times N}$ are computed as the scaled dot product of queries and keys:

$$\mathbf{S}_h = \frac{\mathbf{Q}_h \mathbf{K}_h^\top}{\sqrt{d_h}}$$

To convert these scores into a probability distribution reflecting attention weights, a softmax function is applied along the last dimension:

$$\mathbf{A}_h = \text{softmax}(\mathbf{S}_h)$$

The output of each attention head is then a weighted sum of the values, aggregating information from all tokens:

$$\mathbf{Z}_h = \mathbf{A}_h \mathbf{V}_h$$

The outputs from all heads are concatenated and linearly projected to produce the final MHSA output:

$$\text{MHSA}(\mathbf{X}) = \text{Concat}(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_H)\mathbf{W}^O$$

where $\mathbf{W}^O \in \mathbb{R}^{d \times d}$ is a learnable projection matrix.

This output is then combined with the input through a residual connection and layer normalization, allowing gradient flow and stable training (optionally, the dropout of 30% can also be activated, although its disablement proved better results for this research). Additionally, the CLS token attends to all patch tokens in action during MHSA, progressively aggregating global image information across the 12 layers.

- **Feed-Forward Network**: following the MHSA block, this network is then applied independently to each token to enhance its non-linear representation. Given input $\mathbf{Y} \in \mathbb{R}^{N \times d}$ from the post-layer-normalization output, the FFN firstly applies a linear transformation for expansion(4x; 768 $\rightarrow$ 3072). Next, the GELU activation function is applied, followed by an optional dropout layer of 30% chance. A linear projection then reduces the dimensionality (1/4x; 3072 $\rightarrow$ 768), finalizing with another optional identical Dropout layer and a residual connection succeeds, which retrieves a structured output with enhanced representation.

CLS Token Extraction and Classifier Head: After the 12th transformer layer, the output is $[B, 197, 768]$, where position 0 corresponds to the CLS token. The extraction is performed by indexing: `features = self.dinov2(x).last_hidden_state:, 0`, which yields $B, 768$. Unlike previous CNNs that use spatial grouping functions such as global average grouping, DINOv2 relies on the learned aggregation capabilities of the CLS token. This token has no fixed semantics, as it learns through self-supervised training to aggregate task-relevant information via attention. The then extracted 768-dimensional CLS features are passed to a custom three-layer MLP classification head: it begins with a linear layer expanding input features from 768 to 1024 dimensions, followed for Batch Normalization and ReLU activation layers and then followed by a 30% dropout to prevent overfitting. The next linear layer expands the dimensions once again from 1024 to 2048, followed in a similar fashion by Batch Normalization, ReLU, and 30% chance dropout. Finally, a last linear layer projects the dimensionality from 2048 channels to 3 output logits corresponding to the 3 EGGIM classes. All these layers enable FMs to refine and learn specific decision boundaries on top of DINOv2's general-purpose representations.

5.4. Training Methodologies

5.4.1 Progressive Unfreezing and Learning Rates

Progressive unfreezing addresses the "fine-tuned adaptability to pre-trained general models" challenge through a controlled layer-by-layer adaptation mechanism that preserves learned representations while enabling task-specific refinement. The backbone starts as completely frozen, training only the newly initialized classification head for several epochs to establish stable predictions based on pre-trained features, but gradually, deeper layers are systematically unfrozen in reverse order, allowing gradual adaptation without catastrophic forgetting of the fundamental representations encoded during the pre-training weights knowledge.

Discriminative learning rates are also strategically implemented, applying progressively smaller learning rates to earlier layers that specialize more on encoding general features like edges and textures. For example, in the ResNet-50 fine-tuned configuration (Figure 5.2), layer 4/stage 5 receives a learning rate of 3×10^{-5}, layer 3/stage 4 uses 1×10^{-5}, and layer 2/stage 3 employs 5×10^{-6}, while the classification head operates at 3×10^{-4}. This differential rate assignment prevents early layers from excessive modification, while permitting deeper layers to adapt to the gastric lesion domain. Combined with techniques such as exponential moving average weight tracking and batch normalization stabilization through evaluation mode maintenance during fine-tuning, this learning rate own tuning ensures controlled adaptation without destabilizing pre-trained representations.

Progressive unfreezing and discriminative learning rates are techniques applied exclusively to fine-tuned configurations of the three architectures (ResNet-50, ConvNeXt, DINOv2).

5.4.2 Advanced Data Augmentation

Robust model generalization in medical imaging demands augmentation strategies that extend beyond conventional transformations, really pushing the model capabilities to the maximum, as precision is key in this environment. This study implements a hierarchical augmentation framework differentiated by the training paradigm.

In pre-trained configurations opt to employ more conservative augmentations, such as:

- **RandomResizedCrop:** Extracts random crops on certain scales, then resizing them to 224×224 using bicubic interpolation (downscales images while minimizing quality loss by creating a smoother and more detailed representation). This technique simulates zoom variations and enhances compositional diversity, facilitating lesion recognition on multiple scales without explicitly altering resolution.
- **RandomRotation:** Rotates images by 5–15° depending on configuration, taking into account variations in the angle of the camera during endoscopy. Smaller angles in modern architectures suggest superior data efficiency and stability.
- **ColorJitter:** Adjusts brightness, contrast, saturation, and hue independently. Endoscopic imaging exhibits lighting variations due to the setting of equipment and the anatomical positioning, which also mimics the inconsistencies of the real-world clinic data acquisition.

- **GaussianBlur:** Applies Gaussian blur: simulates focus variations and motion blur common during rapid endoscope passes, teaching robust lesion classification despite soft focus.
- **RandomErasing:** Erases 2–33% of the image area with random pixel values, simulating partial data occlusions while forcing the model to classify from incomplete visual information, mirroring clinical scenarios.
- **RandAugment:** Applies one random operation (shear, translation, rotate, etc.) from a predefined set with a certain magnitude, adding diversity without overwhelming the training signal.
- **Progressive Resolution Curriculum:** Gradually increases input resolution from 192$\rightarrow$224$\rightarrow$256 pixels during training. Initial epochs train on lower resolutions (faster), gradually moving to higher resolutions, which enhances convergence while managing computational cost.

These transformations preserve anatomical structure while introducing sufficient variability to reduce overfitting without disrupting pre-learned feature statistics.

However, in fine-tuned configurations, conservative augmentations are applied alongside other more aggressive augmentation techniques, which force the model to learn under more substantial perturbations. One of the more aggressive augmentation techniques used is **MixUp**, proposed by Zhang et al., which, rather than traditionally augmenting images, generates new training samples by blending pairs of images and their labels. With this, the model is more encouraged to learn smooth decision boundaries by training combinations of classes, reducing overconfidence and improving robustness to ambiguous or noisy inputs, being especially effective when training data is limited and unbalanced [79]. **CutMix** by Yun et al. is another one of these techniques, which expands on MixUp's principle by replacing rectangular image regions with patches from other samples, with label mixing proportional to the pixel area contributed by each image. Random erasing removes rectangular regions (between 2% and 33% of the image area) during training, simulating occlusion, and improving robustness to incomplete visual information [80].

These aggressive techniques (Figure 5.5) are applied with controlled probability (higher probability in early training, progressively reduced with time) to balance the benefits of regularization against the stability of training.

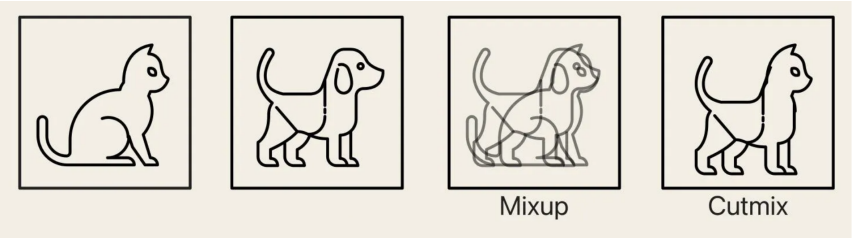


Figure 5.5: Advanced Data Augmentation Techniques: Mixup vs Cutmix (Adapted from [81])

5.4.3 Class Imbalance Mitigation

As previously explained, the EGGIM dataset exhibits a severe class imbalance, Class 1 representing only 19.2% of the training samples, while Classes 0 and 2 comprise 43.7% and 36.9%, respectively. Addressing this imbalance requires strategic interventions beyond the standard cross-entropy loss, so **weighted random sampling** is used to ensure that each training batch contains balanced class representation by oversampling minority classes during data loading (in a balanced way and proportional to the density of loss functions and augmentations per configuration). The class weights derived from scikit-learn's `compute_class_weight` function provide inverse frequency weighting: $w_i = \frac{N}{N_c \cdot n_i}$, where N is the total sample count, N_c is the number of classes, and n_i represents samples in class i. These weights initialize the scaling factors of the loss function, emphasizing minority class errors during optimization, with this mitigation strategy being applied in all six configurations.

5.4.4 Advanced Loss Functions

A plethora of different loss functions were tested for each learning configuration, with three main custom loss functions ultimately tested to be implemented over the final 6 pipelines:

- **Focal Loss with Label Smoothing**: introduced by Lin et al., addresses class imbalance by reducing the weight assigned to easily classified samples and focusing on hard-to-classify cases [82]. The formulation extends standard cross-entropy with a modulating factor:

$$\mathcal{L}_{\text{focal}} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t)$$

where p_t is the predicted probability for the ground truth class, $\gamma > 0$ is the focus parameter controlling down-weighting strength, and α_t provides class-specific weighting. This approach naturally directs the model's attention to the underrepresented class 1 without requiring manual reweighting, using a focusing parameter $\gamma = 2$ and class-specific weights derived from inverse frequencies. In addition, the label smoothing proposed by Szegedy et al. complements focal loss by preventing overconfident predictions. It transforms hard one-hot labels y_{hot} into soft targets:

$$(y_{\text{smooth}} = (1 - \epsilon)y_{\text{hot}} + \frac{\epsilon}{K}$$

where $\epsilon = 0.1$ is the smoothing factor and K is the number of classes. This regularization distributes a small probability mass to incorrect classes, preventing the model from becoming overconfident while improving calibration, which is particularly important for clinical decision support where prediction confidence must reflect true uncertainty [83]. The combined focal loss with label smoothing is employed in ConvNeXt fine-tuned and DINOv2 fine-tuned.

- **Balanced Softmax Loss**: developed by Ren et al., aims to tackle long-tailed recognition problems, incorporating training class frequencies into the softmax computation, accounting for distribution shifts between training and deployment. Balanced softmax adjusts the activation function as:

$$\phi_j = \frac{n_j e^{\eta_j}}{\sum_{i=1}^{k} n_i e^{\eta_i}}$$

where n_j is the training sample count for class j and η_j represents the model's logit output for class j. Unlike standard cross-entropy that implicitly assumes balanced test distributions, this loss weights each class prediction proportionally to its training sample count. This design naturally compensates for imbalance without requiring post-hoc calibration or threshold adjustment, maintaining consistency with balanced evaluation metrics. The implementation derives class counts directly from the training set, making it parameter-free beyond the loss itself [84]. The loss function then becomes:

$$\mathcal{L}_{\text{balanced}} = -\log\left(\frac{n_y e^{\eta_y}}{\sum_{i=1}^{k} n_i e^{\eta_i}}\right)$$

- **Logit Adjustement Loss**: proposed by Menon et al., this method introduces post-hoc or training-time corrections to model logits, enforcing larger margins between rare and common classes [85]. The adjusted softmax incorporates class prior information directly:

$$\mathcal{L}_{\text{logit-adj}} = -\log\left(\frac{e^{\eta_y+\tau\log\pi_y}}{\sum_{j=1}^{k} e^{\eta_j+\tau\log\pi_j}}\right)$$

  where π_j represents the prior probability of class j (estimated as $\pi_j = n_j/N$), and τ is a temperature parameter controlling adjustment strength (typically $\tau = 1$). This formulation demands larger decision margins for rare classes by adding $\tau\log\pi_y$ to logits, preventing dominant classes from overwhelming minority predictions. The adjustment term $\log(\pi_{y'}/\pi_y)$ establishes a relative margin between classes y and y', making sure the model learns discriminative boundaries appropriate for the training distribution.

The remaining configurations (all pre-trained pipelines and ResNet-50 fine-tuned baseline) did not benefit substantially from additional custom loss functions and consequently retained variants of the standard **Weighted Cross-Entropy with Label Smoothing** implementation provided by the *torch* library.

5.5. Evaluation Framework and Testing Methodologies

5.5.1 Test-Time Augmentation

Test-Time Augmentation (TTA) enhances prediction reliability of the predictions by creating several augmented versions of each test image and combining their predictions. In contrast to training-time augmentation that increases dataset diversity, TTA introduces the model to geometric and photometric transformations of the same input during inference, resulting in ensemble predictions from a single model. This approach improves robustness to small perturbations and reduces the variance of the prediction without requiring additional model training.

The TTA strategy in use utilizes dihedral transformations of D4, which include eight types of symmetries:

1. Original image;
2. Horizontal flip;
3. Vertical flip;
4. 90-degree rotation;
5. 180-degree rotation;
6. 270-degree rotation;
7. 90-degree rotation with horizontal flip;
8. 90-degree rotation with vertical flip.

Additionally, multiscale evaluation at resolutions [224, 256, 288] complements geometric augmentations, with predictions being averaged over these scales. The horizontal flip is used to account for the non-directional nature of the anatomical features in the gastric mucosa, while rotations are intended to manage potential variations in camera orientation during endoscopy. TTA is applied in different proportions for the six configurations during the final testing and evaluation.

5.5.2 Ensemble Learning

Ensemble learning combines multiple model predictions from several models to decrease prediction variance and improve generalization beyond what individual models achieve. This technique leverages diversity in architecture and training, with different configurations and random elements causing a variety of learned patterns. Three ensemble techniques were evaluated and applied at different stages of the pipeline:

- **Simple averaging**: baseline approach applied immediately after validation, treating all models on an equal basis, and averaging the predicted class probabilities within the ensemble. This technique does not necessitate calibration and serves as a sanity check: if it fails to outperform the individual models, it might indicate a lack of architectural or training diversity;

- **Weighted averaging**: applied during model selection when validation performance varies significantly across configurations, more importance is given to models that perform well on validation data. This weighting reflects confidence that models with better validation results should have a more significant influence on the final prediction. The weights are determined only once after the validation process is finished and remain constant during the test evaluation;

- **Logit-space ensembling**: implemented in the final evaluation pipeline before softmax normalization is applied. Instead of averaging probabilities after softmax, the averaging occurs directly on the raw model output (logits), with a softmax function being applied once right after. This method tends to yield better-calibrated confidence scores given that the averaging is done in the linear logit space, prior to the introduction of non-linearity.

The ensemble functions during the post-training evaluation phase for all six pipelines. Once each model has finished training and validation, their test data predictions are combined based on the selected strategy. The most effective ensemble configuration is determined by assessing the validation set performance across all potential combinations.

5.5.3 Regularization and Optimization Stability

Training deep networks for small medical imaging datasets requires handling optimization dynamics with a range of regularization and stabilization strategies. These approaches extend past traditional dropout and data augmentation to address unpredictable gradient behaviors and model overfitting, ensuring stable and efficient fine-tuning of large pretrained models.

- **Exponential Moving Average (EMA)**: EMA maintains a slowly updated copy of the model weights by blending current parameters with an exponential moving average ($\text{EMA}_{\text{new}} = \tau \cdot \text{EMA}_{\text{old}} + (1 - \tau) \cdot \text{weight}_{\text{current}}$, where $\tau = 0.999$). These smoothed weights are used for the validation evaluation, but do not participate in the gradient computation. EMA stabilizes the performance metrics by reducing the noise from batch-to-batch fluctuations and acts as an implicit regularizer, preventing the model from committing to noisy solutions. Applied to all fine-tuned configurations during training, EMA weights are swapped in before validation to produce more reliable performance estimates.

- **Gradient Clipping**: Global norm clipping (`torch.nn.utils.clip_grad_norm_`) caps gradient magnitudes to a maximum norm, typically $\|\nabla\| \leq 1.0$. This prevents exploding gradients that destabilize training, particularly problematic when fine-tuning with small learning rates and accumulated gradients: when gradients exceed the threshold, they are scaled down uniformly. Applied uniformly across all fine-tuned configurations, gradient clipping prevents weight divergence during backpropagation, allowing tighter discriminative learning rates without training collapse.

- **Mixed Precision Arithmetic (AMP):** Selectively applied to ResNet-50 and ConvNeXt models fine-tuned on CUDA-enabled GPUs, AMP optimizes computation by using float16 or bfloat16 for forward passes, whilst maintaining float32 precision for loss calculations and backward passes (technique only employed when utilizing the kempo2 GPU server; in Apple Silicon (MPS), AMP is not utilized due to numerical instabilities at lower precisions). A `GraphicScaler` adjusts the loss magnitude before backpropagation to prevent issues with gradient underflow at lower precision, allowing this method to cut GPU memory usage by about 50%, enabling faster batch processing and improved computational speed. In contrast, DINOv2 fine-tuned continues using full float32 precision to ensure stability, as its self-supervised pre-training, which relies on precise gradient estimates.

- **Temperature Scaling in Logit Adjustment**: When incorporating class priors into logits using logit adjustment, loss involves a temperature parameter τ that governs the strength of the adjustment by scaling the log-prior contribution. Increasing the value of τ increases the influence of the priors, while a τ of 1.0 implies the use of unscaled priors. This method adjusts the decision boundaries to align with the characteristics of the training distribution without altering the loss function itself. It is selectively employed in the ResNet-50 pre-trained configuration.

- **Batch Normalization Freezing**: During fine-tuning, backbone batch normalization layers remain in evaluation mode to preserve learned statistics while allowing deeper layers to adapt. Applied to ResNet-50 fine-tuned and ConvNeXt fine-tuned when unfreezing backbone layers (DINOv2 does not need this as it uses layer normalization on the classifier).

5.5.4 Evaluation Metrics

The evaluation of the model goes beyond aggregate accuracy to capture the distinct performance characteristics of each class, which proves essential for medical applications. Primary metrics include overall accuracy, macro-averaged F1-score, weighted F1-score, and per-class precision, recall, and F1-score, defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \quad \text{Recall}_c = \frac{TP_c}{TP_c + FN_c}, \quad \text{F1}_c = 2 \cdot \frac{\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}$$

where c denotes the class index, and TP, TN, FP, FN represent true positives, true negatives, false positives, and false negatives, respectively. The macro-averaged F1-score treats each class equally regardless of support, emphasizing minority class performance, while weighted F1 accounts for class frequency in the dataset.

In addition to this, confusion matrices are generated, offering detailed insights into specific misclassification patterns within all six pipelines. These matrices indicate whether errors occur between adjacent EGGIM risk categories (Class 0 versus Class 1, or Class 1 versus Class 2) or involve fundamental classification errors. This detailed analysis helps identify whether particular architecture-training pairs only confuse certain class transitions or demonstrate broader misinterpretations.

Dimensionality Reduction via t-SNE: Upon completion of model training, the features of the penultimate layer of the test set are extracted and projected into a 2D space through t-distributed Stochastic Neighbor Embedding (t-SNE) employing standard hyperparameters, where perplexity is adjusted based on the size of the dataset and the learning rate is set automatically. These visualizations illustrate the separation of classes into distinct clusters in the learned representation space, while also identifying potential class overlap or dataset artifacts. Misclassified samples are highlighted distinctly, highlighting areas where class boundaries are undefined even in high-dimensional space [86]. Individual t-SNE plots are produced for each of the six configurations, allowing for a direct visual comparisons of the learned representations and enabling a qualitative evaluation of the features' quality beyond numerical metrics.

All metrics are computed consistently across configurations to enable rigorous comparative analysis.

5.6. Results

Each of the six training configurations was meticulously trained and evaluated to assess classification performance in the three EGGIM risk categories. This section details the specific training and testing pipeline parameters for each configuration, followed by quantitative performance metrics, confusion matrices, and t-SNE visualizations that reveal the learned feature space structure.

5.6.1 ResNet-50 Pre-trained Configuration

Pipeline Specifications

Training Setup:

- Epochs: 100, with early stopping (patience=20);
- Unfreezing Strategy: Backbone frozen for 2 epochs, then gradual unfreezing (layer4/stage 5 at epoch 3, layer3/stage 4 at epoch 7, earlier layers remain frozen);
- Optimization: AdamW with discriminative Learning Rate (LR)s (head=3×10^{-3}, backbone=3×10^{-4}, weight decay=1×10^{-4});
- LR Scheduler: 5-epoch warmup + cosine decay, then reduce-on-plateau after unfreezing. Final warmup+cosine phase after second unfreeze;
- Checkpointing: None.

Training Augmentations:

- Traditional Augmentation: Resize(256), RandomResizedCrop(224, scale=0.85-1.0), RandomHorizontalFlip (p=0.5), RandomGrayscale (p=0.05);
- Advanced Augmentation: None.

Loss Functions & Class Imbalance:

- Loss Functions: Cross Entropy with label smoothing ($\epsilon = 0.05$) and balanced class weights;
- Class Imbalance: No class weighting or oversampling.

Training Regularization

- General Regularization: Gradient clipping (max_norm=1.0) and Dropout (p=0.5 in head);
- Advanced Regularization: None;

Testing/Inference:

- Testing Augmentation: Identity and Horizontal Flip;
- Testing Regularization: None;
- Prediction Ensembling: Simple predictions averaging.

Performance Results

Table 5.1: Per-Class Metrics for ResNet-50 Pre-trained Configuration

Metric	Class 0	Class 1	Class 2	Overall
Precision	61.94%	25.00%	57.55%	-
Recall	70.94%	8.70%	65.59%	-
F1-Score	66.14%	12.90%	61.31%	46.78% (macro)
Accuracy	-	-	-	57.81%

Table 5.2: Confusion Matrix for ResNet-50 Pre-trained Configuration

Predicted\True	Class 0	Class 1	Class 2
Class 0	83	10	24
Class 1	21	4	21
Class 2	30	2	61

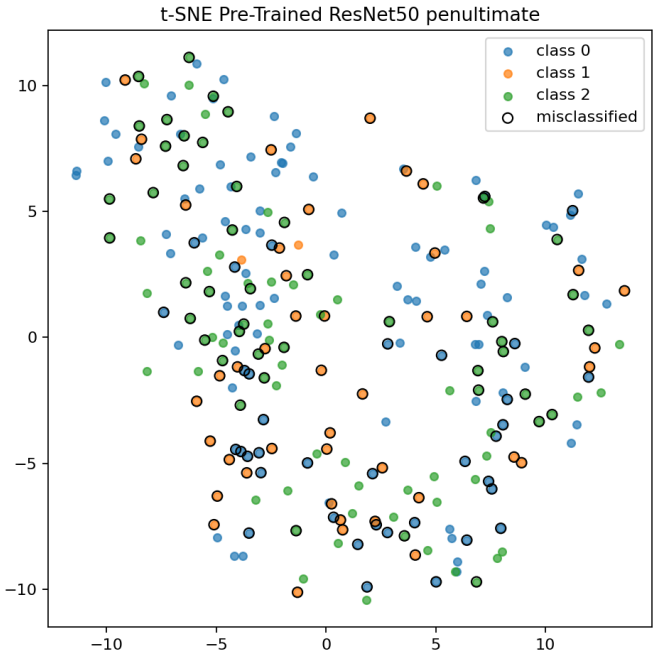


Figure 5.6: t-SNE visualization for ResNet-50 Pre-trained

5.6.2 ResNet-50 Fine-tuned Configuration

Pipeline Specifications

Training Setup:

- Epochs: 220, with early stopping (patience=30, macro-F1 based);
- Unfreezing Strategy: Progressive (head-only for 8 epochs, layer4/stage 5 at epoch 9, layer3/stage 4 at epoch 15, layer2/stage 3 at epoch 21);
- Optimization: Stochastic Gradient Descent (SGD) with Nesterov momentum (0.9), discriminative LRs (head=3×10^{-4}, backbone=5×10^{-6}, weight decay=1×10^{-5});
- LR Scheduler: Warmup from 0 during head-only, then cosine decay after unfreezing;
- Checkpointing: Top-5 checkpoint ensembling (macro-F1 based).

Training Augmentations:

- Traditional Augmentation: Progressive resolution (192$\rightarrow$224$\rightarrow$256), RandomResizedCrop (scale=0.7-1.0, bicubic interpolation), RandomHorizontalFlip (p=0.5), RandomVerticalFlip (p=0.1), RandomRotation(150°, p=0.3), GaussianBlur (p=0.15);
- Advanced Augmentation: RandAugment (1 operation, magnitude=5), ColorJitter (brightness/contrast/saturation=0.2, hue=0.05, p=0.3), RandomErasing (p=0.25), MixUp/CutMix (45% probability, epochs 1-16 only).

Loss Functions & Class Imbalance:

- Loss Functions: Cross Entropy with label smoothing ($\epsilon = 0.05$);
- Class Imbalance: WeightedRandomSampler (inverse frequency) during early training, switches to uniform sampling after epoch 60.

Training Regularization:

- General Regularization: Gradient clipping (max_norm=1.0);
- Advanced Regularization: EMA (decay=0.999), AMP with GradScaler;

Testing/Inference:

- Testing Augmentation: D4 dihedral group (8 variants);
- Testing Regularization: Automatic temperature scaling;
- Prediction Ensembling: Multiscale ensembling (either by specifying multiple weight files or a log of top models, which are loaded and combined using logit averaging).

Performance Results

Table 5.3: Per-Class Metrics for ResNet-50 Fine-tuned Configuration

Metric	Class 0	Class 1	Class 2	Overall
Precision	65.03%	23.53%	62.50%	-
Recall	79.49%	8.70%	64.52%	-
F1-Score	71.54%	12.70%	63.49%	49.24% (macro)
Accuracy	-	-	-	61.33%

Table 5.4: Confusion Matrix for ResNet-50 Fine-tuned Configuration

Predicted\True	Class 0	Class 1	Class 2
Class 0	93	7	17
Class 1	23	4	19
Class 2	27	6	60

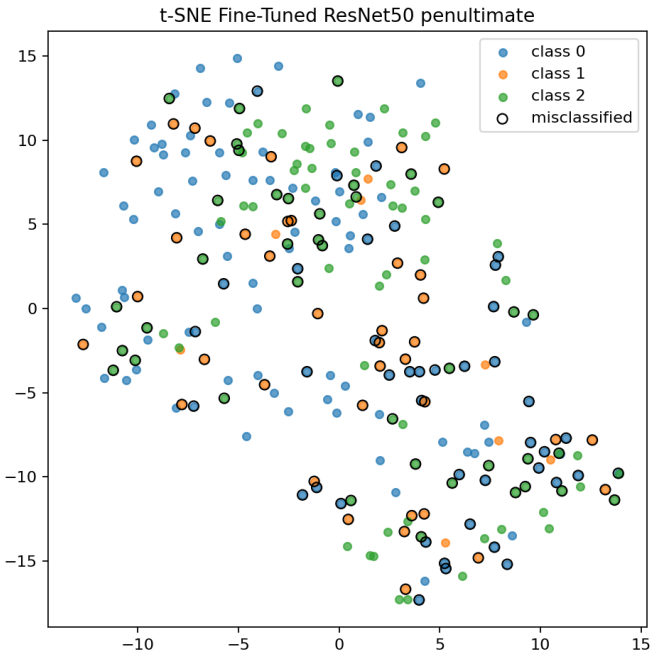


Figure 5.7: t-SNE visualization for ResNet-50 Fine-tuned

5.6.3 ConvNeXt Pre-trained Configuration

Pipeline Specifications

Training Setup:

- Epochs: 100;
- Unfreezing Strategy: Backbone frozen for 5 epochs, then unfreeze last 2 stages;
- Optimization: AdamW (LR=5×10^{-5}, weight decay=0.01);
- LR Scheduler: Cosine Annealing;
- Checkpointing: None.

Training Augmentations:

- Traditional Augmentation: RandomResizedCrop(224, scale=0.85-1.0), RandomHorizontalFlip (p=0.5), RandomRotation(5°), RandomErasing (p=0.25);
- Advanced Augmentation: ColorJitter ($\sim$8%), and MixUp (40% probability).

Loss Functions & Class Imbalance:

- Loss Functions: Cross Entropy with label smoothing ($\epsilon = 0.1$) and balanced class weights;
- Class Imbalance: Class 1 oversampled 3 times via WeightedRandomSampler.

Training Regularization:

- General Regularization: Gradient clipping (max_norm=1.0);
- Advanced Regularization: None;

Testing/Inference:

- Testing Augmentation: Random crops, horizontal flips, 25° rotation, color jitter (4 variants total);
- Testing Regularization: Just MixUp;
- Prediction Ensembling: Simple prediction averaging.

Performance Results

Table 5.5: Per-Class Metrics for ConvNeXt Pre-trained Configuration

Metric	Class 0	Class 1	Class 2	Overall
Precision	70.50%	16.67%	65.71%	-
Recall	83.76%	4.35%	74.19%	-
F1-Score	76.56%	6.90%	69.70%	51.05% (macro)
Accuracy	-	-	-	66.02%

Table 5.6: Confusion Matrix for ConvNeXt Pre-trained Configuration

Predicted\True	Class 0	Class 1	Class 2
Class 0	98	3	16
Class 1	24	2	20
Class 2	17	7	69

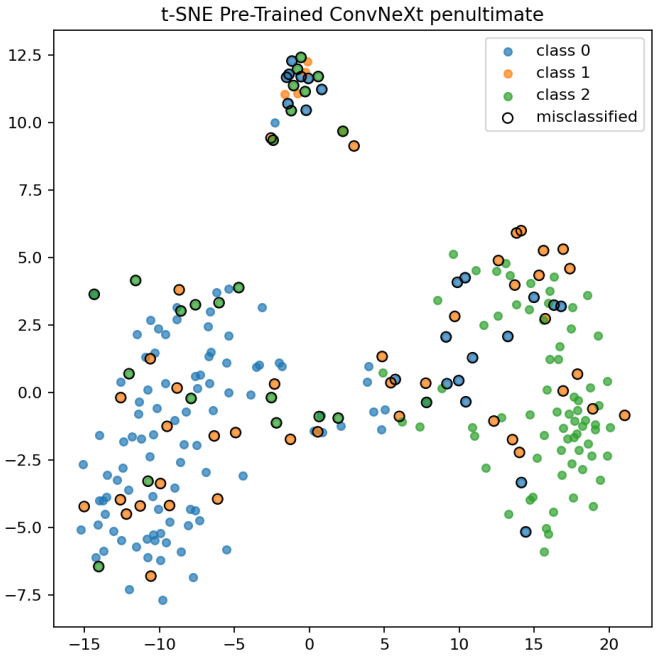


Figure 5.8: t-SNE visualization for ConvNeXt Pre-trained

5.6.4 ConvNeXt Fine-tuned Configuration

Pipeline Specifications

Training Setup:

- Epochs: 150, with early stopping (patience=25, macro-F1 based);
- Unfreezing Strategy: Progressive (head-only for 4 epochs, last stage at epoch 5, all stages by epoch 9);
- Optimization: AdamW with discriminative LRs (head=1×10^{-3}, deeper layers=4.5×10^{-5});
- LR Scheduler: Tiny warmup + cosine annealing;
- Checkpointing: None.

Training Augmentations:

- Traditional Augmentation: RandomResizedCrop, RandomHorizontalFlip, RandomRotation(5°), RandomErasing (p=0.25);
- Advanced Augmentation: ColorJitter ($\leq$8%), just MixUp (20% probability, head only phase).

Loss Functions & Class Imbalance:

- Loss Functions: Weighted Cross Entropy with label smoothing ($\epsilon = 0.02$-0.05), power law scaling factor 1.1 for class weights;
- Class Imbalance: Powered weighting sampler (inverse class frequency).

Training Regularization:

- General Regularization: Gradient clipping (max_norm=1.0);
- Advanced Regularization: EMA (decay=0.999), AMP with GradScaler;

Testing/Inference:

- Testing Augmentation: Multiple rotations: 5, 10, and 15 degrees
- Testing Regularization: Temperature tuning and per-class bias adjustment;
- Prediction Ensembling: Ensemble with pre-trained ConvNeXt checkpoint (adjustable blend ratio), multiscale evaluation.

Performance Results

Table 5.7: Per-Class Metrics for ConvNeXt Fine-tuned Configuration

Metric	Class 0	Class 1	Class 2	Overall
Precision	67.55%	30.77%	70.65%	-
Recall	87.18%	8.70%	69.89%	-
F1-Score	76.12%	13.56%	70.27%	53.32% (macro)
Accuracy	-	-	-	66.80%

Table 5.8: Confusion Matrix for ConvNeXt Fine-tuned Configuration

Predicted\True	Class 0	Class 1	Class 2
Class 0	102	3	12
Class 1	27	4	15
Class 2	22	6	65

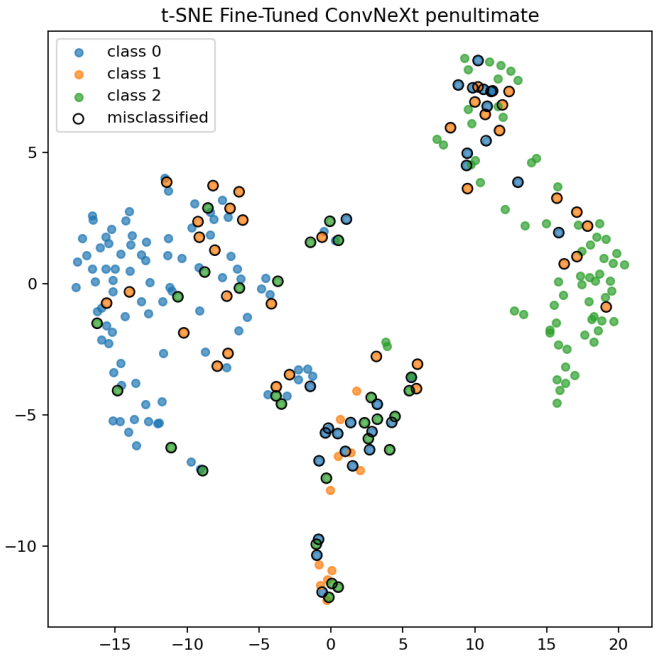


Figure 5.9: t-SNE visualization for ConvNeXt Fine-tuned

5.6.5 DINOv2 Pre-trained Configuration

Pipeline Specifications

Training Setup:

- Epochs: 100-150;
- Unfreezing Strategy: Backbone frozen for 5 epochs, then unfreeze last 2 transformer layers;
- Optimization: AdamW (LR=5×10^{-5}, weight decay=0.01);
- LR Scheduler: Cosine annealing;
- Checkpointing: None.

Training Augmentations:

- Traditional Augmentation: RandomResizedCrop and RandomHorizontalFlip;
- Advanced Augmentation: Just MixUp (40% probability).

Loss Functions & Class Imbalance:

- Loss Functions: Smooth Cross Entropy ($\epsilon = 0.1$) with class-balanced weights (computed inversely proportional to class frequency);
- Class Imbalance: Class 1 oversampled 3 times using WeightedRandomSampler.

Training Regularization:

- General Regularization: Gradient Clipping (max_norm=1.0);
- Advanced Regularization: None;

Testing/Inference:

- Testing Augmentation: Random crops, horizontal flips, 25° rotation, color jitter;
- Testing Regularization: None;
- Prediction Ensembling: Simple prediction averaging.

Performance Results

Table 5.9: Per-Class Metrics for DINOv2 Pre-trained Configuration

Metric	Class 0	Class 1	Class 2	Overall
Precision	65.71%	15.79%	64.95%	-
Recall	78.63%	6.52%	67.74%	-
F1-Score	71.60%	9.23%	66.32%	49.05% (macro)
Accuracy	-	-	-	61.72%

Table 5.10: Confusion Matrix for DINOv2 Pre-trained Configuration

Predicted\True	Class 0	Class 1	Class 2
Class 0	92	9	16
Class 1	25	3	18
Class 2	23	7	63

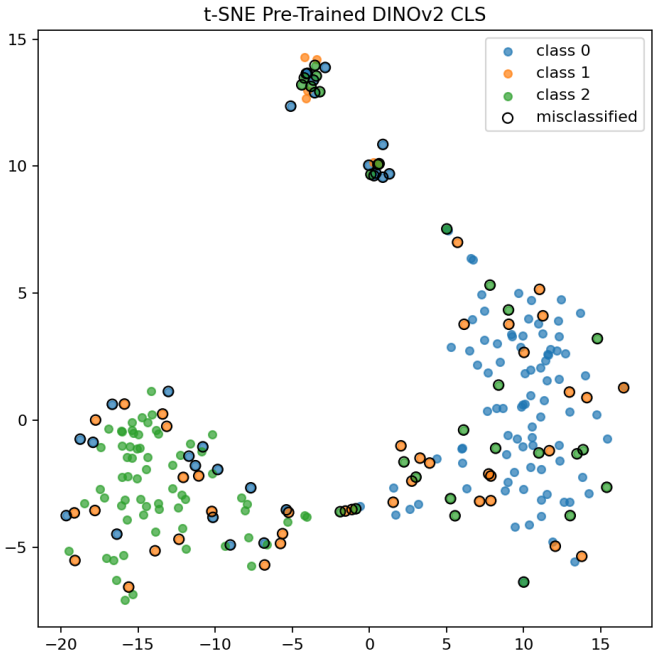


Figure 5.10: t-SNE visualization for DINOv2 Pre-trained

5.6.6 DINOv2 Fine-tuned Configuration

Pipeline Specifications

Training Setup:

- Epochs: 250, with early stopping (patience=25);
- Unfreezing Strategy: Backbone frozen for 4 epochs, then unfreeze last 4 transformer layers;
- Optimization: AdamW (initial head-only LR=5×10^{-5}, after unfreezing LR=7×10^{-5}, weight decay=0.01);
- LR Scheduler: Cosine annealing, reset after unfreezing;
- Checkpointing: Checkpoint averaging (best model + top-5 epochs).

Training Augmentations:

- Traditional Augmentation: RandomResizedCrop(224, scale=0.75-1.0), RandomHorizontalFlip, RandomVerticalFlip, RandomRotation(10°);
- Advanced Augmentation: MixUp ($\alpha = 0.4$, 25% probability, disabled after epoch 35), CutMix ($\alpha = 1.0$, 25% probability, disabled after epoch 35), ColorJitter (brightness/contrast/saturation=0.2). CutMix and MixUp are never applied to the same image.

Loss Functions & Class Imbalance:

- Loss Functions: Focal loss with label smoothing ($\gamma = 2.0$, $\epsilon = 0.08$) and balanced class weights with 30% boost on Class 1's weight;
- Class Imbalance: Class 1 oversampled 3 times and Class 2 oversampled 2 times by the sampler.

Training Regularization:

- General Regularization: None;
- Advanced Regularization: None.

Testing/Inference:

- Testing Augmentation: Two modes (light: identity + horizontal flip + 10° rotation; stable: resized crop, horizontal/vertical flips, 25° rotation, color jitter);
- Testing Regularization: None;
- Prediction Ensembling: Simple prediction averaging.

Performance Results

Table 5.11: Per-Class Metrics for DINOv2 Fine-tuned Configuration

Metric	Class 0	Class 1	Class 2	Overall
Precision	74.77%	36.96%	64.08%	-
Recall	68.38%	36.96%	70.97%	-
F1-Score	71.43%	36.96%	67.35%	58.58% (macro)
Accuracy	-	-	-	63.67%

Table 5.12: Confusion Matrix for DINOv2 Fine-tuned Configuration

Predicted\True	Class 0	Class 1	Class 2
Class 0	80	15	22
Class 1	14	17	15
Class 2	13	14	66

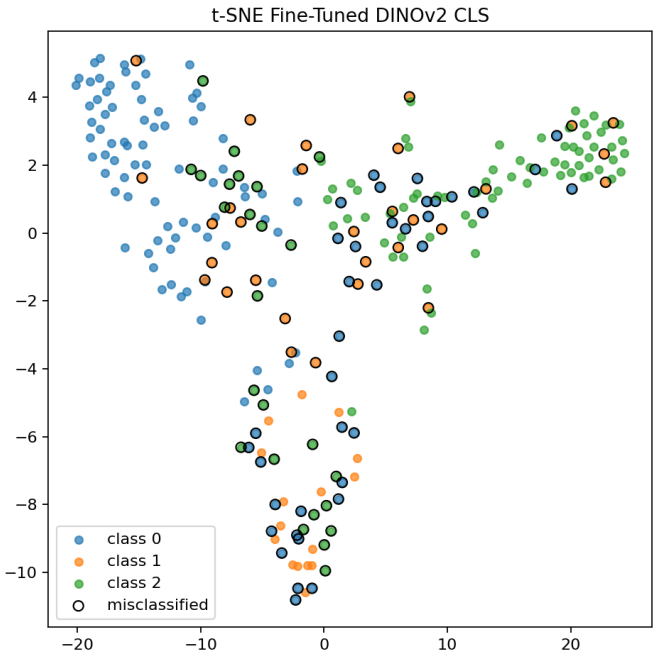


Figure 5.11: t-SNE visualization for DINOv2 Fine-tuned

5.6.7 Comparative Performance Analysis

Follows a comparison of the results of the six learning configurations in standardized metrics, highlighting architectural strengths and the impact of progressive unfreezing strategies on EGGIM classification performance.

Table 5.13: Comprehensive Performance Comparison Across All Configurations

Configuration	Class 0 F1	Class 1 F1	Class 2 F1	Macro F1	Accuracy
ResNet-50 Pre	66.14%	12.90%	61.31%	46.78%	57.81%
ResNet-50 FT	71.54%	12.70%	63.49%	49.24%	61.33%
ConvNeXt Pre	**76.56%**	6.90%	69.70%	51.05%	66.02%
ConvNeXt FT	76.12%	13.56%	**70.27%**	53.32%	**66.80%**
DINOv2 Pre	71.60%	9.23%	66.32%	49.05%	61.72%
DINOv2 FT	71.43%	**36.96%**	67.35%	**58.58%**	63.67%

Table 5.14: Per-Class Precision Comparison Across All Configurations

Configuration	Class 0	Class 1	Class 2	Avg Class 0+2	Avg Class 1+2	Class 1 Precision
ResNet-50 Pre	61.94%	25.00%	57.55%	59.75%	41.28%	25.00%
ResNet-50 FT	65.03%	23.53%	62.50%	63.77%	43.02%	23.53%
ConvNeXt Pre	**70.50%**	16.67%	65.71%	68.11%	41.19%	16.67%
ConvNeXt FT	67.55%	30.77%	**70.65%**	**69.10%**	**50.71%**	30.77%
DINOv2 Pre	65.71%	15.79%	64.95%	65.33%	40.37%	15.79%
DINOv2 FT	**74.77%**	**36.96%**	64.08%	69.43%	**50.52%**	**36.96%**

Table 5.15: Per-Class Recall Comparison Across All Configurations

Configuration	Class 0	Class 1	Class 2	Avg Class 0+2	Avg Class 1+2	Class 1 Recall
ResNet-50 Pre	70.94%	8.70%	65.59%	68.27%	37.15%	8.70%
ResNet-50 FT	79.49%	8.70%	64.52%	72.01%	36.61%	8.70%
ConvNeXt Pre	83.76%	4.35%	**74.19%**	**78.98%**	39.27%	4.35%
ConvNeXt FT	**87.18%**	8.70%	69.89%	78.54%	39.30%	8.70%
DINOv2 Pre	78.63%	6.52%	67.74%	73.19%	37.13%	6.52%
DINOv2 FT	68.38%	**36.96%**	70.97%	69.68%	**53.97%**	**36.96%**

Table 5.16: Confusion Matrix Summary - True Positives by Configuration

Configuration	Class 0 TP	Class 1 TP	Class 2 TP	Total TP	Accuracy
ResNet-50 Pre	83	4	61	148	57.81%
ResNet-50 FT	93	4	60	157	61.33%
ConvNeXt Pre	98	2	**69**	169	66.02%
ConvNeXt FT	**102**	4	65	**171**	**66.80%**
DINOv2 Pre	92	3	63	158	61.72%
DINOv2 FT	80	**17**	66	163	63.67%

5.7. Discussion

The main goal of this research was to systematically evaluate the potential of leveraging Foundation Models for the classification of precancerous gastric lesions within the EGGIM framework. By conducting a comparative analysis of three architectural paradigms in both their pre-trained and fine-tuned configurations, ResNet-50 (traditional Convolutional Neural Networks), ConvNeXt (hybrid CNN-transformer), and DINOv2 (self-supervised Vision Transformers), this research provides valuable insights into the strengths and limitations of FMs approaches in endoscopic imaging applications.

5.7.1 Essential Performance Indicators

The experimental results demonstrate a clear hierarchy in model performance that both supports and questions the initial assumptions about the dominance of FMs. DINOv2 Fine-tuned emerged as the top performer, achieving a macro F1-score of 58.58%, which is a 25% improvement compared to the baseline ResNet-50 Pre-trained model score of 46.78%. This result supports the hypothesis that self-supervised FMs, when properly adapted through progressive fine-tuning, can develop robust representations that generalize better to specialized endoscopic and medical imaging tasks [7].

The findings reveal a more complex scenario when assessing specific performance metrics: ConvNeXt setups reached the highest overall accuracy rates (66.80% for those fine-tuned and 66.02% for those pre-trained). This outcome implies that the hybrid CNN-transformer architecture excels in accurately classifying the majority of samples, primarily due to the outstanding identification of Class 0 (normal mucosa) and Class 2 (advanced metaplasia). In these categories, the ConvNeXt pre-trained model attained 83.76% and 74.19%, respectively, the highest scores among all setups. Nonetheless, this advantage comes with a notable drawback: the ConvNeXt Pre-trained model identified only 4.35% of Class 1 cases, representing the lowest sensitivity among all the models.

This trade-off highlights a fundamental issue in medical AI: aiming for overall accuracy can sometimes overlook the most crucial clinical cases. Within the EGGIM classification, Class 1 (growing metaplasia up to 30% IM) represents the intervention window where early detection has the most significant impact on patient outcomes [2]. DINOv2 Fine-tuned achieved a 36.96% Class 1 F1-score, reflecting a perfect balance between precision and recall, marking a 171% improvement over ConvNeXt Fine-tuned (13.56%) and

a 435% improvement over ConvNeXt Pre-trained (6.90%), which indicates that FMs pre-trained through self-supervised learning develop feature representations that are better at recognizing subtle pathological patterns in underrepresented classes.

The performance gap between pre-trained and fine-tuned configurations varied significantly across different architectures, revealing insights into how different pre-training paradigms interact with task-specific adaptation. The ResNet-50 architecture exhibited a modest increase of 2.46 percentage points in macro F1, while ConvNeXt showed even smaller improvements of 2.27 points, suggesting that supervised pre-training on ImageNet already offers features that are fairly adaptable for the medical imaging task at hand. On the other hand, DINOv2 showed the most substantial fine-tuning advantage with an increase of 9.53 points, highlighting that self-supervised FMs, although adept at learning versatile general-purpose representations, still require extensive domain-specific adaptation to excel in specialized medical classification tasks [77].

These results were obtained using base model checkpoints from all three architectures, in the same fixed seed and model's last convolutional layer size (2048 channels), representing the smallest configurations available for each paradigm. This approach facilitated a fair computational comparison and set baseline performance expectations. The results indicate that even smaller variants of FMs can provide significant performance enhancements over traditional architectures, although the extent of these benefits varies greatly depending on the pretraining method and class distribution characteristics.

5.7.2 Critical Analysis of Class 1 and t-SNE Insights

The severe underperformance of most configurations in Class 1 detection warrants a detailed examination, as it represents both the primary technical challenge and the most clinically consequential limitation of this study. With only 46 Class 1 samples in the test set (compared to 117 Class 0 and 93 Class 2), the dataset reflects natural intermediate IM patterns, but creates an extreme imbalance scenario in which Class 1 comprises merely 18% of the testing data [2].

Traditional metrics highlight the magnitude of this issue: a mere 8.70% Class 1 recall were achieved by the ResNet-50 configurations, correctly identifying only 4 of 46 cases of growing metaplasia. ConvNeXt Pre-trained performed even worse, with a recall of just 4.35% recall (2 correct detections), while ConvNeXt Fine-tuned equaled ResNet-50 at 8.70%. These dismal figures suggest that standard supervised learning techniques,

even when paired with class-weighted loss functions and oversampling tactics, struggle fundamentally when minority class prevalence falls below certain thresholds.

DINOv2 Fine-tuned's notable progress to 36.96% recall for Class 1 (17 of 46 correct detections) deserves detailed examination: this setup used a combination of techniques: focal loss with $\gamma = 2.0$ to emphasize hard examples, tripled oversampling of Class 1 in the training sampler, a 30% weight increase for Class 1 in the loss function, and intensive MixUp/CutMix augmentation in the early training epochs [79]. However, strategies of similar density were applied in other setups without achieving similar Class 1 performance, indicating that the self-supervised FMs' learned representations inherently aid in detecting rare patterns.

The t-SNE visualizations offer further clinical evidence supporting both conclusions: In every setup, Class 1 samples show greater intraclass variance compared to Classes 0 and 2, suggesting that GIM presents a more diverse range of visual features than normal tissue or more advanced lesions. This heterogeneity could support the notion that intermediate-stage precancerous lesions show different morphological characteristics depending on location, extent, and progression dynamics [3], and, in that regards, Foundation Models broader feature space may better accommodate this variability than narrowly specialized architectures.

However, even DINOv2's performance leaves 63% of Class 1 cases undetected, raising fundamental questions about whether current approaches can achieve clinically acceptable sensitivity without substantially larger and more balanced training datasets: the confusion matrix reveals that 15 of the 29 (51.7%) misclassified Class 1 cases were predicted as Class 0 (normal), representing a possible critical failure where precancerous lesions are dismissed as healthy tissue, aggravating the patient situation to a point where in a short future it would be too late to effectively treating the disease.

5.7.3 Architecture Analysis and Clinical Deployment

A comparison of architectural designs highlights distinct performance profiles with specific clinical utility scenarios. ConvNeXt demonstrates outstanding performance in detecting Class 0 and Class 2, achieving an average recall of 78.98% in its pre-trained form, making it an effective tool for screening and ruling out conditions where high detection sensitivity to normal tissues and advanced metaplasia is crucial for initial classification processes. In a two-step diagnostic process (metaplasia / non-metaplasia), ConvNeXt has the potential

to filter images that require expert review, thus reducing clinical workload by accurately classifying definitive positive and negative cases, while sending ambiguous cases to specialists. Unfortunately, ConvNeXt's systematic failure in Class 1 detection fundamentally limits its use in comprehensive EGGIM stratification scenarios. With a pre-trained recall of only 4.35%, it would fail to detect 95% of growing metaplasia cases, missing the precise intervention window where early detection most benefits patients [1]. Even when fine-tuned, the model's recall improves only to 8.70% recall remains unacceptably low, as it would overlook 91% of these critical lesions.

On the other hand, DINOv2 Fine-tuned offers a complementary performance profile, with balanced multi-class detection being its biggest strength as it performs significantly better on minority classes. The 36.96% recall for Class 1, although not very high, marks a notable improvement in detection capability on such a small and unbalanced endoscopic clinical classification dataset of 3 classes like EGGIM. This model, integrated into a hybrid human-AI workflow and with more depth and adjustment to the architecture, has some potential to pinpoint suspicious areas in Class 1-like cases, accelerating expert attention to the most time-sensitive diagnostic situations [1], although the significant computational costs that FMs pose may require more inference time and memory than CNN-based architectures [5].

5.7.4 Technical Insights

This research implemented technicalities that contribute to form generalizable insights for AI in endoscopic imaging:

- **Unfreezing and Feature Adaptability**: Both convolutional architectures opted for a more progressive unfreezing strategy, standing by the premise that effective learning strategies critically depend on the pre-training paradigm. ResNet-50's ImageNet initialization benefited from gradual layer-by-layer adaptation (unfreezing layer4/stage 5 at epoch 9, layer3/stage 4 at epoch 15, layer2/stage 3 at epoch 21), preventing catastrophic forgetting of edge and texture features while allowing semantic adaptation [75]. In contrast, DINOv2 achieved a better overall performance with aggressive unfreezing (last 4 transformer layers after only 4 head-only epochs), which could indicate that self-supervised representations exhibit greater domain-shift robustness than more traditional approaches. This observation is also consistent with the actual

DINOv2 publication by Oquab et al., which poses that models with contrastive learning approaches like Foundation Models develop more transferable features and are more robust than supervised classification models [77].

- **Augmentation and its Benefits**: The application of augmentation strategies could be sorted into two distinct cases. Advanced augmentation techniques, such as MixUp and CutMix, improved macro F1 by approximately 2%, showing utility for small medical datasets where the diversity of the training dataset is the key to more robust models. However, benefits plateaued beyond Epoch 16, suggesting that synthetic sample generation, in these small clinic dataset scenarios, reach diminishing returns as models overfit to enhancement patterns rather than learning genuine pathological features [79]. The more traditional approaches were more discrete in training setups, but while on testing configurations, methods like TTA through D4 dihedral transformations provided consistent 1 to 3 percentage points with greater gains in minority classes, which could suggest that the selective application of TTA in uncertain cases is more practical than the universal application for conturbed data environments such as those for clinical deployment.

- **Class Imbalance is Difficult**: The severe class imbalance (43.7% Class 0, 19.2% Class 1, 36.9% Class 2) of this EGGIM dataset stress-tested mitigation intentions and strategies, despite multiple countermeasures such as weighted losses, oversampling, focal loss, custom sampling, etc. The best accuracy results being a densely tuned ConvNeXt as simple 66.8%, with the addition that all configurations were unable to get satisfactory results in Class 1 detection, the most crucial clinic class to be extracted, ultimately could affect the strength of the conclusions obtained. Still, it should be taken into consideration the use of base models for each of the 3 architectures explored in this analysis.

- **Focal Loss Effective on FMs**: Between a diversity of loss functions experimented, Focal Loss in DINOv2 Fine-tuned ($\gamma = 2.0$, reducing the weight of easy examples by 4x) demonstrated to be the most distinctly effective strategy: by concentrating on challenging misclassified cases instead of frequently learning from numerous easy ones, focal loss seems especially complementary to the fine-tuning of FMs. The blend of focal loss emphasis and DINOv2's rich broad representations could foster a positive cycle overall, wherein the model gradually refines specialized features for difficult minority-class patterns.

- **Ensembling boosts Stability, not Fairness**: Ensemble experiments showed that top-5 averaging improved checkpoints enhanced the accuracy of ResNet-50 Fine-tuned by 1.2 points, it barely affected Class 1 detection. Multiscale ensembling, as in ConvNeXt Fine-Tuned, resulted in moderate overall gains but failed to solve the Class 1 issue, indicating that ensemble approaches primarily lessen variance in well-represented classes rather than tackling biases against minority classes.

5.7.5 Foundation Models: Valences and Broader Applications

Foundation Models like DINOv2 reveal superior capability in learning representations for identifying rare patterns in imbalanced medical imaging datasets than their convolutional network counterparts, with the FMs' significant improvement in Class 1 F1-score highlighting the advantage of pre-training on 142 million diverse images, which provides robust feature foundation for subtle pattern discrimination [77]. The effectiveness of self-supervised learning is notable because, unlike supervised approaches, it encourages diverse feature learning without task-specific biases, thus facilitating better adaptation to differences between the domains of the natural and medical image domains [7].

In contrast, considerable limitations restrain enthusiasm for the universal use of these systems in a clinical context, as the computational demands of ViTs architectures present real-world limitations for clinical application: DINOv2's 12-layer transformer with 85 million parameters requires substantially more GPU memory than ResNet-50's 25 million-parameter convolutional backbone, potentially limiting deployment on hospital hardware with limited resources [34]. Although this study did not conduct detailed measurements of inference times, transformer-based models generally took about three times longer to run compared to simpler CNNs models like ResNet-50, which could pose a challenge in high-volume endoscopic screening workflows where processing speed is critical [5].

The results from this EGGIM classification study provide indicators useful for broader GI imaging issues regarding more refined transformer-based structures like FMs, as class imbalance patterns, where intermediate-stage lesions are the hardest to detect, are likely applicable to other precancerous classifications: for example, in colorectal polyp detection, low-grade dysplasia is similarly an intervention-critical but underrepresented category, indicating that FMs could offer similar advantages for colonoscopy AI [71]. In high-throughput screening workflows that prioritize computational efficiency, hybrid CNN-transformer models such as ConvNeXt could show robust performance with manageable

inference costs, while in diagnostic contexts requiring high sensitivity to rare patterns, self-supervised FMs for EGC detection are more suitable due to their computational demands by improving the detection of minority classes [8].

In addition, techniques such as progressive unfreezing and discriminative learning rate strategies, which are both established in transfer learning, were also effective in this challenging domain. As previously determined, self-supervised models like DINOv2 can handle more aggressive unfreezing compared to their supervised counterparts, allowing for potentially faster and more robust adaptation, and although these approaches are not necessarily new, their effectiveness has been validated in relation to this FMs and small, unbalanced endoscopic datasets, as demonstrated by these results [5, 7].

From a patient safety perspective, a 63% Class 1 false negative rate poses significant risks, as a screening tool that fails to detect almost two-thirds of GIM could provide false assurance, delay necessary interventions, and lead to poorer patient outcomes. This grim reality suggests that although the current base Foundation Models methods have potential, they need considerable advancement through larger datasets, more refined or innovative architectures, or a combination of human-AI systems before they are ready for clinical use in these specifically challenging EGGIM classification similar tasks [3].

This work systematically evaluated the possibility of Foundation Models to classify precancerous gastric lesions within the EGGIM framework, comparing the ResNet-50, ConvNeXt, and DINOv2 architectures. The fine-tuned DINOv2 consistently outperformed others in detecting minority cases, especially those with significant clinical relevance for early intervention, yet its results still underscore the significant detection gap that exists for small or rare category data. Although ConvNeXt demonstrated broad accuracy, it struggled with the subtlety and diversity of intermediate-stage lesions. The study used methodical benchmarking, t-SNE analysis, and technical indicators, ranging from discriminative learning rates to targeted enhancements, to show that while FMs offer strong representational power and hold strong promise, they do not overcome the core limitations imposed by these clinical dataset constraints (particularly small dataset size and severe underrepresentation of the most critical class). Ultimately, combining the strengths of these models with better data, innovative tuning methods, and robust learning procedures will be crucial to improve the aid of FMs in diagnosis in medical imaging [5, 7].

6. Conclusion

This chapter consolidates the research results and explores their impacts: initially, it integrates significant contributions to medical AI applications in gastric lesion classification, then addresses the shortcomings of existing frameworks, and finally proposes resumed future practical research paths based on the limitations identified in this study.

6.1. Summary of Contributions

This research makes several concrete contributions to the intersection of Foundation Models and medical imaging, particularly in the context of Gastrointestinal endoscopy and Endoscopic Grading of Gastric Intestinal Metaplasia (EGGIM) classification:

First Systematic Evaluation of Foundation Models for three-class stratification of Intestinal Metaplasia

As far as can be determined from the literature, this work represents the first systematic review of self-supervised FMs specifically targeting the classification of precancerous gastric lesions in endoscopic images (using the EGGIM dataset). Previous work by Kerdegari et al. explored FMs in the context of gastric inflammation in a broader sense, whereas this study focuses explicitly on the specific three-class EGGIM stratification problem (normal mucosa, growing metaplasia, advanced metaplasia), which has direct implications for clinical surveillance protocols and gastric cancer risk assessment [7]. By establishing baseline performance metrics across six unique configurations, this work provides a reference benchmark for future research in this domain.

Architectural Insights from Comparative Analysis

A comparative analysis highlights distinct performance profiles in different architectures, which have significant clinical implications. The ConvNeXt configurations achieved the highest overall accuracy at 66.80%), particularly excelling in detecting Class 0 (normal mucosa) and Class 2 (advanced metaplasia), with recall rates of up to 83.76% and 74.19% respectively. However, this success was offset by a significant shortfall in the detection of Class 1 (growing metaplasia), where the ConvNeXt Pre-trained model only managed a 4.35% recall, correctly identifying just 2 of 46 growing metaplasia cases. In contrast, the DINOv2 Fine-tuned model demonstrated a more balanced performance by achieving a 36.96% recall for Class 1 while maintaining competitive performance in other classes,

representing a 435% improvement over ConvNeXt Pre-trained for the most clinically critical category, which supports the hypothesis that self-supervised feature mapping learns feature representations more effectively for detecting subtle pathological patterns in underrepresented classes [77].

The performance gap between pre-trained and fine-tuned configurations varied considerably across architectures: ResNet-50 showed a modest improvement of 2.46 percentage points in macro F1, ConvNeXt saw an even smaller gain of 2.27 points, while DINOv2 exhibited the most significant increase with a gain of 9.53 points, implying that while self-supervised pre-training creates representations that require more extensive domain adaptation, they can ultimately achieve superior detection in minority classes when properly fine-tuned [7, 77].

Empirical Evaluation for Small and Imbalanced Medical Datasets

This study systematically evaluated and documented a detailed training methodology that combines established techniques to address the issues associated with small and severely unbalanced medical imaging classification datasets. The methodology integrates progressive unfreezing schedules (gradual for supervised models, aggressive for self-supervised), discriminative learning rates (head LR 10-100 times larger than the backbone), targeted augmentation (MixUp/CutMix, test-time augmentation), and strategies for addressing class imbalance mitigation (focal loss, weighted sampling, oversampling). While these techniques are individually well-established, the unique value of this analysis lies in key observations such as the fact that self-supervised FMs can endure significantly more aggressive unfreezing than their supervised counterparts, a comparative insight that provides practical guidance for future medical AI development when adapting FMs to specialized clinical tasks, such as those in endoscopic imaging [7, 77].

Limitations and Assessment of Current Capabilities

It is essential to assess the limitations of the current base FMs for clinical deployment. The best performing configuration, DINOv2 Fine-tuned, managed only a 36.96% recall rate for Class 1, resulting in 63% of cases of growing metaplasia going undetected. Of the 29 misclassified Class 1 cases, 15 (51.7%) were incorrectly identified as normal mucosa, representing a significant failure mode where precancerous lesions are dismissed as healthy tissue. This stark reality highlights that although the current base FMs approaches show promise, they are not yet sufficiently developed for clinical deployment

readiness for EGGIM classification without significant further development [3].

The training time and computational cost also represent significant practical limitations: the transformer architecture of DINOv2 exhibited substantially longer per-epoch training times, approximately three times longer than the ResNet-50's convolutional backbone, based on observational comparison during training; moreover, DINOv2 required more epochs to reach convergence (250 compared to 100). In exploratory medical AI research where computational budgets are limited, these extended training requirements restrict the number of hyperparameter configurations and ablation studies within the project timeline [5].

Visual Analysis through t-SNE Embeddings

The t-SNE visualizations offered crucial insights into the feature representations that were learned, revealing that, across all configurations, samples from Class 1 showed much higher intra-class variance than Classes 0 and 2, suggesting that growing metaplasia manifests diverse visual characteristics depending on its location, extent, and progression dynamics. DINOv2 Fine-tuned's embedding space demonstrated that Class 1 samples formed more cohesive clusters with less overlap, unlike ConvNeXt and ResNet-50, where Class 1 samples were more scattered and intermixed with other classes, which supports the hypothesis that FMs' broader feature space, developed from 142 million diverse images during pre-training, is better suited to represent the morphological variability seen in intermediate-stage precancerous lesions [77].

6.2. Future Research Directions

Future investigations should focus on domain-specific FMs pre-trained in large-scale gastric endoscopy datasets to capture mucosal textures and vascular patterns relevant to metaplasia detection [8], while vision-language FMs integrating endoscopic images with clinical metadata like H. pylori status could leverage contextual information used in clinical procedures [5]. Exploring larger model variants such as ConvNeXt-large or DINOv2-large, limited by computational resources in this work, may reveal different performance trade-offs and enhance minority class detection [77]. Expanding the dataset with diverse Class 1 samples is essential, as the 63% miss rate reflects inherent data limitations. Standardized benchmarking protocols emphasizing minority class recall and clear reporting of computational costs would accelerate progress toward deployable systems [74].

References

[1] Y. Li, C. Xu *et al.*, "Gastric intestinal metaplasia and cancer risk stratification: focus on eggim and olgim systems," *Cancer Cell International*, vol. 21, pp. 1–12, 2021. [Online]. Available: https://cancerci.biomedcentral.com/articles/10.1186/s12935-021-01758-6 [Cited on pages 1, 15, and 84.]

[2] P. Pimentel-Nunes, D. Libânio, J. Lage, D. Abrantes, M. Marques, and M. Dinis-Ribeiro, "A multicenter prospective study of the endoscopic grading of gastric intestinal metaplasia (eggim)," *Endoscopy*, vol. 48, no. 10, pp. 963–969, 2016. [Online]. Available: https://www.thieme-connect.de/products/ejournals/abstract/10.1055/s-0042-108435 [Cited on pages 1, 2, 15, 16, 81, and 82.]

[3] ——, "Endoscopic grading of gastric intestinal metaplasia: multicenter validation of the eggim score," *Endoscopy*, vol. 51, no. 6, pp. 515–521, 2019. [Online]. Available: https://www.thieme-connect.com/products/ejournals/html/10.1055/a-0808-3186 [Cited on pages ix, 1, 15, 16, 83, 87, and 90.]

[4] M. Kim, J. Yun, Y. Cho, K. Shin, R. Jang, H.-j. Bae, and N. Kim, "Deep learning in medical imaging," *Neurospine*, vol. 16, no. 4, pp. 657–668, Dec. 2019. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC6945006/ [Cited on pages 2, 18, 19, 22, 23, and 25.]

[5] T. L. Veldhuizen *et al.*, "Foundation models in medical imaging - a review and outlook," *arXiv preprint*, 2025. [Online]. Available: https://arxiv.org/pdf/2506.09095v1 [Cited on pages 2, 28, 29, 30, 31, 84, 86, 87, and 90.]

[6] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," 2022. [Online]. Available: https://arxiv.org/abs/2201.03545 [Cited on pages 2, 52, and 53.]

[7] H. Kerdegari, K. Higgins, D. Veselkov, I. Laponogov, I. Polaka, M. Coimbra, J. Pescino, M. Leja, M. Dinis-Ribeiro, T. Fleitas, and K. Veselkov, "Foundational models for pathology and endoscopy images: Application for gastric inflammation," *Diagnostics*, vol. 14, p. 1912, 08 2024. [Cited on pages ix, 2, 13, 19, 21, 26, 27, 29, 31, 32, 43, 55, 81, 86, 87, 88, and 89.]

[8] J. Dermyer *et al.*, “Endodino: A foundation model for gi endoscopy,” *arXiv preprint*, 2025. [Online]. Available: https://arxiv.org/pdf/2501.05488 [Cited on pages 2, 31, 32, 42, 87, and 90.]

[9] J. Ferlay, M. Ervik, F. Lam, M. Laversanne, M. Colombet, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram, and F. Bray, “Global cancer observatory: Cancer today. world fact sheet 2022 (version 1.1),” *International Agency for Research on Cancer*, pp. 1–2, Feb. 2024. [Online]. Available: https://gco.iarc.who.int/today [Cited on pages 5 and 7.]

[10] J. S. Tobias and D. Hochhauser, *Cancer and its Management*. London: John Wiley & Sons, 2004, historical trends of gastric cancer discussed on p. 259. [Cited on page 5.]

[11] G. Menon, S. El-Nakeep, and H. M. Babiker, “Gastric cancer,” *StatPearls [Internet]*, p. –, Oct. 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK459142/ [Cited on pages 5, 6, 7, 8, 9, 10, 11, 12, and 13.]

[12] M. B. Piazuelo, F. Carneiro, and M. C. Camargo, “Considerations in comparing intestinal- and diffuse-type gastric adenocarcinomas,” *Helicobacter*, vol. 28, no. 3, p. e12975, Jun. 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/36965033/ [Cited on pages ix and 6.]

[13] National Cancer Institute, “Gastric cancer treatment (pdq®)–health professional version,” https://www.cancer.gov/types/stomach/hp/stomach-treatment-pdq, 2025, provides statistics on gastric adenocarcinoma prevalence, management and therapies. [Cited on pages 6, 8, 12, and 13.]

[14] J. A. Ajani, T. A. D’Amico, D. J. Bentrem, J. Chao, D. Cooke, C. Corvera, P. Das, P. C. Enzinger, T. Enzler, P. Fanta, F. Farjah, H. Gerdes, M. K. Gibson, S. Hochwald, W. L. Hofstetter, D. H. Ilson, R. N. Keswani, S. Kim, L. R. Kleinberg, S. J. Klempner, J. Lacy, Q. P. Ly, K. A. Matkowskyj, M. McNamara, M. F. Mulcahy, D. Outlaw, H. Park, K. A. Perry, J. Pimiento, G. A. Poultsides, S. Reznik, R. E. Roses, V. E. Strong, S. Su, H. L. Wang, G. Wiesner, C. G. Willett, D. Yakoub, H. Yoon, N. McMillian, and L. A. Pluchino, “Gastric cancer, version 2.2022, nccn clinical practice guidelines in oncology,” *Journal of the National Comprehensive Cancer Network*, vol. 20, no. 2, pp. 167–192, 2022, comprehensive source covering anatomy, risk factors, staging, and management of gastric cancer.

[Online]. Available: https://jnccn.org/view/journals/jnccn/20/2/article-p167.xml [Cited on page 6.]

[15] National Comprehensive Cancer Network, “Nccn clinical practice guidelines in oncology: Gastric cancer,” *JNCCN*, vol. 20, no. 2, pp. 167–200, 2022, comprehensive source for anatomical distribution, risk factors, and management. [Online]. Available: https://jnccn.org/view/journals/jnccn/20/2/article-p167.xml [Cited on pages 7, 8, 10, 12, and 14.]

[16] International Agency for Research on Cancer, “Age-standardized rate (world) per 100,000, incidence, both sexes, stomach cancer, 2022,” https://gco.iarc.who.int/today, 2024, data version: Globocan 2022 (Version 1.1), updated 08 Feb 2024. Image reproduced with permission from Cancer TODAY, IARC/WHO. [Cited on pages ix and 8.]

[17] C. Hamashima, “Current issues and future perspectives of gastric cancer screening,” *World Journal of Gastroenterology*, vol. 20, no. 38, pp. 13 767–13 774, Oct. 2014, review of screening methods (endoscopy, serology, screen-and-treat), effectiveness, and implementation phases. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4194560/ [Cited on pages 7, 10, and 14.]

[18] K.-P. Ko, “Risk factors of gastric cancer and lifestyle modification for prevention,” *Journal of Gastric Cancer*, vol. 24, no. 1, pp. 99–107, 2024, open-access review of risk factors (H. pylori, diet, smoking, obesity, alcohol) and prevention strategies. [Online]. Available: https://jgc-online.org/DOIx.php?id=10.5230/jgc.2024.24.e10 [Cited on page 9.]

[19] Stanford Health Care, “Stomach cancer diagram – stomach,” https://stanfordhealthcare.org/content/dam/SHC/conditions/cancer/images/stomach-cancer-diagram-stomach.gif, 2025, educational resource on gastric cancer anatomy and spread. Accessed 2025-08-25. [Cited on pages ix and 11.]

[20] IHH Healthcare Berhad, “Gastric cancer educational graphic,” https://cdn-assets-eu.frontify.com/s3/frontify-enterprise-files-eu/..., 2025, educational infographic provided by IHH Healthcare Berhad. Accessed 2025-08-25. [Cited on pages ix and 11.]

[21] R. Ahlawat *et al.*, “Esophagogastroduodenoscopy (egd),” StatPearls [Internet], 2023, comprehensive overview of patient preparation, sedation, and endoscopic

technique; accessed via NCBI Bookshelf. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK532268/ [Cited on pages 12 and 13.]

[22] Cleveland Clinic, “Egd procedure (upper endoscopy): What it is & what to expect,” Cleveland Clinic Health Library, 2025, patient-oriented overview of upper endoscopy including procedure steps and rationale. [Online]. Available: https://my.clevelandclinic.org/health/procedures/22549-egd-procedure-upper-endoscopy [Cited on pages 12 and 13.]

[23] Mayo Clinic, “Upper endoscopy–about the test,” Mayo Clinic Health Library, 2024, description of upper GI endoscopy, indications, and procedural overview. [Online]. Available: https://www.mayoclinic.org/tests-procedures/endoscopy/about/pac-20395197 [Cited on page 13.]

[24] P. Lakatos and I. Hritz, *Atrophic Gastritis*. Treasure Island (FL): StatPearls [Internet], 2023, available from NCBI Bookshelf. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK563275/ [Cited on page 14.]

[25] N. Uedo, R. Ishihara, H. Iishi, S. Yamamoto, S. Yamamoto, T. Yamada, K. Imanaka, Y. Takeuchi, K. Higashino, S. Ishiguro, and M. Tatsuta, “A new method of diagnosing gastric intestinal metaplasia: narrow-band imaging with magnifying endoscopy,” *Endoscopy*, vol. 38, no. 8, pp. 819–824, 2006. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/17001572/ [Cited on page 14.]

[26] Y. Sano, R. Lambert, and J.-F. Rey, “Paris classification of superficial neoplastic lesions: Esophagus, stomach, and colon,” *Gastrointestinal Endoscopy*, vol. 58, no. 6 Suppl, pp. S3–S43, 2003. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8750452/ [Cited on page 14.]

[27] R. M. Zagari, M. Dinis-Ribeiro *et al.*, “Olga and olgim staging systems for gastritis and gastric cancer risk assessment,” *World Journal of Gastroenterology*, vol. 30, no. 21, pp. 2341–2354, 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11915242/ [Cited on page 14.]

[28] G. P. Veldhuizen, C. Ingmar, J. C. Rijn *et al.*, “Deep learning-based subtyping of gastric cancer histology predicts clinical outcome: a multi-institutional retrospective study,” *Gastric Cancer*, vol. 26, no. 5, pp. 795–806, Sep. 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10361890/ [Cited on pages 17 and 19.]

[29] IBM Corporation, “What is deep learning?” IBM Think Topics, Jun. 2024, online resource defining deep learning concepts and applications. [Online]. Available: https://www.ibm.com/think/topics/deep-learning [Cited on pages 18, 19, 25, and 26.]

[30] Viso.ai. (2025) Computer vision tasks (comprehensive 2025 guide). Viso.ai. DL Families Overview. [Online]. Available: https://viso.ai/deep-learning/computer-vision-tasks/ [Cited on pages ix, 18, 19, 22, 23, 24, and 25.]

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255. [Cited on pages 19, 27, 50, and 52.]

[32] C. C. Chatterjee. (2019, July) Basics of the classic CNN. Accessed: 2025-9. [Online]. Available: https://medium.com/data-science/basics-of-the-classic-cnn-a3dce1225add [Cited on pages ix and 20.]

[33] A. Teramoto, “Detection and characterization of gastric cancer using cascade deep learning model in endoscopic images,” *Diagnostics*, 2022. [Cited on pages 20, 22, 36, 37, 38, and 39.]

[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,” *arXiv preprint*, 2020, viTs. [Online]. Available: https://arxiv.org/abs/2010.11929 [Cited on pages 20, 21, and 86.]

[35] Viso.ai. (2025) Vision transformers (vit) in image recognition. Viso.ai. ViTs. [Online]. Available: https://viso.ai/deep-learning/vision-transformer-vit/ [Cited on page 20.]

[36] GeeksforGeeks. (2025, July) Vision transformer (ViT) architecture. Accessed: 2025-9. [Online]. Available: https://www.geeksforgeeks.org/deep-learning/vision-transformer-vit-architecture/ [Cited on pages ix and 21.]

[37] G. Ayana, H. Barki, and S.-W. Choe, “Pathological insights: Enhanced vision transformers for the early detection of colorectal cancer,” *Cancers*, vol. 16, no. 7, p. 1441, 2024, viTs Classification Example. [Online]. Available: https://www.mdpi.com/2072-6694/16/7/1441 [Cited on page 21.]

[38] T. Liang *et al.*, “Mask r-cnn with r-d net: A detection and segmentation model for early gastric cancer gastroscopy images,” in *2023 3rd International Conference on Electronic Information Engineering and Computer Science (EIECS)*, 2023, pp. 1346–1352. [Cited on pages 22, 36, 37, 38, and 39.]

[39] Z.-H. Cui *et al.*, “Application of improved mask r-cnn algorithm based on gastroscopic image in detection of early gastric cancer,” in *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2022, pp. 1396–1401. [Cited on pages 22, 36, 37, 38, and 39.]

[40] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *arXiv preprint*, 2015. [Online]. Available: https://arxiv.org/pdf/1506.01497 [Cited on page 22.]

[41] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” *arXiv preprint*, 2016. [Online]. Available: https://arxiv.org/abs/1506.02640 [Cited on page 22.]

[42] J. Wan, W. Zhu, B. Chen, L. Wang, K. Chang, and X. Meng, “Crh-yolo for precise and efficient detection of gastrointestinal polyps,” *Scientific Reports*, vol. 14, no. 1, p. 30033, 2024. [Online]. Available: https://www.nature.com/articles/s41598-024-81842-9 [Cited on page 22.]

[43] ResearchGate. (2018) Basic architectures of the considered types of object detectors. Accessed: 2025-9. [Online]. Available: https://www.researchgate.net/figure/Basic-architectures-of-the-considered-types-of-object-detectors_fig1_327315674 [Cited on pages ix and 23.]

[44] D. He, “Dual-branch hybrid network for lesion segmentation in gastric cancer images,” *Scientific Reports*, 2023. [Cited on pages 23, 24, 36, 37, 38, and 39.]

[45] Y. Zhang, Y. Li, and D. He, “U-shaped network with multi-scale feature extraction for lesion region segmentation in gastric cancer images,” in *2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)*, vol. 3, 2023, pp. 234–238. [Cited on pages 23, 24, 36, 37, 38, and 39.]

[46] “From cnn to transformer: A review of medical image segmentation models,” *PMC*, 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11300773/ [Cited on pages ix and 24.]

[47] "Transformers in medical image segmentation: a narrative review," *PMC*, 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10722011/ [Cited on page 24.]

[48] R. Osuala *et al.*, "Medigan: A python library of pretrained generative models for medical image synthesis," *Journal of Medical Imaging*, vol. 10, no. 6, 2023. [Cited on pages 25, 36, 37, 38, and 39.]

[49] B. L. Kidder, "Advanced image generation for cancer using diffusion models," *Biology Methods and Protocols*, vol. 9, no. 1, p. bpae062, 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11387006/ [Cited on pages 25 and 26.]

[50] Y. Skandarani, P.-M. Jodoin, and A. Lalande, "Gans for medical image synthesis: An empirical study," *Journal of Imaging*, vol. 9, p. 69, 03 2023. [Cited on pages ix and 26.]

[51] C. R. Wolfe. (2023, October) Data is the foundation of language models. Accessed: 2025-9. [Online]. Available: https://medium.com/data-science/data-is-the-foundation-of-language-models-52e9f48c07f5 [Cited on pages x and 27.]

[52] TechTarget. (2024) Foundation models explained: Everything you need to know. TechTarget. [Online]. Available: https://www.techtarget.com/whatis/feature/Foundation-models-explained-Everything-you-need-to-know [Cited on pages 27 and 28.]

[53] Encord. (2023) Visual foundation models (VFMs) explained. Accessed: 2025-9. [Online]. Available: https://encord.com/blog/visual-foundation-models-vfms-explained/ [Cited on pages x and 29.]

[54] J. Wu, Y. Wang, Z. Zhong, W. Liao, N. Trayanova, Z. Jiao, and H. Bai, "Vision-language foundation model for 3d medical imaging," *npj Artificial Intelligence*, vol. 1, 08 2025. [Cited on pages x and 30.]

[55] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024. [Online]. Available: https://www.nature.com/articles/s41467-024-44824-z [Cited on pages 30 and 42.]

[56] T. Ferreira, D. Marques, M. Dinis-Ribeiro, and M. Coimbra, "A review of segmentation algorithms for gastrointestinal images: Datasets, annotations and architectures," 09 2025, pp. 5–8. [Cited on pages 33 and 102.]

[57] K. Zhang, “Early gastric cancer detection and lesion segmentation based on deep learning and gastroscopic images,” *Scientific Reports*, 2024. [Cited on pages 36, 37, 38, and 39.]

[58] Z. Li, “Mfa-net: Multiple feature association network for medical image segmentation,” *Computers in Biology and Medicine*, 2023. [Cited on pages 37, 38, and 39.]

[59] Y. Sun, “Lesion segmentation in gastroscopic images using generative adversarial networks,” *Journal of Digital Imaging*, 2022. [Cited on pages 37, 38, and 39.]

[60] W. Du, “Early gastric cancer segmentation in gastroscopic images using a co-spatial attention and channel attention based triple-branch resunet,” *Computer Methods and Programs in Biomedicine*, 2023. [Cited on pages 37, 38, and 39.]

[61] I. Ligato, “Convolutional neural network model for intestinal metaplasia recognition in gastric corpus using endoscopic image patches,” *Diagnostics*, 2024. [Cited on pages 37, 38, and 39.]

[62] Q. Chang, “Esfpnet: Efficient stage-wise feature pyramid on mix transformer for deep learning-based cancer analysis in endoscopic video,” *Journal of Imaging*, 2024. [Cited on pages 37, 38, and 39.]

[63] Q. Lin, W. Tan, S. Cai, B. Yan, J. Li, and Y. Zhong, “Lesion-decoupling-based segmentation with large-scale colon and esophageal datasets for early cancer diagnosis,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 8, pp. 11 142–11 156, 2024. [Cited on pages 37, 38, and 39.]

[64] P. Saxena and A. K. Bhandari, “Context aware automatic polyp segmentation network with mask attention,” *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 7, pp. 3510–3523, 2024. [Cited on page 39.]

[65] W. Li, Y. Liu, X. Liu, and Y. Qi, “Fusing transformer and fcn for polyp segmentation,” in *2023 7th Asian Conference on Artificial Intelligence Technology (ACAIT)*, 2023, pp. 623–628. [Cited on pages 37, 38, and 39.]

[66] N. D. Manh *et al.*, “Endounet: A unified model for anatomical site classification, lesion categorization and segmentation for upper gastrointestinal endoscopy,” in *2022 14th International Conference on Knowledge and Systems Engineering (KSE)*, 2022, pp. 1–6. [Cited on pages 37, 38, and 39.]

[67] N. V. Sai Manoj, M. Rithani, and R. S. SyamDev, “Enhancing gastric cancer diagnosis through ensemble learning for medical image analysis,” in *2023 Seventh International Conference on Image Information Processing (ICIIP)*, 2023, pp. 822–826. [Cited on pages 37, 38, and 39.]

[68] D. Katayama *et al.*, “Development of computer-aided diagnosis system using single fcn capable for indicating detailed inference results in colon nbi endoscopy,” in *2023 International Technical Conference on Circuits/Systems, Computers, and Communications (ITC-CSCC)*, 2023, pp. 1–6. [Cited on pages 37, 38, and 39.]

[69] Y.-T. Chen, E.-H. Yang, W.-L. Chang, J. Lin, H.-C. Cheng, and C.-R. Huang, “Mask focal modulation network for gastric intestinal metaplasia segmentation,” in *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024, pp. 1–8. [Cited on pages 37, 38, and 39.]

[70] A. Garbaz, Y. Oukdach, S. Charfi, M. El Ansari, L. Koutti, and M. Salihoun, “Bleeding segmentation based on a u-formed network with separable contextual feature-guided in wireless capsule endoscopy images,” in *2024 11th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, 2024, pp. 1–6. [Cited on pages 36, 38, and 39.]

[71] P. Jha, M. Smedsrud, D. Riegler, P. Halvorsen, T. de Lange, and H. D. Johansen, “Kvasir-seg: A segmented polyp dataset,” in *Proc. Int. Conf. Multimedia Modeling (MMM)*, Daejeon, Korea, 2020, pp. 451–462. [Cited on pages x, 38, 40, and 86.]

[72] S. Bernal, J. Sánchez, F. Vilarino *et al.*, “Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians,” *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015. [Cited on pages x, 38, and 40.]

[73] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, “Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, no. 2, pp. 283–293, 2014. [Cited on page 38.]

[74] G. Antonelli *et al.*, “Quaide—quality assessment of ai preclinical studies in diagnostic endoscopy,” *Gut*, pp. 1–9, 2024. [Cited on pages 41 and 90.]

[75] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2015. [Online]. Available: https://arxiv.org/abs/1512.03385 [Cited on pages 50, 52, and 84.]

[76] Microsoft, “Resnet-50 model [hugging face],” https://huggingface.co/microsoft/resnet-50. [Cited on pages 50 and 52.]

[77] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, “Dinov2: Learning robust visual features without supervision,” *arXiv preprint arXiv:2304.07193*, 2023. [Cited on pages 54, 55, 82, 85, 86, 89, and 90.]

[78] Facebook Research, “Dinov2-base model [hugging face],” https://huggingface.co/facebook/dinov2-base. [Cited on page 54.]

[79] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” 2018. [Online]. Available: https://arxiv.org/abs/1710.09412 [Cited on pages 59, 83, and 85.]

[80] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, “Cutmix: Regularization strategy to train strong classifiers with localizable features,” 2019. [Online]. Available: https://arxiv.org/abs/1905.04899 [Cited on page 59.]

[81] J. Mok. (2021) Cutout, mixup, and cutmix: Implementing modern image augmentations in pytorch. Accessed: September 2025. [Online]. Available: https://medium.com/data-science/cutout-mixup-and-cutmix-implementing-modern-image-augmentations-in-pytorch-a9d7db3074ad [Cited on pages x and 60.]

[82] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988. [Cited on page 60.]

[83] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826. [Cited on page 61.]

[84] J. Ren, C. Yu, X. Ma, H. Zhao, S. Yi *et al.*, “Balanced meta-softmax for long-tailed visual recognition,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 4175–4186. [Cited on page 61.]

[85] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, “Long-tail learning via logit adjustment,” in *International Conference on Learning Representations (ICLR)*, 2021. [Cited on page 62.]

[86] K. Erdem, “t-sne clearly explained,” Medium, April 2020, accessed: September 2025. [Online]. Available: https://medium.com/data-science/t-sne-clearly-explained-d84c537f53a [Cited on page 66.]

[87] “Portal of the 8th ieee portuguese meeting on bioengineering (enbeng 2025),” in *8th IEEE Portuguese Meeting on Bioengineering (ENBENG 2025)*, IEEE EMBS Portugal Chapter. University of Aveiro, Aveiro, Portugal: IEEE, September 2025. [Online]. Available: https://embs.ieee-pt.org/8th-enbeng-2025/ [Cited on page 102.]

Segmentation Algorithms in GI Imaging: A Systematic Review

The article was published as a contribution to the 8th IEEE Portuguese Meeting on Bioengineering (ENBENG 2025) conference [56, 87], and the respective presentation poster.

A Review of Segmentation Algorithms for Gastrointestinal Images: Datasets, Annotations and Architectures

Tomás Ferreira[1], David Marques[2], Mário Dinis-Ribeiro[3], Miguel Coimbra[4]

***Abstract*— Accurate segmentation of gastrointestinal (GI) images is critical for early detection of cancers and other pathologies, showing a need for the integration of computer-aided diagnostic systems in clinical practice. Despite recent advances in deep learning, research in this area remains fragmented, with limited datasets and diverse annotation strategies. To address this, we conducted a systematic review of GI image segmentation studies published between 2022 and 2024, focusing on three key dimensions: datasets, annotation practices, and model architectures. A total of 919 unique articles were retrieved from PubMed, Scopus, and IEEE Xplore, and screened through a multi-stage selection process involving title and abstract filtering, conflict resolution, and final full-text validation, resulting in 20 studies included for in-depth analysis.**

Our findings highlight the predominance of small and often publicly available datasets, with Kvasir-SEG and CVC-ClinicDB being the most used. Annotation practices varied significantly, with limited transparency regarding annotator expertise and agreement. Most studies favored custom-built deep learning architectures, often based on U-Net variants, while transformer-based and ensemble methods remained relatively rare. These findings emphasize the need for larger, more diverse datasets, standardized annotation protocols, and reproducible benchmarking frameworks to enable more robust and clinically relevant GI image segmentation models.

Keywords—gastrointestinal image segmentation, deep learning, medical imaging

I. Introduction

Gastrointestinal (GI) image segmentation plays a vital role in advancing diagnostic accuracy, particularly for early detection and treatment planning in diseases such as colorectal and gastric cancers. With the growing interest of computer-aided diagnostic tools, accurate and reliable segmentation algorithms are essential for clinical integration, namely in endoscopic workflows (Figure 1). Despite rapid advancements in deep learning, there remains a pressing need to systematically consolidate and evaluate the breadth of segmentation techniques applied to GI imaging.

Motivated by this gap, our work undertakes a rigorous literature review centered on a multi-stage, criteria-driven screening process, aiming to identify the most relevant contributions from the last two years to this topic. More specifically, our study aims to highlight the state of the art in dataset usage, annotation practices, and algorithmic design in GI image segmentation.

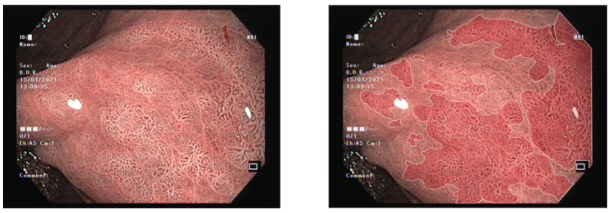


Figure 1 - Example result of a segmentation algorithm applied to a GI image.

To ensure relevance and quality, we conducted a structured database search (PubMed, Scopus, IEEE Xplore) for the years 2022–2024 using a carefully refined query and specific filters. The resulting 919 documents underwent a multi-step selection process, comprising title-based and abstract-based filtering with dual author review, conflict resolution, and final full-text validation, ultimately narrowing the set to 20 selected publications [1-20]. Through this methodology, we aim to map current trends and techniques in GI image segmentation, highlighting benchmark datasets, annotation workflows, and the popular deep learning architectures. Our goal is to provide researchers and clinicians with a cohesive view of the field's most impactful contributions and its current challenges.

II. Review Methodology

A. Literature Search Strategy

The process of constructing this review began with the formulation of a comprehensive search query designed to retrieve recent and relevant scientific literature on segmentation algorithms applied to GI medical imaging. The objective was to capture publications that aligned with the domains of medical image analysis, particularly in relation to cancers of the digestive tract, while ensuring coverage of modern techniques from computer science and engineering disciplines.

Three widely recognized scholarly databases were selected: PubMed, Scopus, and IEEE Xplore. The final query was optimized iteratively to balance sensitivity and specificity and was structured as follows

[1] Tomás Ferreira is the corresponding author and is with the INESC TEC and Faculty of Sciences of the University of Porto, Portugal (email: up202006594@edu.fc.up.pt)

[2] David Marques is a co-author and is with the INESC TEC and Faculty of Sciences of the University of Porto, Portugal (email: up201905574@edu.fc.up.pt)

[3] Mário Dinis-Ribeiro is a co-author and is with the Department of Gastroenterology, Instituto Português de Oncologia do Porto (IPO-Porto), Portugal (email: mario.ribeiro@ipoporto.min-saude.pt)

[4] Miguel Coimbra is a co-author and is with the INESC TEC and Faculty of Sciences of the University of Porto, Portugal (email: mcoimbra@fc.up.pt)

Segmentation AND medical AND image AND cancer AND (colon OR intestine OR gastric OR endoscopy OR 2D)

(Publication Date: 2022–2024 | Filters: Computer Science, Engineering, Medicine)

This search resulted in a total of 537 records from PubMed, 353 from Scopus, and 180 from IEEE Xplore, yielding 919 unique documents after deduplication. The deduplication and review process were facilitated using Rayyan, an AI-assisted platform tailored for systematic review screening and conflict resolution.

B. Inclusion and Exclusion Criteria

To refine the dataset, inclusion and exclusion criteria were defined a priori and applied across two filtering stages: title-based and abstract-based screening.

Inclusion Criteria:

- Keywords present: Segmentation, Medical Image, Gastric, Intestinal, Cancer, 2D, Endoscopy
- Focus on algorithmic or architectural contributions to image segmentation
- Studies involving human GI imaging modalities

Exclusion Criteria:

- Topics focused on Detection, Prediction, Comparison, Rectal (non-GI scope), Tumor biology without image analysis, 3D modalities, Review articles, or non-research documents (e.g., practice guidelines)
- Papers focused on non-GI cancers or non-relevant imaging technologies (e.g., White Light Interferometry)
- Non-English publications or inaccessible full texts

C. Screening Procedure

The review employed a dual-stage filtering protocol, each independently conducted by the two authors, with conflict resolution via consensus meetings.

1) Title-Based Screening

Documents were assessed based on the relevance of their titles:

- Approved by Author 1: 287 (31.2%)
- Approved by Author 2: 196 (21.3%)
- Overlap (approved by both): 180 (19.6%)
- Conflicting decisions: 123 (13.4%)

After resolving conflicts, an additional 27 documents were accepted, resulting in 207 titles advancing to the next stage.

2) Abstract-Based Screening

A total of 194 documents were eligible after removing 13 non-English or unavailable entries. Abstracts were then screened using the same inclusion/exclusion criteria:

- Approved by Author 1: 53 (27.3%)
- Approved by Author 2: 33 (17.0%)
- Overlap: 20 (10.3%)
- Conflicts: 46 (23.7%)

An additional 8 documents were accepted after conflict resolution, bringing the total to 28 documents.

3) Final Eligibility Check

A full-text review ensured alignment with the review's scope. Eight studies were excluded at this stage due to:

- Use of incompatible imaging modalities
- Lack of focus on segmentation or on relevant GI cancers

The final corpus included 20 peer-reviewed publications [1-20], forming the evidence base for the ensuing review and comparative analysis. A summary of the whole process can be inspected in the PRISMA flow diagram of Figure 2.

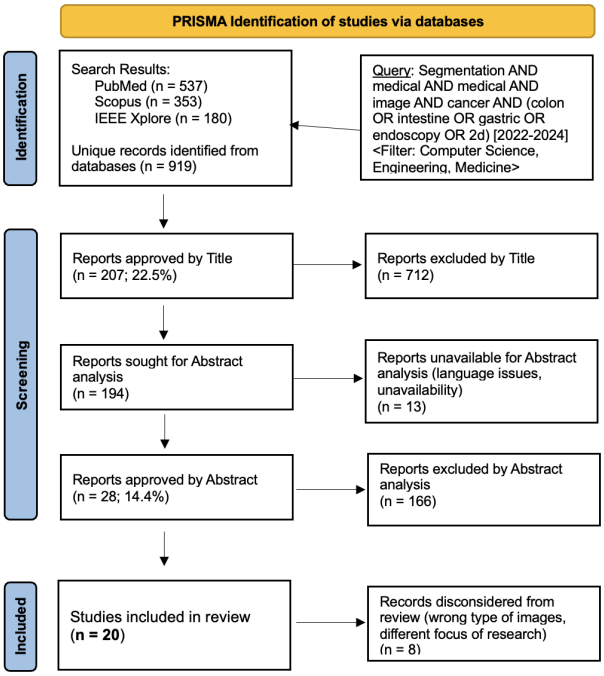


Figure 2 - PRISMA flow diagram detailing the identification, screening, and inclusion of studies for this review.

III. Result

A. Dataset Characteristics

The analysis of dataset sizes across the reviewed papers reveals a significant concentration of studies utilizing relatively small image collections. As shown in Figure 3, most datasets contained fewer than 3,000 images, with the most common range being <= 1,000 images. This highlights a prevailing limitation in current GI segmentation research, where access to large-scale annotated image repositories remains scarce. Despite this, a few outliers reported datasets with over 20,000 images, pushing the maximum to 28,939 images, though the median dataset size was just 1,252 images, and the mean stood at approximately 3,877, indicating a skewed distribution.

In terms of clinical segmentation targets (Figure 4), lesion region segmentation was the most common focus (n=7) [3,5,14,15,16,17,18], followed by early gastric cancer (EGC)

segmentation (n=5) [1,4,6,16,18], and polyp segmentation (n=4) [9,10,11,12]. A smaller subset addressed GI tract boundaries (n=2) [10,13], metaplasia (n=2) [7,19], or other cancer types (n=3) [2,8,10]. These results suggest a growing attention toward high-impact lesion detection tasks, especially for early-stage disease, but also reflect the heterogeneity in segmentation objectives, posing challenges for standardizing datasets or benchmarking algorithms.

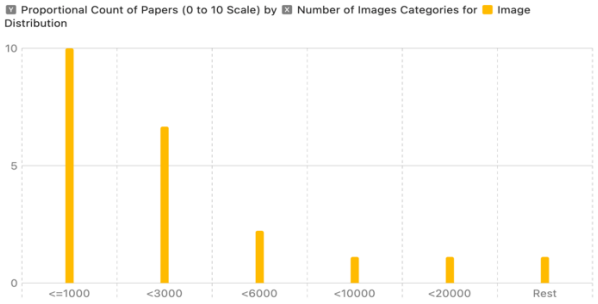


Figure 3 - Distribution of dataset sizes used across the reviewed studies.

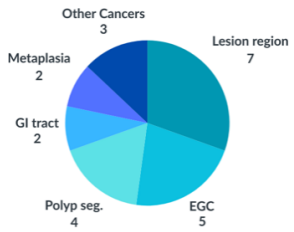


Figure 4 - Primary segmentation targets in the selected studies.

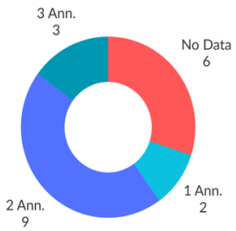


Figure 5 - Reported number of annotators per study.

B. Dataset Annotations

The availability and transparency of datasets emerged as a critical factor in the reviewed studies. As shown in Figure 6, there was an even split between public and private datasets, with 10 papers relying on openly accessible datasets and 10 utilizing private or institution-specific collections. Among the public datasets, Kvasir-SEG [21] was the most frequently used (9 papers) [1,2,3,8][9][10][11][12][14], followed by CVC-ClinicDB [22] (5 papers) [3][8][10][11][12] and ETIS-Larib [23] (2 papers) [8][10], indicating a heavy reliance on a small subset of well-established GI imaging repositories. This limited diversity underscores a broader challenge in the field: the need for expanded public datasets to support generalizability and reproducibility in segmentation research.

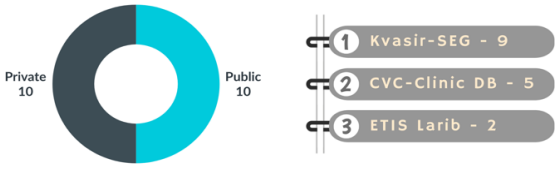


Figure 6 - Proportion of studies using public versus private datasets. An even split was observed, with Kvasir-SEG [21] and CVC-ClinicDB [22] among the most frequently used public repositories.

Regarding dataset annotation practices, a substantial proportion of studies (14 out of 20) explicitly reported involving human annotators in the labelling process (Figure 5). However, transparency about the number of annotators remained inconsistent. Among those that provided information, the most common configuration was two annotators [1][2][7][10][11][12][14][16][17] (9 papers), followed by three annotators [6][8][18] (3 papers), and only two papers [3][4] employed a single annotator. Notably, six studies [5][9][13][15][19][20] provided no information on annotation personnel, which introduces concerns about label quality, inter-observer variability, and dataset reliability. These findings highlight the pressing need for clearer reporting standards in annotation protocols, especially in clinical image segmentation tasks where expert consensus is crucial.

C. Deep Learning Architectures

An analysis of the architectural frameworks adopted across the reviewed studies reveals a strong tendency toward task-specific customization rather than strict adherence to canonical models (Figure 7). The most common approach involved the use of custom-designed models (n=6) [2][6][10][11][19][20], often blending elements from multiple standard architectures to better suit the unique challenges posed by GI imaging data, namely subtle lesion boundaries, varied lighting, or limited dataset sizes.

Among established models, U-Net and its derivatives remained the most popular backbone architecture (n=4) [3][5][13][15], continuing its prominence in medical image segmentation due to its efficient encoder–decoder structure and strong performance on small datasets. Mask R-CNN variants followed (n=3) [1][16][18], often enhanced with modules like bi-directional feature fusion or region discrimination to improve detection in gastroscopic imaging. More novel methods such as Transformers (n=2) [8][12] have begun to emerge, signaling growing interest in leveraging generative or attention-based mechanisms, although their adoption remains limited, likely due to data requirements and training complexity. CNN-based models (n=2) [4][7] and FCN-based approaches (n=1) [17] played more foundational roles, either in baseline tasks or early-stage feature extraction, while still contributing meaningfully to detection and classification tasks. At last, Ensemble methods [14] and GANs [9] were the least utilized (n=1), suggesting that while theoretically promising, either their complexity may hinder widespread adoption as a current research trend, or there could still exist training instability or domain specificity associated with the architecture implementation. As a final comment, only two of the twenty papers inspected provided easily available and reproducible source code [3][9].

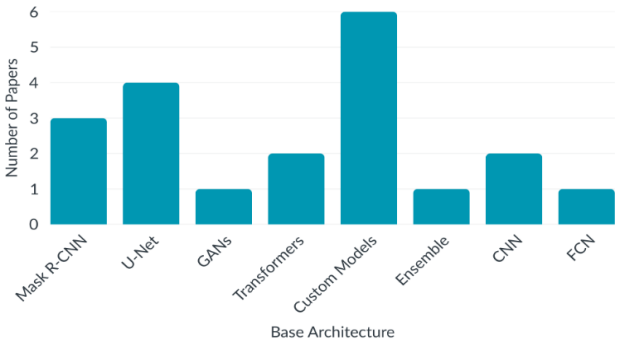


Figure 7 - Types of deep learning architectures employed across the reviewed papers.

IV. Discussion

This review reveals that while segmentation of gastrointestinal images is gaining traction, the field remains constrained by several challenges. Most studies rely on relatively small datasets, often under 3,000 images, with a strong dependency on a handful of public datasets like Kvasir-SEG [21] and CVC-ClinicDB [22]. Annotation practices lack consistency, with limited reporting on the number and expertise of annotators, which raises concerns about label quality and reproducibility. Methodologically, there is a trend toward task-specific, custom-built architectures, reflecting innovation but limiting standardization and comparability. The limited uptake of newer paradigms such as transformers and ensembles suggest that complexity and data demands remain barriers. Overall, the findings underscore a need for larger, more diverse datasets, transparent annotation protocols, and more unified benchmarking efforts to advance reproducible and clinically relevant GI segmentation research. This is reinforced by methodological proposals such as QUAIDE [24], that aim to create standards for pre-clinical trial algorithmic results on this field.

Acknowledgment

This work was supported by European funds within the scope of the Recovery and Resilience Mechanism of the European Union with reference 2024.07584.IACDC/2024, and by the Horizon Europe Framework Programme (Grant Agreement Nº: 101095359) and by the UK Research and Innovation (Grant Agreement Nº: 10058099).

References

[1] Zhang K, "Early gastric cancer detection and lesion segmentation based on deep learning and gastroscopic images", doi: 10.1038/s41598-024-58361-8.

[2] Li Z, "MFA-Net: Multiple Feature Association Network for medical image segmentation", doi: 10.1016/j.compbiomed.2023.106834.

[3] He D, "Dual-branch hybrid network for lesion segmentation in gastric cancer images", doi: 10.1038/s41598-023-33462-y.

[4] Teramoto A, "Detection and Characterization of Gastric Cancer Using Cascade Deep Learning Model in Endoscopic Images", doi: 10.3390/diagnostics12081996.

[5] Sun Y, "Lesion Segmentation in Gastroscopic Images Using Generative Adversarial Networks", doi: 10.1007/s10278-022-00591-1.

[6] Du W, "Early gastric cancer segmentation in gastroscopic images using a co-spatial attention and channel attention based triple-branch ResUnet", doi: 10.1016/j.cmpb.2023.107397.

[7] Ligato I, "Convolutional Neural Network Model for Intestinal Metaplasia Recognition in Gastric Corpus Using Endoscopic Image Patches", doi: 10.3390/diagnostics14131376.

[8] Chang Q, "ESFPNet: Efficient Stage-Wise Feature Pyramid on Mix Transformer for Deep Learning-Based Cancer Analysis in Endoscopic Video", doi: 10.3390/jimaging10080191.

[9] R. Osuala et al., "Medigan: A Python library of pretrained generative models for medical image synthesis," Journal of Medical Imaging, vol. 10, no. 6, 2023, doi: 10.1117/1.JMI.10.6.061403.

[10] Q. Lin, W. Tan, S. Cai, B. Yan, J. Li, and Y. Zhong, "Lesion-Decoupling-Based Segmentation with Large-Scale Colon and Esophageal Datasets for Early Cancer Diagnosis," IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 8, pp. 11142–11156, 2024, doi: 10.1109/TNNLS.2023.3248804.

[11] P. Saxena and A. K. Bhandari, "Context Aware Automatic Polyp Segmentation Network with Mask Attention," IEEE Transactions on Artificial Intelligence, vol. 5, no. 7, pp. 3510–3523, 2024, doi: 10.1109/TAI.2024.3375832.

[12] W. Li, Y. Liu, X. Liu, and Y. Qi, "Fusing Transformer and FCN for Polyp Segmentation - 2023 7th Asian Conference on Artificial Intelligence Technology (ACAIT)," 2023 7th Asian Conference on Artificial Intelligence Technology (ACAIT), pp. 623–628, 2023, doi: 10.1109/ACAIT60137.2023.10528540.

[13] N. D. Manh et al., "EndoUNet: A Unified Model for Anatomical Site Classification, Lesion Categorization and Segmentation for Upper Gastrointestinal Endoscopy - 2022 14th International Conference on Knowledge and Systems Engineering (KSE)," 2022 14th International Conference on Knowledge and Systems Engineering (KSE), pp. 1–6, 2022, doi: 10.1109/KSE56063.2022.9953766.

[14] N. V. Sai Manoj, M. Rithani, and R. S. SyamDev, "Enhancing Gastric Cancer Diagnosis Through Ensemble Learning for Medical Image Analysis - 2023 Seventh International Conference on Image Information Processing (ICIIP)," 2023 Seventh International Conference on Image Information Processing (ICIIP), pp. 822–826, 2023, doi: 10.1109/ICIIP61524.2023.10537790.

[15] Y. Zhang, Y. Li, and D. He, "U-Shaped Network with Multi-Scale Feature Extraction for Lesion Region Segmentation in Gastric Cancer Images - 2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)," 2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), vol. 3, pp. 234–238, 2023, doi: 10.1109/ICIBA56860.2023.10165314.

[16] T. Liang et al., "Mask R-CNN with R-D Net: A Detection and Segmentation Model for Early Gastric Cancer Gastroscopy Images - 2023 3rd International Conference on Electronic Information Engineering and Computer Science (EIECS)," 2023 3rd International Conference on Electronic Information Engineering and Computer Science (EIECS), pp. 1346–1352, 2023, doi: 10.1109/EIECS59936.2023.10435544.

[17] D. Katayama et al., "Development of Computer-Aided Diagnosis System Using Single FCN Capable for Indicating Detailed Inference Results in Colon NBI Endoscopy - 2023 International Technical Conference on Circuits/Systems, Computers, and Communications (ITC-CSCC)," 2023 International Technical Conference on Circuits/Systems, Computers, and Communications (ITC-CSCC), pp. 1–6, 2023, doi: 10.1109/ITC-CSCC58803.2023.10212877.

[18] Z.-H. Cui et al., "Application of Improved Mask R-CNN Algorithm Based on Gastroscopic Image in Detection of Early Gastric Cancer - 2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)," 2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC), pp. 1396–1401, 2022, doi: 10.1109/COMPSAC54236.2022.00221.

[19] Y.-T. Chen, E.-H. Yang, W.-L. Chang, J. Lin, H.-C. Cheng, and C.-R. Huang, "Mask Focal Modulation Network for Gastric Intestinal Metaplasia Segmentation - 2024 International Joint Conference on Neural Networks (IJCNN)," 2024 International Joint Conference on Neural Networks (IJCNN), pp. 1–8, 2024, doi: 10.1109/IJCNN60899.2024.10650100.

[20] A. Garbaz, Y. Oukdach, S. Charfi, M. El Ansari, L. Koutti, and M. Salihoun, "Bleeding Segmentation Based on a U-Formed Network with Separable Contextual Feature-Guided in Wireless Capsule Endoscopy Images - 2024 11th International Conference on Wireless Networks and Mobile Communications (WINCOM)," 2024 11th International Conference on Wireless Networks and Mobile Communications (WINCOM), pp. 1–6, 2024, doi: 10.1109/WINCOM62286.2024.10656577.

[21] P. Jha, M. Smedsrud, D. Riegler, P. Halvorsen, T. de Lange, and H. D. Johansen, "Kvasir-SEG: A segmented polyp dataset," in Proc. Int. Conf. Multimedia Modeling (MMM), Daejeon, Korea, Jan. 2020, pp. 451–462, doi: 10.1007/978-3-030-37731-1_37.

[22] S. Bernal et al., "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," Computerized Medical Imaging and Graphics, vol. 43, pp. 99–111, Jul. 2015, doi: 10.1016/j.compmedimag.2015.02.007.

[23] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," Int. J. Comput. Assist. Radiol. Surg., vol. 9, no. 2, pp. 283–293, Feb. 2014, doi: 10.1007/s11548-013-0926-3.

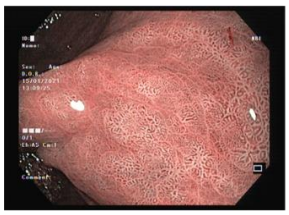
[24] G. Antonelli et al., "QUAIDE—Quality assessment of AI preclinical studies in diagnostic endoscopy," Gut, pp. 1–9, 2024, doi: 10.1136/gutjnl-2024-332820.

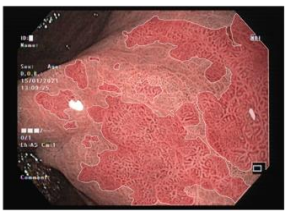
A Review of Segmentation Algorithms for Gastrointestinal Images: Datasets, Annotations and Architectures

Research Focus

Accurate gastrointestinal (GI) image segmentation is critical for early detection of cancers and other pathologies, necessitating robust computer-aided diagnostic systems for clinical practice. Despite significant advances in deep learning methodologies, research in this domain remains fragmented, characterized by limited datasets and diverse annotation strategies across studies.

This systematic review consolidated and evaluated segmentation techniques applied to gastrointestinal imaging from 2022-2024, examining three key dimensions: datasets, annotation practices, and model architectures.



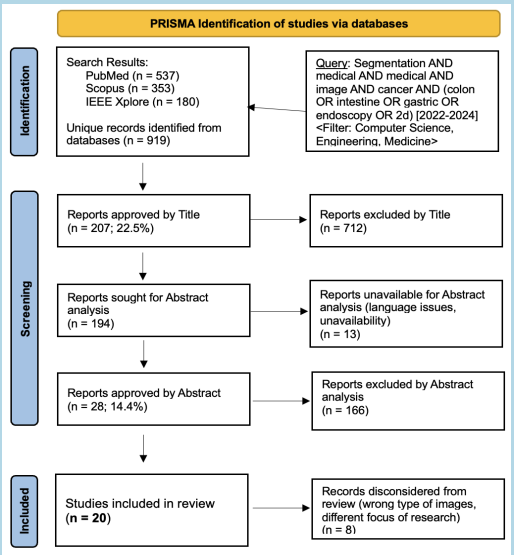


Tomás Ferreira (FCUP/INESC Porto)
David Marques (FCUP/INESC Porto)
Mário Dinis-Ribeiro (IPO Porto)
Miguel Coimbra (FCUP/INESC Porto)

Throughout this methodology, the aim was to identify current trends and challenges in gastrointestinal image segmentation research, while establishing recommendations for advancing clinically relevant and reproducible investigations.

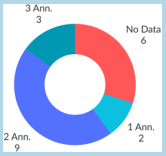
The main goal was to provide researchers and clinicians with a solid assessment of the field's state, addressing the gap where, despite technological advances, research remains fragmented and requires standardization in evaluation metrics and methodological approaches.

Review Methodology



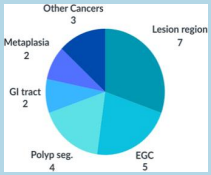
PRISMA flow diagram detailing the identification, screening, and inclusion of studies for this review.

Results: Dataset Annotations

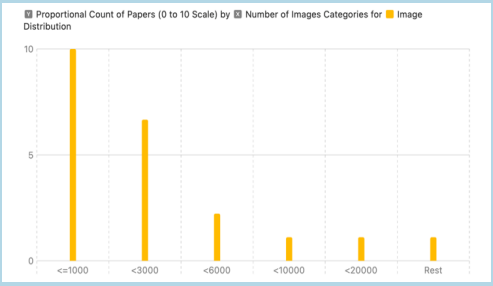


Only 14 of the 20 the studies reported human annotators involvement: limited reporting on annotator expertise and inter-observer agreement might cause quality concerns.

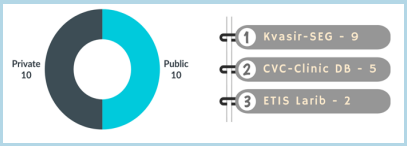
Clinical targets focused primarily on lesion region segmentation, early gastric cancer detection, and polyp identification, reflecting growing attention toward high-impact early-stage disease detection but also highlighting heterogeneity in segmentation objectives that challenges standardization efforts



Results: Dataset Characteristics



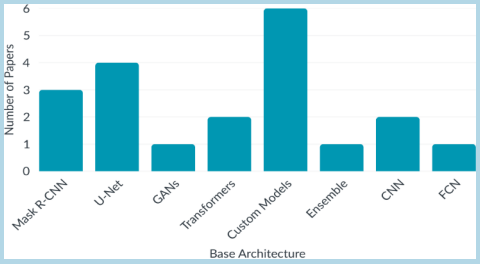
Distribution of dataset sizes used in the studies: the majority of the datasets contained less than 3,000 images.
Median: 1,252 images | Mean: 3,877 images.



Proportion of studies using public versus private datasets. An even split was observed, with Kvasir-SEG and CVC-ClinicDB as the most frequently used public repositories.

Results: Deep Learning Architectures

Architectural analysis revealed that custom-designed approaches dominate currently the field by blending elements from multiple standard architectures to address unique GI imaging challenges. U-Net variants maintained popularity as backbone architectures due to their proven efficiency on small datasets, while newer paradigms like Transformers showed limited adoption likely due to data requirements and training complexity.
As additional note, it was observed that only 10% of studies provide accessible source code, undermining the reproducibility of the techniques in use.



Discussion & Conclusions

1. Most studies rely on less than 3,000 images with heavy dependency few public repositories;
2. Limited transparency about annotator expertise raises label quality concerns;
3. Custom architectures show innovation but limit comparability;
4. Minimal code availability hinders research improvement.

This review underscores the need for research standardization in the field, showing the critical importance of methodological frameworks such as QUAIDE for enhanced pre-clinical algorithmic validation.