UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

Università degli Studi di Modena e Reggio Emilia
Dipartimento di Scienze e Metodi dell'Ingegneria

Master's Degree Course in Digital Automation Engineering

Automated Crohn's Disease Screening in Small-Bowel Capsule Endoscopy: A Data-Centric Pipeline for Reliable Workload-Aware Evaluation

Supervisor:
Prof. Manuel Iori

Candidate:
Michele Ferramola

Co-Supervisor:
Prof. Stefania Monica

Academic Year 2024/2025

Abstract

Small-Bowel Capsule Endoscopy (SBCE) is a valuable tool for Crohn's disease assessment, but its clinical use is limited by the large number of frames that must be reviewed and by the difficulty of identifying subtle, sparse, and heterogeneous lesions. At the same time, many published AI studies report high performance under evaluation settings that are not fully representative of realistic screening conditions. This thesis addresses automated Crohn's disease screening in SBCE from a data-centric perspective, with the aim of understanding how dataset construction, split design, and evaluation protocol affect the reliability of reported results.

A unified experimental pipeline was developed by combining three public SBCE datasets with different annotation schemes, grouping structures, and acquisition devices. The pipeline included label harmonization into a common screening-oriented label space, embedding-guided redundancy reduction, mitigation of Blood dominance, and patient-wise split optimization formulated as a constrained combinatorial problem. Frozen DINOv2 embeddings were used as generic visual representations, and classification experiments were conducted with linear models and lightweight non-linear heads under realistic low-prevalence conditions.

The results showed that linear classifiers were consistently limited by a severe Crohn recall–precision trade-off, while a shallow MLP improved both recall and precision, indicating that the embedding space contained useful non-linear discriminative structure. However, broader increases in model complexity did not overcome the main bottlenecks, which remained tied to low disease prevalence, limited patient diversity, and frame-level independence. At pathology level, workload-aware triage experiments based on Recall@10% per patient provided a more clinically meaningful evaluation setting and showed a non-trivial triage signal, with the binary erosion-versus-rest formulation outperforming the three-class alternative. Pathology-level experiments further showed that split assignment could have a larger impact on performance than hyperparameter variation. Finally, zero-shot testing with MedGemma did not surpass the supervised pipeline and showed strong Crohn over-prediction, highlighting the limits of large medical vision–language models when applied without task-specific supervision and carefully controlled evaluation.

Overall, the thesis shows that, in realistic SBCE Crohn screening, performance is shaped less by model scale alone than by the interaction between data curation, patient-level evaluation, and task formulation. Its main contribution is therefore not the proposal of a deployment-ready detector, but the definition of a reproducible framework for obtaining more trustworthy performance estimates under clinically meaningful constraints.

Sommario

L'endoscopia capsulare del tenue (SBCE) è uno strumento prezioso per la valutazione della malattia di Crohn, ma il suo impiego clinico è limitato dall'elevato numero di frame da esaminare e dalla difficoltà di identificare lesioni sottili, sparse ed eterogenee. Al contempo, molti studi di intelligenza artificiale pubblicati riportano prestazioni elevate in condizioni di valutazione non pienamente rappresentative di uno screening realistico. Questa tesi affronta lo screening automatizzato della malattia di Crohn in SBCE da una prospettiva data-centric, con l'obiettivo di comprendere come la costruzione del dataset, la progettazione degli split e il protocollo di valutazione influenzino l'affidabilità dei risultati riportati.

È stata sviluppata una pipeline sperimentale unificata combinando tre dataset SBCE pubblici con schemi di annotazione, strutture di raggruppamento e dispositivi di acquisizione differenti. La pipeline comprende l'armonizzazione delle etichette in uno spazio comune orientato allo screening, la riduzione della ridondanza guidata da embedding, la mitigazione della dominanza della classe Blood e l'ottimizzazione degli split per paziente formulata come problema combinatorio vincolato. Embedding DINOv2 congelati sono stati utilizzati come rappresentazioni visive generiche, e gli esperimenti di classificazione sono stati condotti con modelli lineari e teste non lineari leggere in condizioni realistiche di bassa prevalenza.

I risultati mostrano che i classificatori lineari sono sistematicamente limitati da un severo trade-off recall–precisione per la classe Crohn, mentre una MLP shallow migliora sia recall sia precisione, indicando che lo spazio embedding contiene struttura discriminativa non lineare utile. Tuttavia, incrementi ulteriori nella complessità del modello non superano i colli di bottiglia principali, che restano legati alla bassa prevalenza della malattia, alla limitata diversità dei pazienti e all'indipendenza frame-level. A livello patologico, esperimenti di triage workload-aware basati su Recall@10% per paziente hanno fornito un contesto di valutazione clinicamente più significativo e mostrato un segnale di triage non banale, con la formulazione binaria erosione-versus-rest superiore all'alternativa a tre classi. Gli esperimenti a livello patologico hanno inoltre mostrato che l'assegnazione degli split può avere un impatto sulle prestazioni maggiore della variazione degli iperparametri. Infine, il test zero-shot con MedGemma non ha superato la pipeline supervisionata e ha mostrato una forte sovrapredizione della classe Crohn, evidenziando i limiti dei grandi modelli medici vision–language quando applicati senza supervisione task-specific e valutazione attentamente controllata.

Nel complesso, la tesi mostra che, nello screening realistico della malattia di Crohn in SBCE, le prestazioni sono determinate meno dalla scala del modello in sé che dall'interazione tra curazione dei dati, valutazione a livello paziente e formulazione del task. Il contributo principale non è quindi la proposta di un rilevatore pronto per il deployment clinico, ma la definizione di un framework riproducibile per ottenere stime di prestazione più affidabili sotto vincoli clinicamente significativi.

Contents

1 Introduction and Problem Context

1.1 Crohn's Disease and Small-Bowel Capsule Endoscopy

Crohn's disease (CD) is a chronic inflammatory bowel disease characterized by transmural inflammation that may involve any segment of the gastrointestinal tract, with a strong predilection for the small bowel. Clinically, patients frequently present with abdominal pain, chronic or intermittent diarrhea, weight loss, fatigue, and iron-deficiency anemia. In more advanced stages, complications such as strictures, ulcerations, fistulae, and bleeding may occur. The heterogeneity of clinical manifestations reflects the patchy and progressive nature of the inflammatory process [1].

Small-bowel capsule endoscopy (SBCE) has become an established diagnostic modality for evaluating suspected Crohn's disease, particularly when ileocolonoscopy is negative and no obstructive symptoms are present. European clinical guidelines recommend SBCE in this context because it enables direct visualization of mucosal abnormalities in segments of the small bowel that are otherwise difficult to access [2]. Compared with cross-sectional imaging, SBCE offers high sensitivity for early mucosal lesions, making it particularly valuable in cases of subtle inflammatory activity.

However, SBCE examinations generate tens of thousands of frames per patient, and diagnostic interpretation relies heavily on reader expertise and sustained attention. This introduces variability and cognitive burden, motivating the development of computer vision systems to support standardized reading, screening, and prioritization rather than replacing clinical reasoning [2].

Figure 1 illustrates representative examples of ulcers and erosions as captured by SBCE. Both lesion types are hallmarks of Crohn's-related mucosal damage, yet their visual distinction is inherently difficult. Ulcers tend to present as deeper, more irregularly bordered defects, while erosions typically appear as superficial breaks in the mucosa; however, the discriminative features separating the two are often subtle—small differences in depth, border regularity, and surrounding mucosal texture—and can be partially or entirely obscured by variable imaging conditions such as illumination angle, distance from the mucosal surface, presence of debris or bile, and partial occlusion of the field of view. This difficulty is not specific to human observers: it equally affects any automated feature extractor, since the visual signatures that characterize these lesions occupy a narrow range of the overall appearance space and may fall below the representational resolution of encoders not specifically adapted to mucosal pathology. The challenge is therefore intrinsic to the domain rather than to the observer, and it underscores the importance of both the representation quality and the downstream decision strategy in any screening pipeline.

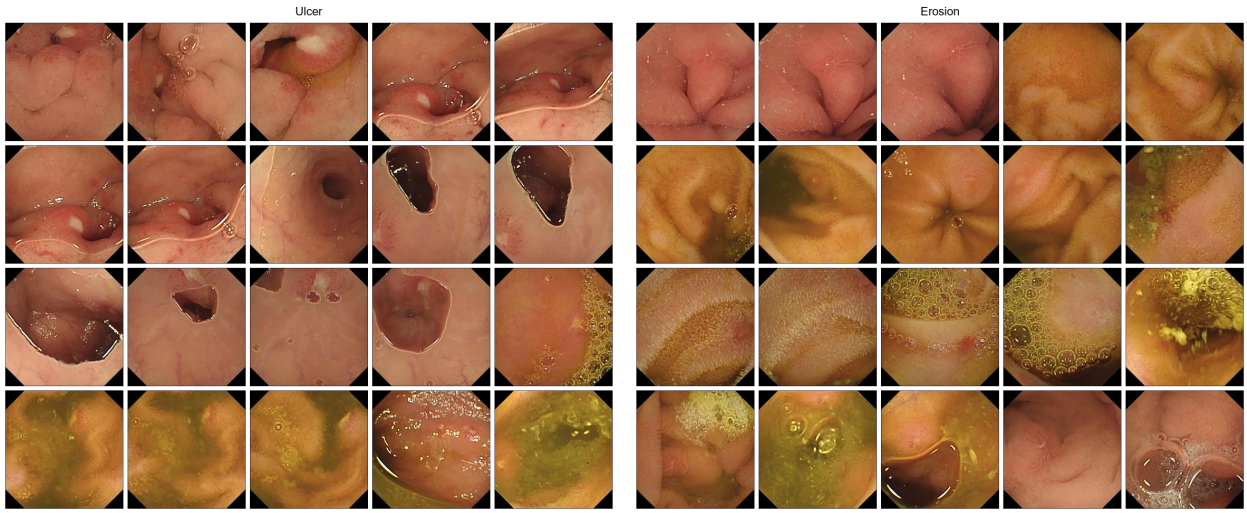


Figure 1: Representative SBCE frames of ulcers (left) and erosions (right) from the Kvasir-Capsule dataset. The discriminative features between the two lesion types are fine-grained and easily obscured by variable illumination, viewing angle, and tissue visibility, making reliable differentiation challenging for both human readers and automated feature extractors.

1.2 The Screening Problem and Workload Reduction

In screening scenarios, the clinical objective is often to maximize sensitivity for relevant pathology while reducing the review workload imposed on clinicians. A typical capsule endoscopy examination generates thousands of frames, the vast majority of which are normal; the gastroenterologist must review all of them to identify the small minority containing diagnostically relevant findings. Workload reduction, in this context, refers to the use of automated systems to filter or rank frames so that the clinician reviews only a subset—ideally the most suspicious one—while retaining the ability to detect clinically significant lesions. The underlying trade-off is between the fraction of frames reviewed (the workload budget) and the fraction of true lesions captured within that budget (the detection sensitivity).

This paradigm has gained increasing traction in the recent literature as an operational objective that complements, and in some settings supersedes, traditional classification accuracy. Ding et al. demonstrated that a deep-learning model for small-bowel capsule endoscopy could reduce reading time by 96.6% while maintaining gastroenterologist-level diagnostic accuracy, and explicitly defined the reading-time reduction rate as a primary evaluation metric alongside sensitivity and specificity [3]. Aoki et al. proposed using a CNN as a first-pass screening filter, allowing clinicians to focus review on the frames flagged as abnormal and thereby compressing the effective reading workload without sacrificing detection performance [4]. More recently, a multicentre prospective study across 14 European centres showed that AI-assisted reading reduced median reading time from 33.7 to 3.8 minutes (-89%) while simultaneously improving diagnostic yield from 62.4% to 73.7% [5]. These studies collectively establish that the clinical value of AI in capsule endoscopy is measured not only by how accurately the model classifies individual frames, but by how effectively it reduces the human effort required

to reach a diagnostic conclusion.

Under strong class imbalance, conventional accuracy can be dominated by the majority class and fail to reflect clinically meaningful performance. ROC analysis summarizes the sensitivity–specificity trade-off [6], but in highly imbalanced settings precision–recall curves can be more informative because they directly capture positive predictive value across recall levels [7]. Moreover, a classifier that reaches high recall only by flagging many frames may not reduce workload in practice. This motivates evaluation that explicitly models workload constraints—such as recall at a fixed percentage of frames reviewed—rather than relying only on threshold-independent metrics.

1.3 Research Questions and Contributions

The present work is motivated by a sequence of interrelated methodological questions that arise when attempting to build a reliable automated screening system for Crohn's disease from heterogeneous capsule endoscopy data. The first question concerns the harmonization problem: given that publicly available SBCE datasets adopt incompatible labelling conventions, overlapping terminologies, and different annotation granularities, how can these heterogeneous sources be reconciled into a unified, clinically grounded label space that is suitable for screening-oriented modelling? Addressing this question, the thesis proposes a three-superclass mapping (Crohn findings, Other findings, Normal) anchored in established clinical interpretation of capsule endoscopy findings, together with a conservative priority rule for resolving multi-label co-occurrences.

The second question addresses the structural pathologies of the raw data: how can the extreme redundancy introduced by dense temporal frame extraction and the dominance of specific subclasses—most notably Blood within Other findings—be controlled without sacrificing the visual diversity and clinically meaningful co-occurrence patterns that the dataset must preserve? In response, the thesis develops an embedding-guided redundancy reduction strategy operating in the DINOv2 latent space, with a regime-aware clustering approach that adapts granularity to per-patient visual diversity, complemented by a dedicated reduction protocol for Blood that explicitly protects multi-label frames from pruning.

The third question turns to the dataset splitting stage: how can the assignment of patients and videos to training, validation, and test splits be formulated as a constrained optimization problem that simultaneously enforces group integrity, prevents data leakage, and stabilizes the balancing conditions required for effective model training? The thesis formalizes this task as a multidimensional multi-way number partitioning problem (MDMWNPP) with an explicit multi-criteria objective function, solves it exactly using the CP-SAT constraint programming solver from OR-Tools, and provides a quantitative comparison against a greedy best-insertion baseline that demonstrates the structural advantages of the exact approach. The pipeline produces a reproducible physical dataset build in WebDataset format, accompanied by a suite of formal validation checks (C1–C6: group integrity, absence of data leakage,

superclass balancing, and Crohn distribution targets).

The fourth and final question concerns evaluation alignment: how should performance assessment be designed so that it reflects the operational realities of a screening workflow—low disease prevalence, strong temporal redundancy, and an explicit trade-off between sensitivity and reading workload—rather than relying on conventional metrics that may be misleading under these conditions? The thesis addresses this through a screening-oriented framing of evaluation metrics and a workload-aware perspective that connects model selection to clinically meaningful operating points.

1.4 Structure of the Thesis

The document is organized as follows. Chapter 2 reviews clinical and methodological background. Chapter 3 covers dataset construction, including source description, label harmonization, early curation, and structural challenges. Chapter 4 details redundancy reduction and pruning. Chapter 5 characterizes the post-curation dataset. Chapter 6 formulates and solves patient-wise splitting as a combinatorial optimization problem. Chapter 7 introduces feature extraction with DINOv2 embeddings and data augmentation. Chapter 8 presents experiments with linear classifiers. Chapter 9 extends the classification to non-linear models. Chapter 10 addresses pathology-level classification and workload-based triage. Chapter 11 presents a zero-shot comparison with MedGemma. Chapter 12 discusses findings, contextualizes results with the existing literature, and identifies limitations. Chapter 13 concludes.

2 Clinical and Methodological Background

2.1 Visual Correlates of Crohn's Disease in SBCE and Lesion Relevance as a Modelling Proxy

In SBCE, Crohn's disease is inferred primarily through visual identification of inflammatory and ulcerative patterns. These range from mild mucosal changes to overt mucosal breaks. Importantly, not all findings carry equal diagnostic weight. Subtle features such as mild edema or hyperemia may reflect non-specific inflammation or drug-induced injury, whereas mucosal breaks—particularly when multiple or deep—are more consistently interpreted as clinically relevant in suspected Crohn's disease [8].

This clinical hierarchy is reflected in established SBCE scoring systems. The Lewis score and the Capsule Endoscopy Crohn's Disease Activity Index (CECDAI) quantify inflammatory burden by assigning greater weight to ulceration extent, depth, and distribution [9]. Although these scores were designed primarily to grade disease severity in known Crohn's disease rather than establish diagnosis, their structure reinforces the principle that ulcerations and erosions are central markers of Crohn-related mucosal pathology.

Accordingly, using ulcerations and erosions as primary modelling targets in automated SBCE analysis is clinically grounded. These lesions represent visually robust, pathophysiologically meaningful markers of Crohn suspicion, reducing ambiguity compared with weaker inflammatory signs that lack specificity [9].

2.2 Computer Vision in Capsule Endoscopy: From Classical Methods to Deep Learning

Early attempts at automated capsule endoscopy analysis relied on handcrafted features such as color histograms, texture descriptors, and edge-based measures combined with conventional classifiers including support vector machines and random forests. These approaches were constrained by limited dataset size and feature engineering capacity. The emergence of deep learning, particularly convolutional neural networks (CNNs), transformed the field by enabling end-to-end feature learning directly from raw image data. CNN-based approaches have demonstrated strong performance in lesion detection and classification tasks across gastrointestinal endoscopy, including ulcer and bleeding detection in capsule endoscopy. A systematic review and meta-analysis by Soffer et al. [10] confirmed the diagnostic potential of deep learning in wireless capsule endoscopy across multiple pathology types, while also highlighting frequent methodological shortcomings in the existing literature—including inadequate dataset splitting, limited external validation, and insufficient reporting of class imbalance—that compromise the reliability of published performance figures.

In the specific context of Crohn's disease, dedicated deep learning systems have reported high frame-level detection accuracy. Klang et al. [11] trained a CNN on single-center capsule endoscopy data and reported an AUC of 0.94–0.99 for Crohn ulcer detection, while Majtner et al. [12] proposed a framework for autonomous detection and classification of Crohn lesions in the small bowel and colon, achieving sensitivity of 95.7% and specificity of 99.8%. However, both studies relied on single-center data without patient-wise splitting guarantees, and their evaluation protocols do not account for the low prevalence and extreme class imbalance characteristic of screening scenarios. These methodological limitations motivate the rigorous dataset construction and evaluation framework developed in the present thesis. Despite these advances, several structural challenges remain in capsule endoscopy analysis: temporal correlation between consecutive frames, severe class imbalance, visual redundancy, and dataset heterogeneity across acquisition devices and clinical sites.

2.3 Foundation Models and the Role of DINOv2 in Endoscopic Representation Learning

A major recent development in computer vision is the rise of foundation models: large-scale encoders pre-trained on vast datasets using self-supervised objectives, yielding general-purpose visual representations transferable to downstream tasks. DINOv2 is a prominent example, demonstrating that large-scale self-supervised training produces high-quality visual embeddings that can be adapted efficiently to new domains [13].

In capsule endoscopy, foundation models are particularly attractive because curated, expertly annotated medical datasets are limited. Instead of training deep models from scratch, one may extract embeddings from a strong pre-trained encoder and adapt them using lightweight heads or parameter-efficient fine-tuning. Recent work suggests that adapting DINOv2-based encoders to capsule endoscopy tasks can achieve competitive performance under limited supervision [14]. Parallel efforts propose endoscopy-native foundation models trained on large-scale gastrointestinal corpora (e.g., EndoDINO) [15]. Moreover, long-range temporal context may improve semantic understanding in endoscopic video analysis [16]. Taken together, these works motivate a representation-centric pipeline based on DINOv2 embeddings, combined with clinically grounded lesion proxies.

However, feature extraction is a particularly sensitive step in the endoscopic domain. The visual signatures that distinguish pathological categories—especially within the Crohn spectrum—are fine-grained: subtle differences in mucosal texture, lesion depth, border sharpness, and colour variation that occupy a narrow region of the overall appearance space. A generic encoder pre-trained on natural images may not allocate sufficient representational capacity to these micro-features, effectively compressing clinically relevant distinctions into a subspace that downstream classifiers cannot reliably separate. This makes the choice of encoder, its training domain, and its representational granularity a potential bottleneck for the entire pipeline, and motivates careful empirical assessment of whether frozen general-purpose em-

beddings capture enough discriminative signal for screening-level performance, or whether domain-adapted representations are ultimately required.

2.4 Datasets as an Enabling Factor and Selection Rationale

Public datasets have played an essential role in advancing computer vision research in gastrointestinal endoscopy, and the landscape of available resources has grown considerably over the past decade. The systematic review by Zhu et al. [17] provides a comprehensive survey of publicly available imaging datasets for artificial intelligence in gastrointestinal endoscopy, cataloguing resources that span different imaging modalities, anatomical segments, annotation formats, and clinical objectives. This review served as the starting point for the dataset selection process undertaken in the present work.

The search was focused specifically on datasets composed of frames or videos acquired through small-bowel capsule endoscopy systems, as opposed to conventional endoscopy or colonoscopy. Priority was given to datasets that covered the full extent of the gastrointestinal tract as captured by capsule devices, rather than restricting to individual anatomical segments, in order to ensure that the resulting training data would encompass the full range of visual appearances that an automated screening system must handle in clinical deployment. Additional selection criteria included the availability of frame-level pathological annotations, sufficient scale to support deep learning experimentation, and—where possible—patient-level or video-level grouping metadata to enable rigorous split construction without data leakage.

Within this search space, four candidate datasets were initially identified as potentially suitable: GALAR (Gastrointestinal Lesion Assessment Repository) [18], Kvasir-Capsule [19], CrohnIPI [20], and KID. Two further resources were considered but subsequently excluded: the AICE project was judged to have limited relevance to the specific screening objective of this work, and KID was excluded due to concerns regarding the methodological rigour of its production process. After this evaluation, three datasets were retained as the empirical foundation of the thesis: GALAR, which provides the largest and most richly annotated multi-label capsule endoscopy resource currently available; Kvasir-Capsule, which contributes medically verified single-label annotations organised by video and has become a widely adopted benchmark in the capsule endoscopy community; and CrohnIPI, which offers a smaller but focused collection of Crohn-related findings and normal mucosa frames annotated by multiple clinical experts. The complementary characteristics of these three sources—in terms of scale, annotation philosophy, and structural organisation—motivated their joint use and informed the harmonisation protocol described in Chapter 3.

3 Dataset Construction: Sources, Harmonization, and Structural Challenges

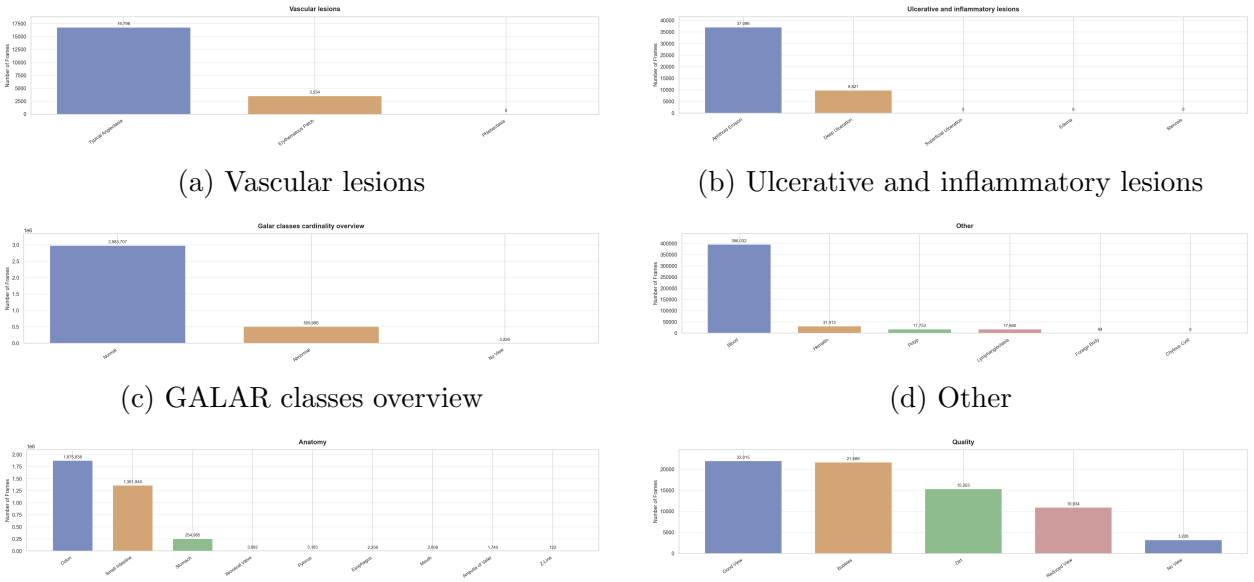
3.1 Description of the Source Datasets

The experimental framework of this thesis relied on three publicly available capsule endoscopy datasets: GALAR (Gastrointestinal Lesion Assessment Repository), Kvasir-Capsule, and CrohnIPI. These datasets differed substantially in size, organization, annotation structure, and acquisition characteristics, and their heterogeneity strongly influenced the design of the harmonization and curation protocol.

GALAR represented the largest and structurally richest source [18]. It was organized at patient level and originated from full-length small-bowel capsule endoscopy videos acquired using multiple devices, including PillCam SB2, PillCam SB3, PillCam COLON2, and Olympus E10 systems. Each patient corresponds to an entire examination, and frame-level annotations follow a multi-label taxonomy. The original archive contained more than 3.5 million frames, making it one of the most extensive capsule endoscopy datasets available at the time this project was initiated. Frames are extracted at high sampling rates from complete videos, which results in pronounced temporal redundancy and large contiguous segments of visually similar images. The dataset also included a non-negligible proportion of low-visibility frames, sometimes labeled under quality-related categories such as "No view" and "Reduced view." Available metadata include patient identifiers, patient age and gender, device information, and frame-level labels, allowing strict patient-wise grouping and advanced structural analyses. Figure 2 illustrates the complexity of the original GALAR label space before any harmonization.

Kvasir-Capsule, published in 2021 as a benchmark dataset for capsule endoscopy research, was organized by video rather than patient [19]. In this thesis, each video was treated as a pseudo-patient to prevent temporal leakage across splits. The dataset comprised 43 videos and 47,238 annotated frames in the configuration adopted here. Compared to GALAR, Kvasir-Capsule contained fewer frames and exhibited lower overall redundancy, but it provided medically verified annotations across a diverse set of pathological categories. The acquisition originated from clinical PillCam systems. While metadata were less extensive than in GALAR, video-level identifiers allowed group-wise split enforcement. Figure 3 shows the category distribution in the curated configuration.

CrohnIPI was considerably smaller and was organized at frame level without patient or video grouping metadata [20]. It contained 3,498 frames acquired with a PillCam SB3 device, focused on Crohn-related findings and normal mucosa. The absence of grouping identifiers prevented patient-wise splitting, which led to its allocation through a secondary, loose assignment step in the splitting optimization protocol. Although limited in size, CrohnIPI contributed additional Crohn-specific evidence and complemented the larger datasets. Fig-



(a) Vascular lesions

(b) Ulcerative and inflammatory lesions

(c) GALAR classes overview

(d) Other

(e) Anatomy

(f) Quality

Figure 2: Initial GALAR taxonomy overview panels. The six plots are paired into three rows (2 panels per row) to summarize the original label space before mapping into the three superclasses (Crohn / Other / Normal).

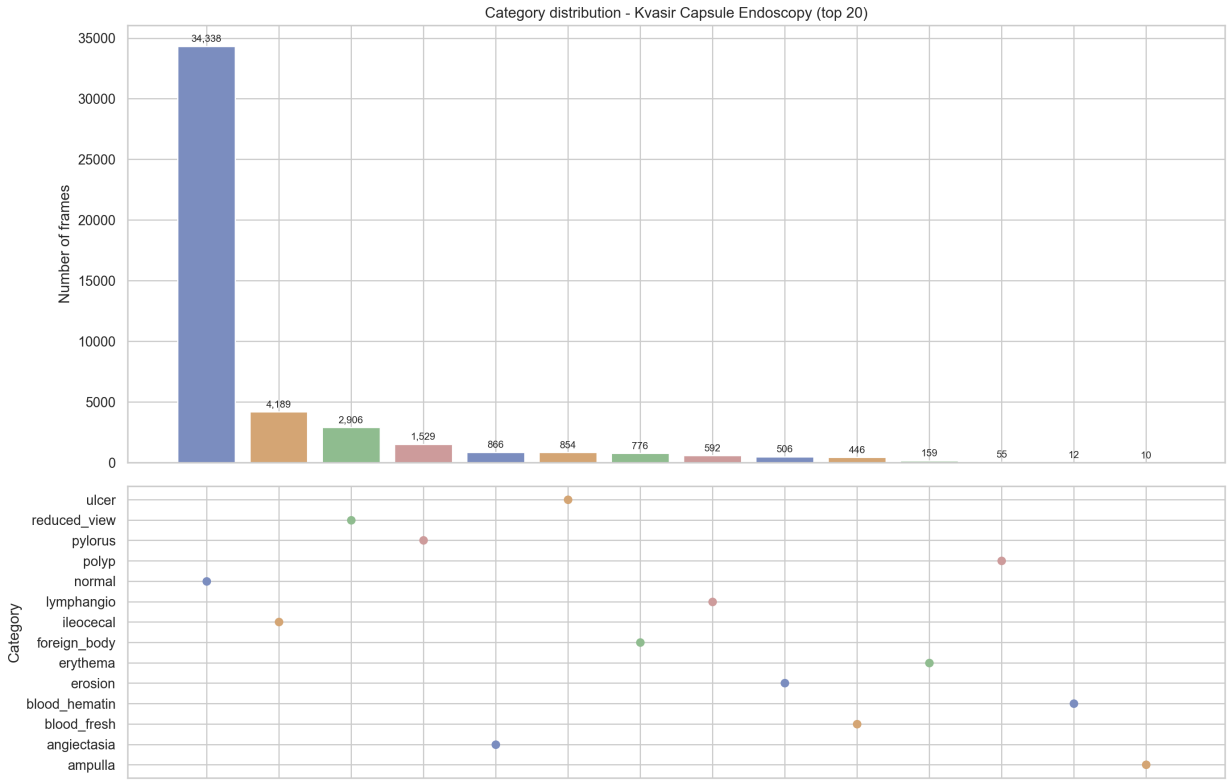


Figure 3: Kvasir-Capsule category distribution after mapping into the unified superclass framework.

ure 4 shows the category distribution.

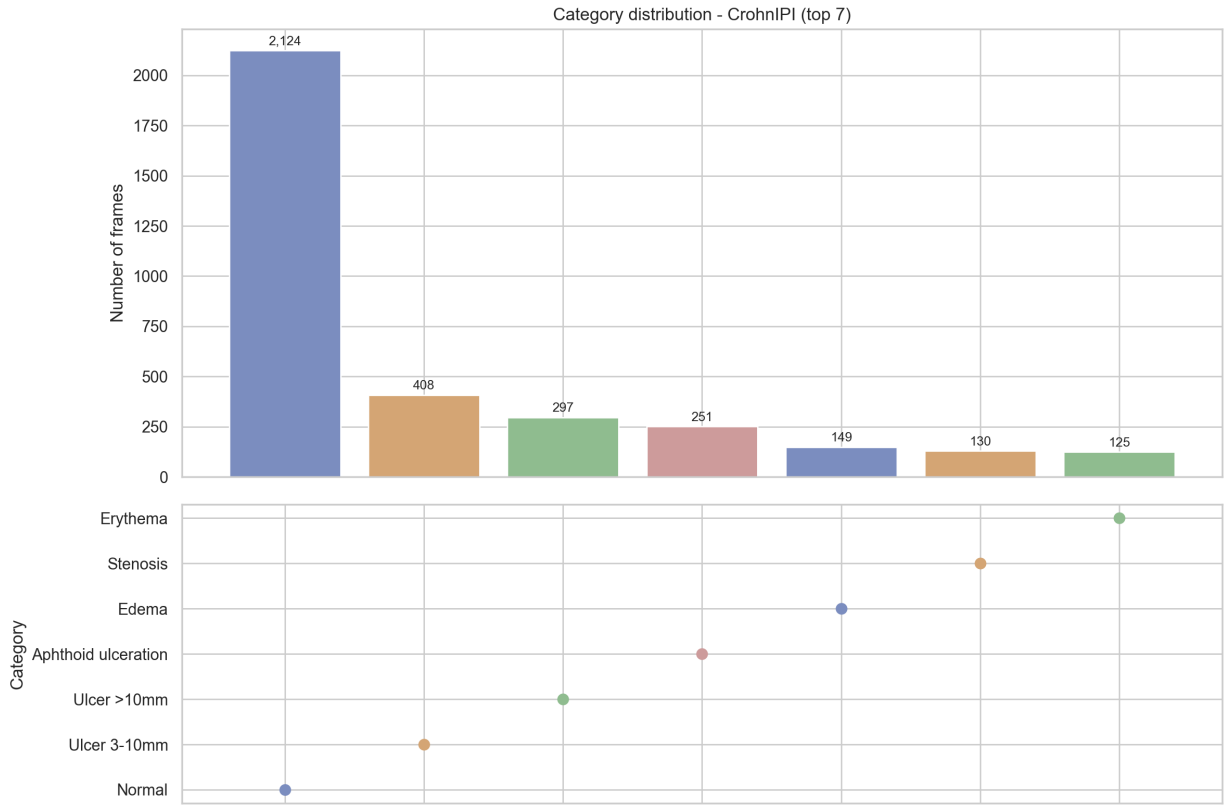


Figure 4: CrohnIPI category distribution after mapping into the unified superclass framework.

These three datasets differed not only in scale but also in annotation philosophy, metadata richness, and structural organization. GALAR was multi-label and highly redundant, Kvasir-Capsule was single-label and video-structured, and CrohnIPI was frame-level without grouping information. This heterogeneity required explicit harmonization into a unified superclass framework before downstream modelling.

3.2 Harmonization and Superclass Definition

Across the available sources, a fundamental practical challenge emerged: each dataset adopted its own labeling rules and terminology. Pathologies are often described using synonyms, partially overlapping categories, or secondary manifestations, making direct aggregation across datasets ambiguous. This heterogeneity motivated the adoption of a unified labeling system in which all source-specific annotations are mapped into a consistent recognition scheme before any downstream modelling.

I defined three primary superclasses:

$$d \in \{C, O, N\}$$

where C denotes Crohn findings, O denotes Other findings, and N denotes Normal frames. This aggregation reflected the high-level triage objective of screening: reliably separating clinically relevant Crohn-related findings from normal frames and non-Crohn abnormalities.

The clinical rationale for this target definition draws on established guidance regarding the interpretation and relevance of capsule endoscopy findings [8]. Mucosal breaks—specifically ulcerations and erosions—represent the most reliable visual proxies for Crohn-related pathology, as reflected in severity scoring systems such as CECDAI and the Lewis score, which assign greater weight to ulceration extent and mucosal damage [9]. Subtle inflammatory signs (e.g., mild edema, hyperemia) may overlap with non-specific inflammation or drug-related injury and are therefore less suitable as primary modelling targets. Focusing on ulcerations and erosions as the core of the Crohn superclass ensures alignment between the computational objective and clinically grounded interpretation.

Because GALAR was multi-label, a conservative priority rule was adopted to resolve co-occurring annotations:

$$C > O > N$$

This rule assigned any frame containing at least one Crohn-related label to superclass C, even when other labels co-occurred. This choice was not merely semantic: in GALAR, approximately 3,488 multi-label frames exhibited Crohn-related findings co-occurring with Other findings (most frequently Blood). Applying the priority rule increased the effective Crohn count from 45,092 to 48,580 (+7.7%) and correspondingly reduced Other from 114,902 to 111,414 (−3.0%). This reallocation had a direct structural consequence for the splitting optimizer. Because the Crohn superclass was the scarcest (∼48,000 frames out of ∼371,000), it acted as the dominant limiting constraint of the balancing problem: the 60/20/20 distribution constraint forced approximately 29,000 Crohn frames into Train, which in turn capped the maximum size of a balanced Train split. The +7.7% increase from the priority rule therefore enlarged the feasible region of the optimizer, allowing larger and better-balanced Train and Validation sets than would otherwise have been achievable under the same group integrity constraints.

Multi-label medical classification settings benefit from explicitly modelling both local label presence and global co-occurrence priors, as demonstrated in recent work on chest X-ray classification, where joint modelling of label interactions improves robustness and interpretability [21]. Although this thesis adopts a priority-based collapsing strategy for screening purposes, the underlying multi-label structure remains clinically relevant.

Figure 5 provides empirical support for the priority rule. The vast majority of Crohn-related frames (31,811) carried only ulcer or erosion labels without any Other co-occurrence. Among the co-occurring frames, Blood was overwhelmingly dominant: the three largest mixed combinations (erosion+blood, ulcer+blood, and ulcer+erosion+blood) accounted for

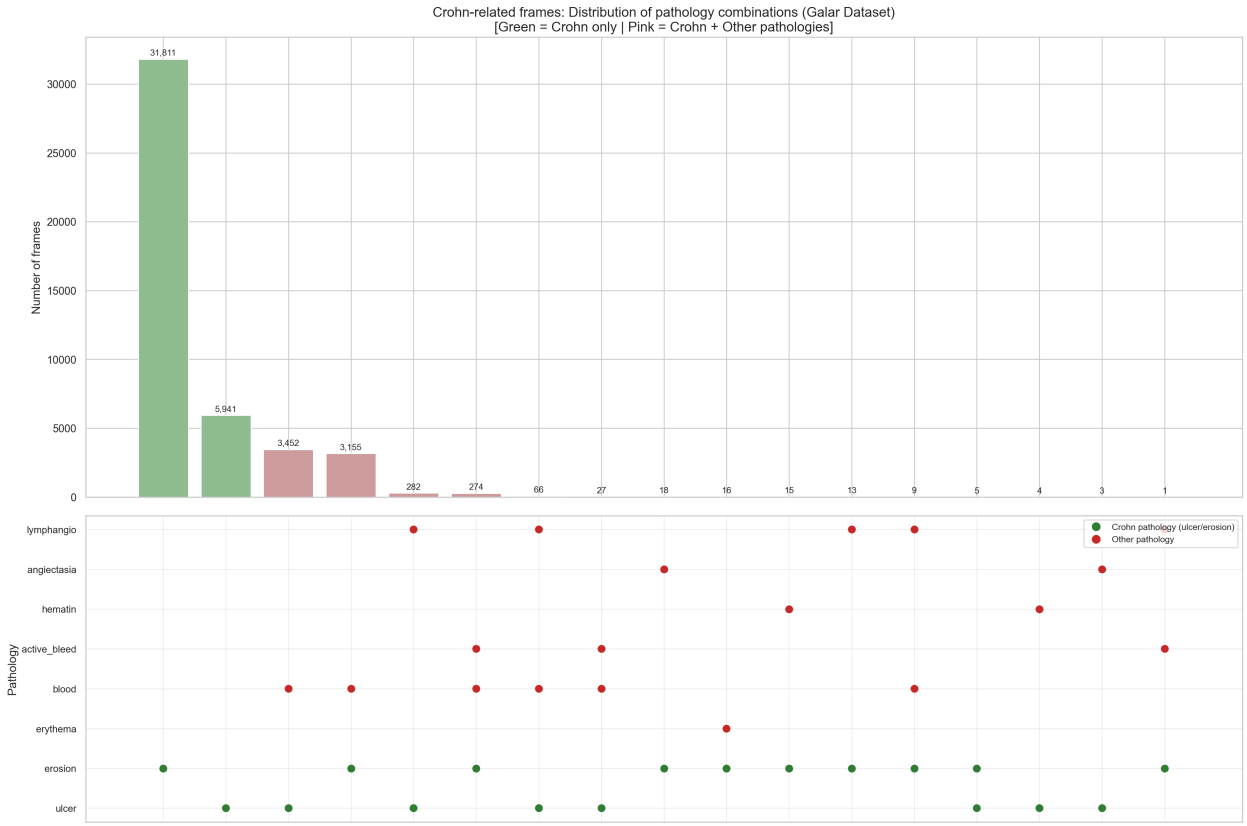


Figure 5: Distribution of pathology combinations in Crohn-related frames (GALAR). Green bars denote frames carrying only Crohn labels (ulcer/erosion); pink bars denote frames where Crohn labels co-occur with Other pathologies. The dot matrix below indicates which pathologies are present in each combination.

over 12,500 frames, while all non-Blood co-occurrences were marginal (< 300 frames each). This pattern had two implications. First, the priority rule reassigned a population that was small relative to the total Crohn count, limiting the perturbation on the Other superclass. Second, Blood emerged as the only Other pathology with substantial intersection with Crohn findings, motivating the dedicated Blood reduction protocol described in Chapter 4, which explicitly protects these multi-label frames from pruning.

3.3 Early Curation Strategy on GALAR

From the outset, the primary analytical focus was placed on GALAR. This choice was motivated by its scale and structural properties. With more than 3.5 million original frames and high sampling density from full-length videos, GALAR presented both the greatest opportunity for controlled redundancy reduction and the most significant computational challenge. The dataset's size required careful management of memory and storage resources during local processing, and its redundancy necessitated structured pruning to avoid overwhelming the learning signal with near-duplicate frames.

The earliest curation phase consisted of removing frames associated with image-quality labels, specifically "No view" and "Reduced view." This decision should not be interpreted as a dismissal of low-visibility frames in clinical practice. Rather, it emerged from uncertainty regarding label semantics and class intersections. Empirical inspection revealed that both quality labels exhibited non-empty intersections with pathological categories. In particular, frames labeled "Reduced view" occasionally co-occurred with Crohn-related findings or other pathological annotations. Although such overlaps may be clinically plausible, the criteria governing mutual exclusivity were not clearly documented. Given this ambiguity, these frames were excluded from training in order to avoid introducing annotation uncertainty that could not be formally resolved within the modelling framework.

A second pruning step removed frames labeled "IBD" and "Cancer." These categories were peripheral to the screening objective of distinguishing Crohn-related findings from other abnormalities and normal mucosa. Their limited relevance to the central task and relatively low frequency justified their exclusion in order to maintain task coherence and reduce label noise.

Subsequently, a patient-level filtering procedure was introduced. Rather than pruning solely at frame level, entire patients were evaluated based on their contribution to label diversity and pathological relevance. A greedy strategy was employed to identify patients that predominantly contributed frames from common or non-pathological categories while offering limited representation of rare or clinically informative classes. Particular attention was paid to rare categories such as Active bleeding, Erythema, and Foreign Body. Patients exhibiting minimal representation of these rare categories, minimal Crohn-related labels (ulcers and erosion), and a disproportionately high ratio of non-pathological frames were considered candidates for removal. In an initial phase, a threshold of approximately 5% pathological

frames per patient was used for low-volume patients. This was followed by a more granular, patient-by-patient inspection emphasizing large patients with high absolute frame counts but limited informational diversity.

Patients 5, 8, 9, 13, 14, and 22 were exempted from removal because they were the only patients containing image-quality labels, making them structurally important for understanding annotation behavior. Prior to actual deletion, simulated removal scenarios were explored using spreadsheet-based heuristics to estimate the effect of candidate removals on label frequency distributions. This exploratory phase reduced the risk of inadvertently eliminating structurally informative patients and provided empirical justification for the greedy selection strategy.

This early curation phase preceded the embedding-based redundancy reduction described in the next chapter and ensured that subsequent pruning operated on a dataset already filtered for semantic consistency and task alignment.

3.4 Structural Challenges and Curation Goals

Rather than treating dataset curation as a cleaning or filtering stage, I frame the entire dataset construction pipeline as a *constrained multi-objective optimization problem*, jointly balancing clinical alignment (Crohn-priority preservation), statistical representativeness (inter-class balance across splits), and computational tractability (redundancy reduction under realistic hardware budgets). The design choices documented in this and the subsequent chapters should therefore be interpreted as optimization decisions—each motivated by an explicit objective or constraint—rather than heuristic preprocessing steps.

Three interrelated structural pathologies characterize the raw data and jointly motivate the curation pipeline.

Class imbalance and low prevalence. In VCE, clinically relevant findings are rare and normal anatomy can account for the vast majority of frames. This leads to severe class imbalance at both the frame level and group level (a small subset of patients can dominate the positive class). Under these conditions, naive training and evaluation can be dominated by the majority class, and the effective difficulty becomes maintaining high sensitivity without overwhelming false positives. Severe class imbalance is a well-recognized challenge in medical imaging applications, where minority pathological classes are often underrepresented and require dedicated sampling, reweighting, or structural mitigation strategies to avoid biased learning dynamics [22]. Importantly, this imbalance is not solely a reflection of clinical prevalence. It is structurally amplified by dense frame sampling from full-length capsule videos, where extended segments of normal mucosa generate massive frame redundancy. A single SBCE examination can produce tens of thousands of frames, of which the vast majority depict unremarkable intestinal wall; pathological frames, by contrast, are sparse and temporally concentrated around lesion sites. The observed class skew is therefore both

epidemiological (Crohn lesions are clinically rare) and *sampling-induced* (the acquisition protocol over-represents normal anatomy by design). Recognising this dual origin is essential because it implies that class-level rebalancing alone is insufficient: redundancy reduction must also target the sampling-induced component to avoid training on near-duplicate frames that inflate the majority class without adding informational content.

Patient-level data leakage. VCE data are not i.i.d. at the frame level: consecutive frames from the same patient/video share lighting conditions, mucosal appearance, device characteristics, and often near-duplicates. If frames from a given patient appear in multiple splits, the model can exploit patient-specific cues and overestimate generalization. This risk is well recognized in the endoscopic imaging community. Smedsrud et al. explicitly enforce video-wise separation in Kvasir-Capsule, ensuring that "no video is shared between splits" so that training frames remain independent from validation and test frames [19]. Jaspers et al. adopt the same principle at the patient level, implementing "a strict split on a patient basis ... to avoid data leakage and intra-patient bias" [23]. Following these established practices, patient-wise (or video-wise) splitting is treated as a hard constraint for sources that provide grouping identifiers (GALAR patients and Kvasir videos).

Temporal redundancy. Temporal redundancy is structural: large contiguous segments show minimal change. In a typical SBCE examination, the capsule may traverse several centimetres of visually similar mucosa, producing hundreds of frames that differ only in minor illumination or peristaltic shifts. Training on all frames is computationally wasteful and can bias learning toward trivial patterns; evaluation may be inflated if redundant neighbors appear in training. Preprocessing therefore targets redundancy reduction using embedding-guided clustering, as described in Chapter 4.

3.5 From Raw Data to a Curated Pool: Dataset Understanding

After the early curation described above, the remaining pool still exhibits extreme skew across macro-categories: Normal (1,835,772 frames) dominates; Other findings has 297,817 frames (Blood alone contributes 206,831); Crohn findings has 49,955 frames. This motivates dedicated reduction of Blood and aggressive pruning of Normal while keeping Crohn findings close to intact. In addition to pruning-based rebalancing, I have planned from the outset to evaluate data augmentation strategies aimed at increasing the effective representation of rare classes, particularly Crohn findings, in later experimental stages. Figure 6 shows the per-patient frame distribution across the initial pool, encompassing patients and videos from all three source datasets (GALAR, Kvasir-Capsule, and CrohnIPI); the distribution is bimodal, with a small subset of patients contributing a very large number of frames while many patients contribute comparatively fewer, underscoring the need for patient-aware curation strategies.

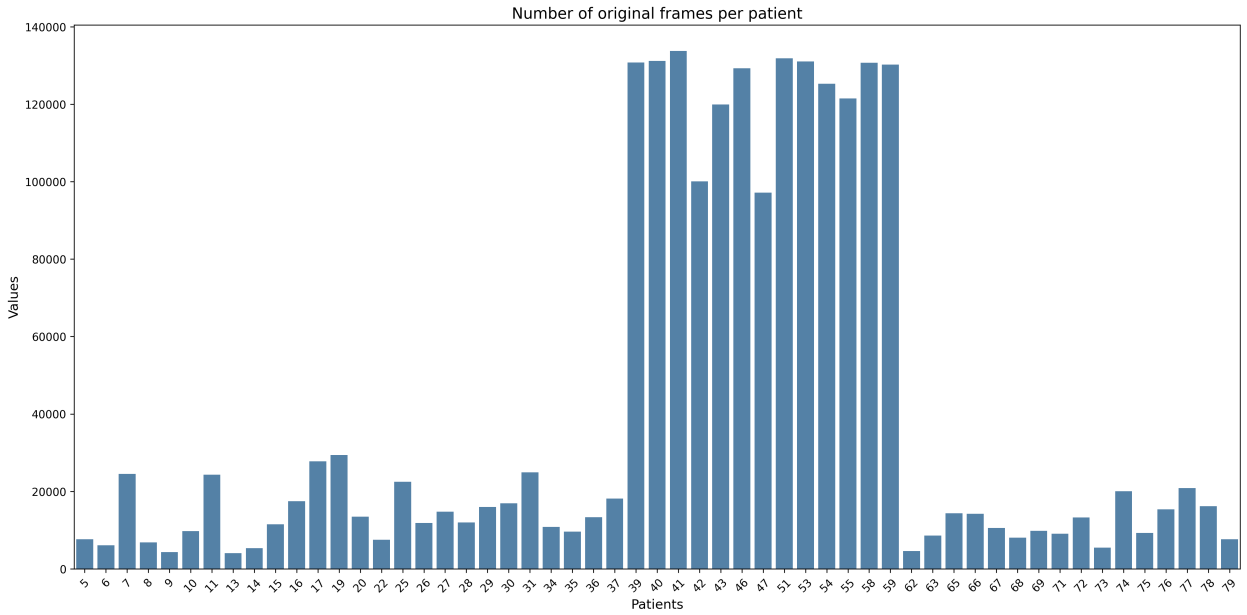


Figure 6: Total number of original frames per patient in the initial dataset pool.

3.6 Initial Dataset State Before Curation

To make the dataset construction defensible and reproducible, I report the pre-curation superclass counts used as the baseline for subsequent decisions: The Normal class accounts for 1,835,772 frames; Other findings amount to 297,817 frames, of which Blood accounts for 206,831; and Crohn findings total 49,955 frames.

The dominance of Blood within Other and the extreme redundancy of Normal introduce a risk of shortcut learning and gradient domination, where the model can over-focus on frequent patterns (e.g., normal mucosa texture or blood appearance) while under-learning rarer but clinically critical Crohn-related findings. This motivates both (i) dedicated reduction of Blood and (ii) embedding-guided redundancy reduction for Normal frames, while keeping Crohn findings close to intact. This risk is consistent with the broader phenomenon of spurious correlation and shortcut learning in machine learning, where models exploit statistically dominant but clinically irrelevant cues instead of learning robust disease representations, a behavior extensively documented in recent surveys on Clever Hans effects in deep models [24].

Figure 7 breaks down the per-patient distribution restricted to Normal frames across all three source datasets, revealing a trimodal structure—small, medium, and very large Normal-frame patients—which motivates patient-aware pruning rather than a uniform subsampling strategy. Figure 8 shows the analogous distribution for blood-only frames, again spanning GALAR, Kvasir-Capsule, and CrohnIPI; the set of patients visible in each figure differs because not all patients contribute frames to every class. This confirms that blood is not uniformly distributed across patients and can dominate the pathological signal, further motivating the dedicated reduction strategy described in Chapter 4.

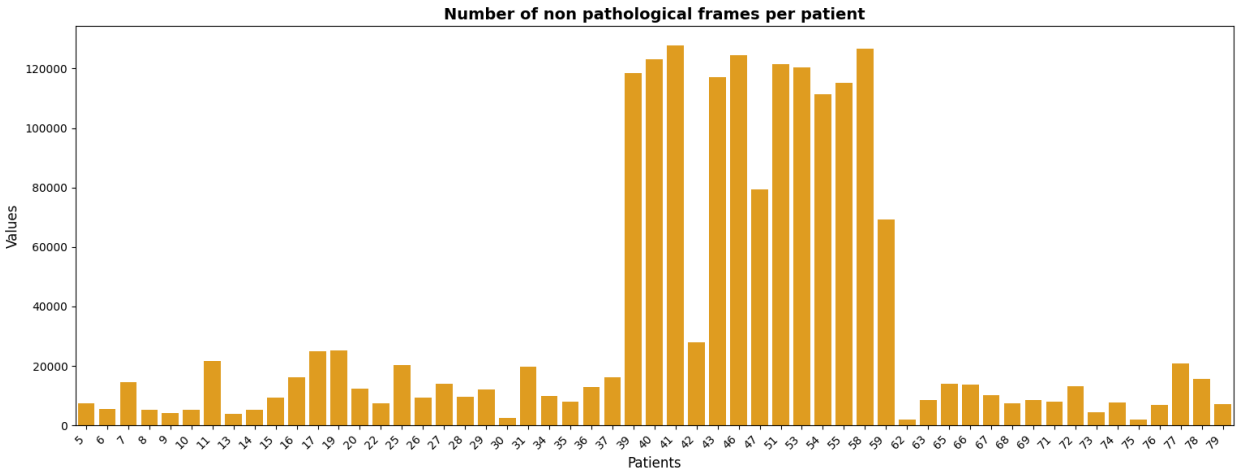


Figure 7: Number of original non-pathological (Normal) frames per patient.

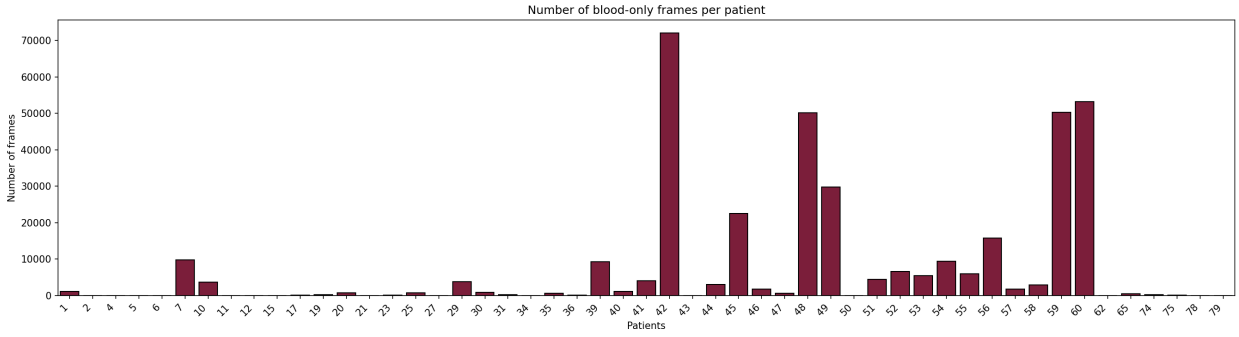


Figure 8: Distribution of blood-only frames per patient (frames where Blood is the only pathological label).

4 Redundancy Reduction and Pruning

Beyond statistical balancing, redundancy reduction also addressed *computational tractability*. Training on full-length SBCE streams would impose substantial I/O overhead, inflate memory requirements, and increase the risk of overfitting to near-duplicate frames that carry no additional discriminative signal. Pruning therefore served a dual role: *statistical correction*—bringing class proportions closer to a regime that supports balanced learning—and *computational stabilisation*—reducing the effective dataset size to a scale that is trainable under realistic hardware constraints without sacrificing the diversity of the underlying visual distribution.

4.1 Normal-Frame Pruning via Embedding-Guided Clustering

To reduce redundancy of Normal frames without discarding rare appearances, clustering was performed in DINOv2 embedding space and only representative frames were retained. The procedure was applied *per patient* (GALAR) and *per video* (Kvasir), never across groups, so that the pruning decision respected the group integrity constraints enforced by the splitting optimizer. The following subsubsections describe the diagnostic study that guided the selection of clustering parameters, the algorithm and its hyperparameters, and the final regime-aware configuration.

Embedding-Based Diagnostic Study and Regime-Aware Configuration

Before fixing pruning configuration, a diagnostic study was run on Normal-labeled embeddings to understand clustering behaviour as K increased. Four representative GALAR patients spanning very different Normal-frame counts (approximately 4k, 7k, 15k, and 120k) were used: patient 51 (large), patients 5 and 9 (small), patient 7 (medium).

Evaluation Metrics

Four complementary metrics characterized the effect of K. Mean intra-cluster distance measured compactness and decreased with K with diminishing returns. The Davies–Bouldin (DB) index captured the trade-off between separation and compactness, where lower values indicate better clustering. The Calinski–Harabasz (CH) index quantified between-cluster versus within-cluster dispersion, with higher values being preferable. Finally, coverage $p95$ reported the 95th percentile of point-to-centroid distances and served as a robustness-oriented coverage criterion. The resulting metric curves are reported for small patients in Figure 9, for the medium patient in Figure 10, and for the large patient in Figure 11.

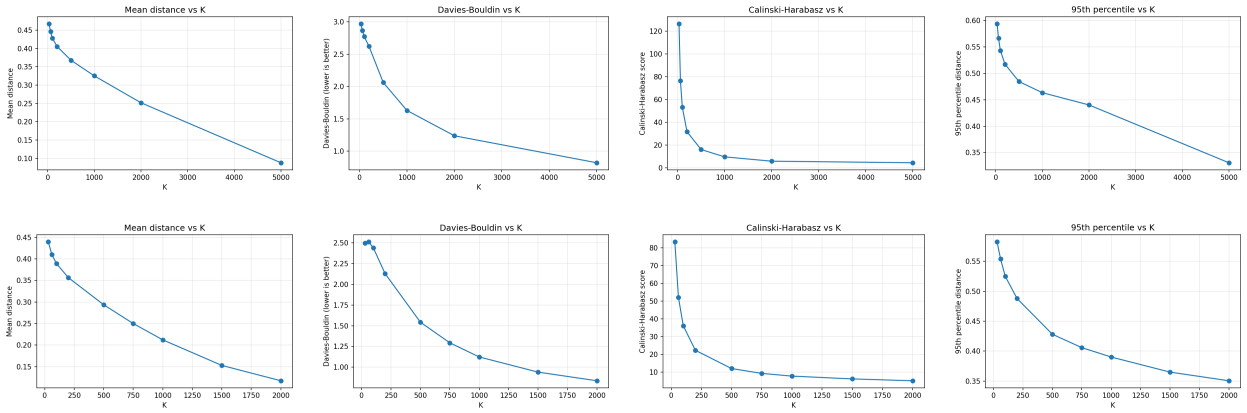


Figure 9: Small-patient Normal-embedding clustering diagnostics as a function of K (patients 5 and 9). Each row shows mean intra-cluster distance, Davies–Bouldin (DB), Calinski–Harabasz (CH), and coverage $p95$.

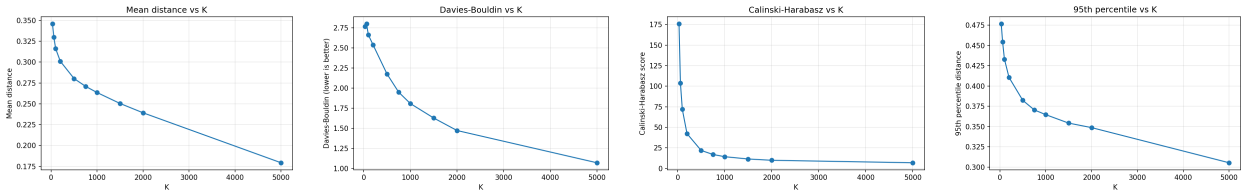


Figure 10: Medium-patient Normal-embedding clustering diagnostics as a function of K (patient 7).

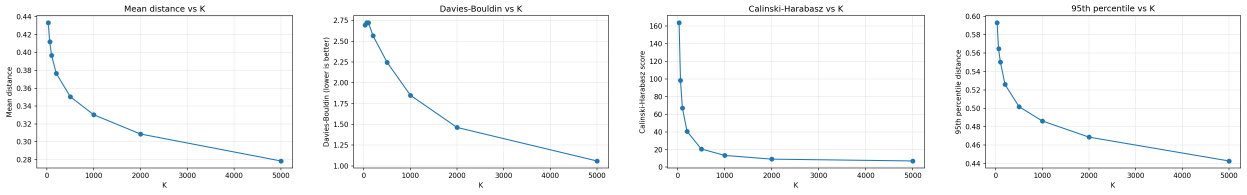


Figure 11: Large-patient Normal-embedding clustering diagnostics as a function of K (patient 51).

The diagnostic study naturally led to an iterative configuration process. The pruning configuration was not established in a single pass. An initial setup using high values of K combined with relatively high per-cluster retention fractions produced severe over-clustering: a large proportion of clusters contained only 1–3 frames. Under a per-cluster retention rule, such micro-clusters were effectively immune to pruning (a cluster of one frame always retains its only member), rendering the reduction step largely ineffective. Moreover, the excessive fragmentation distorted intra-patient variability estimates and reduced the statistical significance of centroid distances.

A systematic diagnostic study was therefore conducted, examining cluster-size distributions, distance-to-centroid distributions, coverage $p95$, Davies–Bouldin, and Calinski–Harabasz curves as functions of K. Among these metrics, mean intra-cluster distance and the Davies–Bouldin index proved the most informative for elbow identification: their curves exhibited clear inflection points that aligned consistently across patient sizes. Coverage $p95$ corroborated these findings in a supporting role, while Calinski–Harabasz was found unsuitable for selecting K on this dataset because the penalty term $(n - K)/(K - 1)$ dominates the ratio as K grows, producing a monotonically decreasing curve regardless of actual clustering quality. The analysis revealed a clear elbow region around $K \approx 300$–600 for small and medium patients, leading to the final configuration of $K = 600$ with a 20% retention fraction for this regime. Patients that had already been clustered under the earlier configuration were retroactively re-processed by applying the revised parameters directly to the original `.npz` embedding files. Because the DINOv2 encoder is deterministic (a given frame always produces the same embedding vector), the set of frames discarded under the first configuration is guaranteed to be a subset of those discarded under the second, more aggressive configuration. This *subset property* ensured that retroactive correction introduced no inconsistencies and required no re-extraction of embeddings.

Clustering Algorithm and Hyperparameters

Clustering uses k-means in embedding space with Euclidean distance (L2). Final runs use:

- Algorithm: MiniBatchKMeans (scikit-learn).
- Distance: Euclidean (L2) in 768-D DINOv2 embedding space.
- Initialization: k-means++.
- Restarts: $n_{\text{init}} = 10$.
- Iterations: max_iter = 300 (or convergence).
- Mini-batch size: 2048.
- Random seed: 42.

Regime-Aware Choice of K

The choice of K reflected a fundamental trade-off between *representativeness* and *efficiency*: overly coarse clustering risks collapsing semantically distinct visual regions into a single centroid, discarding frames that capture genuine anatomical or pathological variability; overly fine clustering preserves near-duplicate frames and undermines pruning efficiency. The regime-aware strategy adapts K to the size and diversity of each patient's embedding distribution, without fragmenting the embedding manifold into micro-clusters that would render the reduction step ineffective.

Empirically, elbow points occur at relatively high K: around 600–800 for small/medium patients, and up to 1500–2000 for the largest patients. The trend is broadly compatible with a sublinear heuristic $K \approx c\sqrt{N}$, but not applied rigidly because sampling rates differ across patients, affecting redundancy and effective diversity. In particular, some patients exhibit coarser effective sampling (larger temporal stride), reducing near-duplicates but increasing frame-to-frame variability; choosing K purely from N can under-partition their embedding distribution and discard genuinely distinct frames. For this reason, small patients intentionally use relatively high K.

Patient regime	Normal-frame count (approx.)	K	Keep fraction
Small / medium	$N < 25k$	600	20%
Intermediate	$25k \leq N \leq 100k$	1000	10%
Large	$N > 100k$	1500	5%

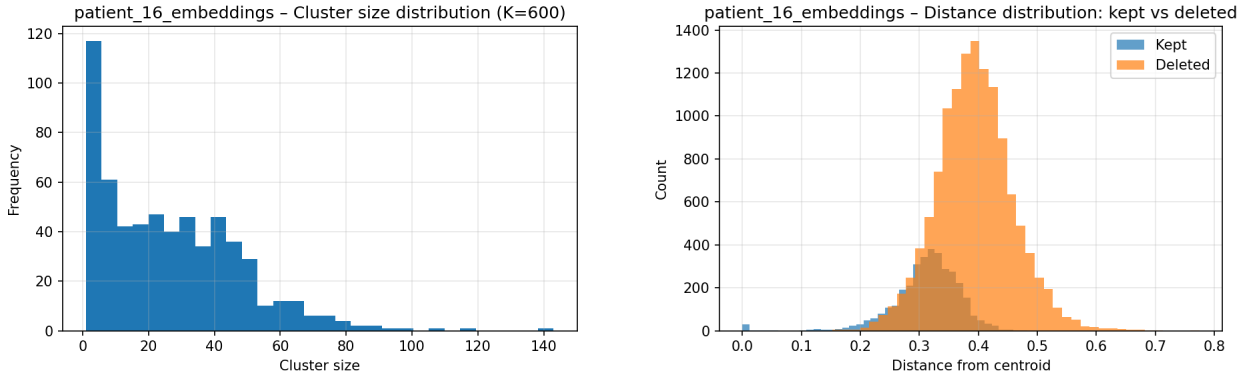
This discretized rule summarized the sweep evidence (mean distance, DB, CH, coverage $p95$). The elbow analysis showed that $K \approx 600$ was sufficient for small and medium patients, while larger patients required higher K to capture the additional visual diversity introduced by denser temporal sampling. The keep fraction decreases with patient size to achieve comparable absolute reduction targets across regimes.

4.2 Global Distance-Based Pruning

To address over-clustering artifacts—specifically the proliferation of micro-clusters containing only 1–3 frames, which are effectively immune to per-cluster retention quotas—per-cluster retention quotas were replaced by a global pruning strategy. After clustering, all frames in a group were ranked by distance to assigned centroid; the top-M most central frames were retained to match target cardinality. This preserved representativeness while avoiding disproportionate preservation of tiny clusters, and ensured that the pruning rate was governed by the overall embedding geometry rather than by cluster fragmentation artifacts.

Why global ranking is conceptually more robust.

Global ranking defined representativeness relative to the entire embedding distribution rather than imposing artificial per-cluster quotas under fragmented micro-clusters. Figure 12 illustrates the effect of global ranking on a representative patient: the cluster-size distribution shows the fragmentation pattern in embedding space, while the distance-to-centroid comparison confirms that the global rule retains central, representative frames and discards peripheral ones.



(a) Cluster-size distribution

(b) Kept vs deleted distances

Figure 12: Normal-frame pruning diagnostics (patient 16). Left: cluster-size distribution in embedding space. Right: distance-to-centroid distributions for kept vs deleted frames.

The pruning strategy concentrated reduction on structurally dominant classes (Normal and Blood) while preserving the already scarce Crohn signal to the greatest extent possible. Crohn findings were kept nearly intact, reflecting a clinically grounded prioritisation whereby, in a screening-oriented setting, sensitivity to Crohn-related lesions outweighs symmetric class reduction. Other findings (non-blood) were largely retained as-is, with only extremely over-represented patterns targeted.

4.3 Dedicated Reduction of the Blood Class Under Multi-Label Constraints

The dedicated reduction of Blood constituted a *bias mitigation strategy* targeting two distinct failure modes: shortcut learning and gradient domination. Blood was visually salient and over-represented within Other findings, accounting for 206,831 of 297,817 frames (69.4%). Left unaddressed, this dominance risked shortcut learning — where the model over-relies on blood appearance as a proxy for pathology — and gradient domination during training. The reduction procedure therefore targets blood-only frames specifically, while preserving multi-label co-occurrences.

Multi-label protection.

The reduction operated *exclusively* on blood-only frames, defined as frames where Blood (category 008) is the sole pathological label. Frames carrying Blood together with Crohn-related labels (categories 001–005) are mapped to the Crohn superclass by the priority rule (Section 3.2) and are never candidates for pruning. Similarly, frames with Blood co-occurring with other non-Crohn pathologies are retained in full. This design preserves clinically relevant co-occurrence patterns and avoids introducing spurious correlations between the reduction of one class and the diagnostic signal of another.

Design target for Blood reduction.

The target is to reduce blood-only frames from approximately 206,000 to an order of magnitude comparable with other pathological subclasses (approximately 30k–50k), balancing diversity preservation against gradient domination by the majority subclass during training.

Quantitative rationale and anti-bias constraint.

To set the reduction target on a principled basis, a multi-label distribution analysis was conducted on all GALAR frames carrying the Blood label (category 008). This analysis distinguished two populations: (i) frames where Blood co-occurs with at least one other pathological label (approximately 17,500 frames), and (ii) frames where Blood is the sole pathological annotation—*blood-only* frames (approximately 189,000 frames). The first group is protected by the multi-label rule described above and is never a candidate for pruning.

The reduction target for blood-only frames was set at approximately 30,000–50,000 rather than a lower value. This choice was motivated by the volume of co-occurring frames: if the blood-only count were reduced too aggressively—for instance, well below the $\sim$17,500 co-occurrence count—the resulting dataset would exhibit a disproportionate association between blood presence and Crohn-related labels (which dominate the co-occurrence group via the priority rule). Such an imbalance risks teaching the model the spurious shortcut "blood $\Rightarrow$ Crohn," undermining the clinical validity of predictions. Maintaining a blood-only volume that substantially exceeds the co-occurrence volume ensures that the Blood signal remains predominantly associated with the Other superclass, preserving the intended label semantics and preventing bias propagation into Crohn classification.

Contrast with Normal-frame pruning: from per-cluster quotas to global ranking.

Blood-only pruning shared the same embedding backbone (DINOv2-Base, ViT-B/14) and clustering algorithm (k-means in 768-dimensional embedding space) used for Normal-frame

pruning. However, two key aspects differ: the *frame selection rule* and the *parameterisation* of clustering.

Frame selection rule. For Normal frames, the pruning step applies a *per-cluster retention quota*: after clustering, each cluster independently retains a fixed percentage of its most central frames (those closest to the cluster centroid). This guarantees that every cluster — including small ones — contributes at least one representative to the reduced set. The rationale is that Normal frames span anatomically distinct regions of the gastrointestinal tract (oesophagus, stomach, small intestine, colon, ileocaecal valve) as well as varying image quality conditions; a small cluster may legitimately represent a rare but genuine anatomical pattern, and discarding it entirely would reduce the visual coverage of the pruned dataset.

For blood-only frames, the per-cluster logic is replaced by a **global distance-based ranking**. After clustering, *all* blood-only frames of a patient are sorted globally by Euclidean distance to their assigned centroid, and only the $M = \max(\lceil 0.15 \cdot N \rceil,\ 50)$ most central frames are retained, regardless of which cluster they belong to. The clusters serve solely to identify each frame's reference centroid and compute its distance; they impose no retention quota. This means that entire small or isolated clusters can be discarded if their frames are far from any centroid.

The motivation for this change was twofold. First, blood-only frames were visually homogeneous — they depict mucosal surfaces with diffuse or localised blood, without the macroscopic anatomical diversity of Normal frames — so small clusters are more likely to contain outliers than rare meaningful patterns. Second, with fewer blood-only frames per patient than Normal frames, the per-cluster quota mechanism would over-preserve peripheral frames and limit the effective reduction rate. The global ranking eliminates both problems, retaining only the most representative frames of the overall distribution.

Clustering parameterisation. For Normal frames, the number of clusters K was determined per patient regime (small/medium/large) through diagnostic sweeps that revealed multimodal elbow behaviour (Section 4.1). For blood-only frames, the diagnostic study described below shows that a single value of K suffices for all patients regardless of frame count, owing to the unimodal nature of the blood-only embedding distribution.

Why blood-only embeddings are expected to be unimodal.

Normal frames span anatomically distinct regions of the gastrointestinal tract — oesophagus, stomach, small intestine, colon, ileocaecal valve — as well as varying image quality conditions (bubbles, reduced view, dirt). This structural heterogeneity produces a *multimodal* embedding distribution: the first clusters segregate these macroscopic visual categories, causing metrics such as Davies–Bouldin to drop sharply at low K (the characteristic elbow). The

number of clusters needed to capture this diversity therefore scales with patient size, motivating the regime-aware strategy described in Section 4.1.

Blood-only frames, by contrast, depict mucosal surfaces with diffuse or localised blood as the sole pathological finding. They lack the macroscopic anatomical diversity of Normal frames: the visual signal is dominated by colour (red/dark haematic) and texture (mucosal surface), with comparatively little structural variation across different gastrointestinal segments. Figure 13 provides visual evidence of this homogeneity: fifty randomly sampled blood frames from the Kvasir-Capsule dataset exhibit a strikingly uniform appearance, dominated by haematic colour and mucosal texture with minimal structural variation across frames. I therefore *hypothesised* that blood-only embeddings followed a unimodal distribution in which (i) clustering metrics improve incrementally rather than exhibiting elbows, and (ii) the functional form of these metrics is largely independent of patient frame count.

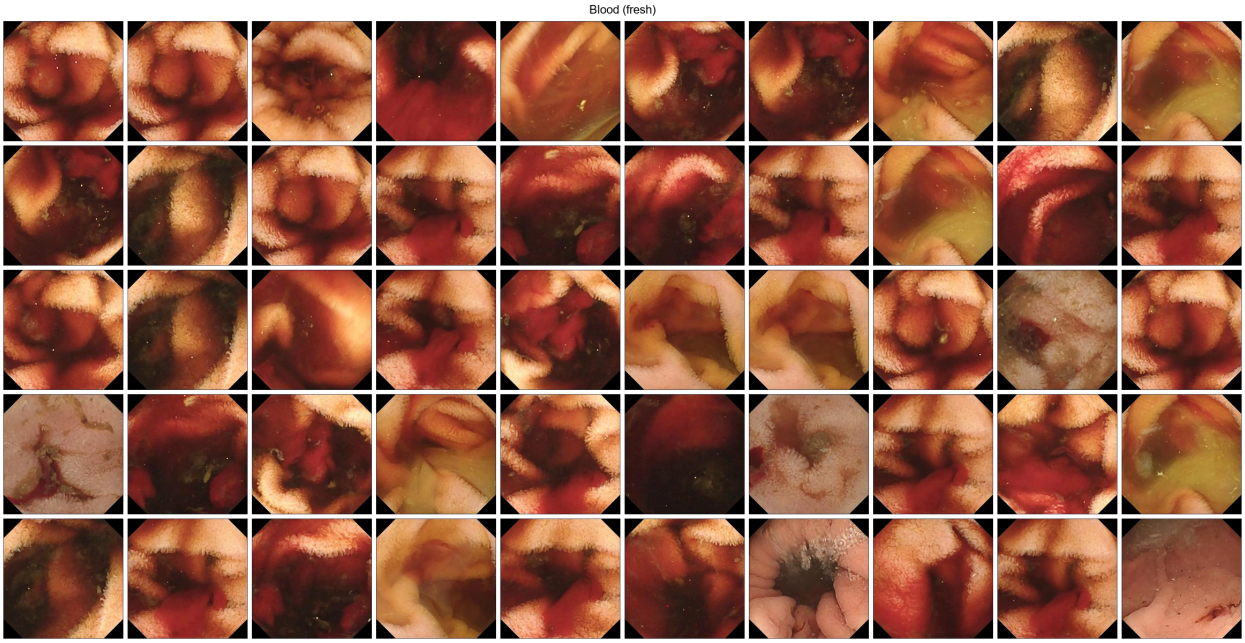


Figure 13: Fifty randomly sampled blood (fresh) frames from the Kvasir-Capsule dataset. The visual homogeneity—dominated by haematic colour and mucosal texture with minimal structural variation—supports the hypothesis that blood-only embeddings form a unimodal distribution in the latent space.

Empirical confirmation.

To test this hypothesis, the same diagnostic sweep conducted for Normal frames was repeated on blood-only embeddings for two patients at opposite extremes of the frame-count spectrum: patient 42 (72,037 blood-only frames) and patient 54 (9,416 frames — a 7.6× ratio).

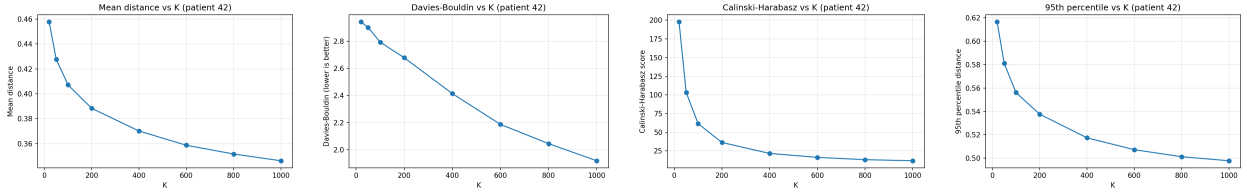
The results (Figure 14 and Figure 15) confirm both predictions. Davies–Bouldin decreases

approximately linearly with K, without any elbow, consistent with the absence of macroscopic visual categories: each additional cluster refines micro-variations rather than segregating qualitatively distinct structures. In the Normal case, by contrast, the first few clusters segregate macrostructures (e.g., different anatomical regions), causing DB to drop rapidly at low K. Mean intra-cluster distance starts at higher absolute values than for Normal frames—reflecting the tighter overall similarity of blood-only frames, which makes even coarse partitions reasonably compact—but decreases rapidly beyond $K \approx 400$; for Normal frames, the mean distance starts lower (diverse macrostructures are already well-separated at small K) but continues to decrease slowly, as higher K captures progressively finer distinctions within each anatomical region. Calinski–Harabasz declines monotonically in both the Normal and blood-only cases, confirming its unsuitability for K selection on this dataset: the penalty term $(n-K)/(K-1)$ dominates the ratio regardless of actual cluster quality, and the phenomenon presents itself in both patients with nearly overlapping curve shapes. Finally, coverage $p95$ follows a smooth, concave decreasing curve without inflection points, consistent with a single continuous distribution rather than a mixture of distinct visual regimes.

Crucially, the curves for patients 42 and 54 exhibit *the same functional form*, differing only by a vertical translation attributable to the different frame count N. The sole minor deviation is a slightly more concave Davies–Bouldin curve for patient 54: with fewer blood-only frames, the haematic regions are likely more spatially dispersed along the gastrointestinal tract, producing embeddings with greater inter-cluster separation and thus a faster DB improvement at low K. This deviation is small and does not alter the overall unimodal character.

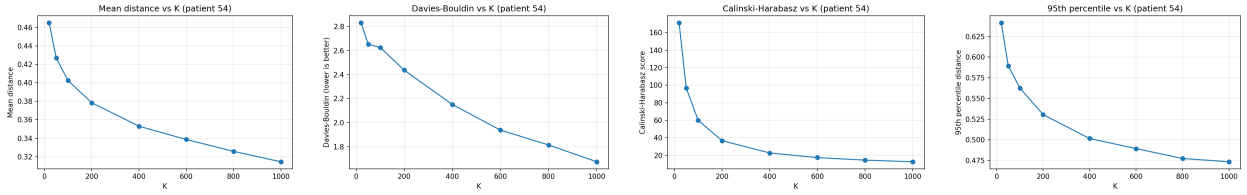
Operational conclusion.

The near-identical behaviour across a 7.6× range of frame counts confirms the unimodal nature of blood-only embeddings and eliminates the need for a regime-aware K strategy. A unified configuration was therefore adopted for all blood-only patients: $K = \min(N-1,\ 400)$, where 400 is the practical upper bound beyond which all metrics show diminishing returns; the retention fraction is set to 15%; and a minimum floor of 50 frames per patient ensures adequate representation even for patients with very few blood-only frames.



(a) Mean distance (b) Davies–Bouldin (c) Calinski–Harabasz (d) Coverage $p95$

Figure 14: Blood-only embedding clustering diagnostics as a function of K for patient 42 (72,037 blood-only frames — large patient). All four metrics show smooth, monotonic trends without elbows, consistent with a visually homogeneous embedding distribution.



(a) Mean distance (b) Davies–Bouldin (c) Calinski–Harabasz (d) Coverage $p95$

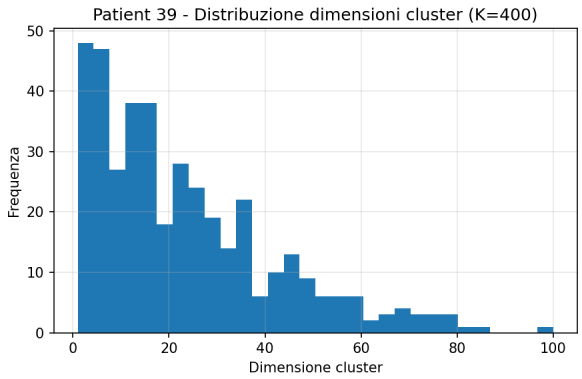
Figure 15: Blood-only embedding clustering diagnostics as a function of K for patient 54 (9,416 blood-only frames — small patient). The curves are nearly identical in shape to those of patient 42 (Figure 14), differing only by a vertical shift due to the smaller frame count. This confirms the unimodal nature of blood-only embeddings and justifies a single K=400 configuration for all patients.

Pruning outcome.

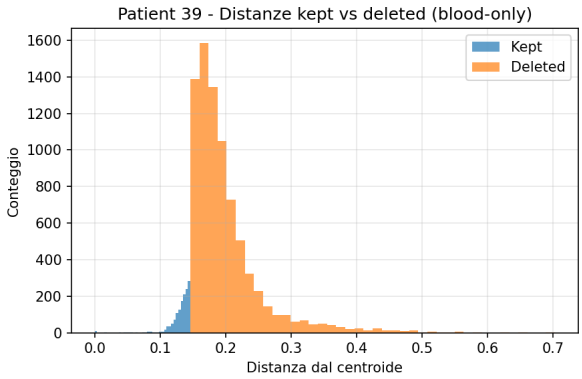
Pruning was applied to the 25 GALAR patients whose blood-only frame count warranted reduction. Across these patients, the procedure reduced blood-only frames from 188,746 to 28,416 (an overall reduction of 84.9%). The effective retention rate was approximately 15% for most patients; smaller patients reached the minimum floor of 50 frames, resulting in slightly higher retention percentages (e.g., 35.7% for patient 17 with 140 original frames). Patient 27, with only 15 blood-only frames, fell below the minimum floor and was retained in full. Figure 16 shows the cluster-size distribution and the kept-versus-deleted distance-to-centroid profiles for an example patient, confirming that the global ranking rule preferentially retains the most central frames. Figure 17 compares the per-patient blood-only frame counts before and after pruning across all 25 effectively pruned patients, illustrating the magnitude of the reduction while preserving cross-patient coverage. Finally, Figure 18 contrasts the multi-label co-occurrence patterns of the Blood label before and after pruning: the post-pruning distribution is reshaped to reduce the dominance of blood-only patterns while retaining clinically relevant combinations such as Blood co-occurring with Crohn-related findings.

4.4 Outcome Targets

The target ranges for the curated dataset are: Crohn findings 50k–70k frames, Other findings 120k–180k (with dedicated Blood reduction), and Normal 150k–200k (via aggressive redundancy pruning). These ranges were not arbitrary but reflected a deliberate trade-off between two competing requirements. On one hand, the curated dataset must substantially increase the relative frequency of Crohn findings compared to the raw pool, where Crohn accounts for fewer than 2.3% of frames; without this correction, the learning signal from clinically critical lesions would be dominated by the overwhelming majority of Normal and Blood frames. On the other hand, the target distribution must preserve the natural frequency hierarchy $N > O > C$, since inverting or equalising the class prevalences would create an ar-

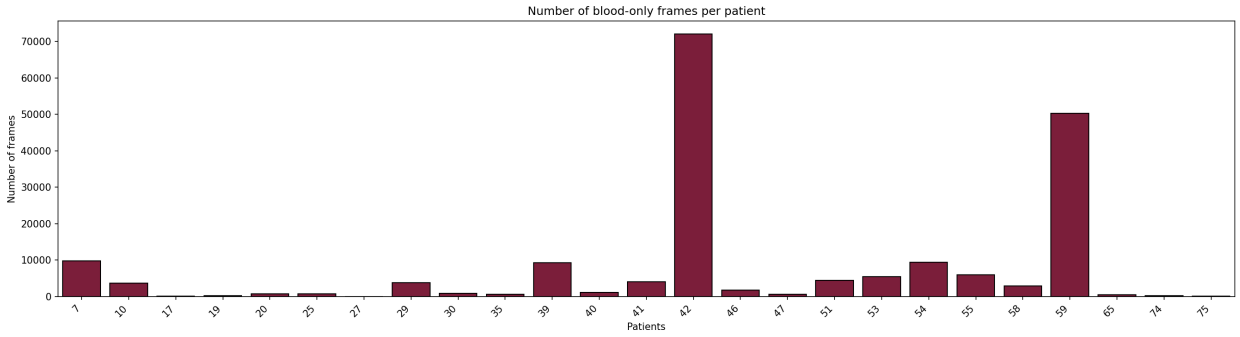


(a) Cluster-size distribution

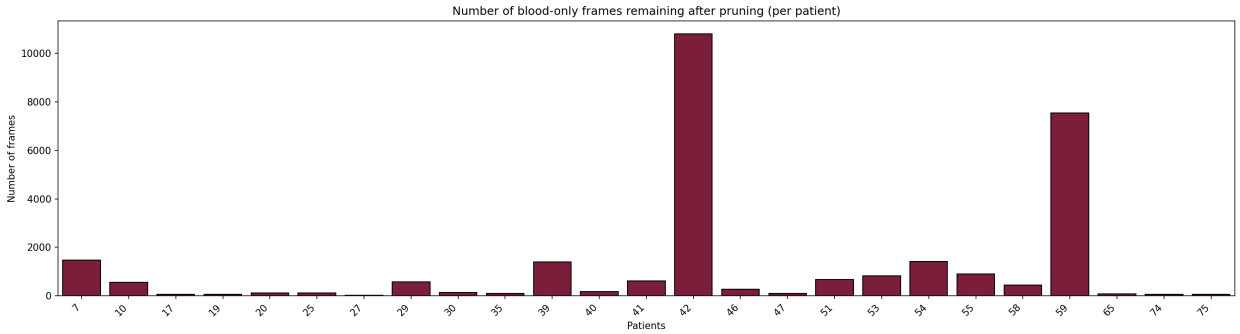


(b) Kept vs deleted distances

Figure 16: Blood-only pruning diagnostics (example patient). Left: cluster-size distribution in embedding space. Right: distance-to-centroid distributions for kept vs deleted frames under the global ranking rule.

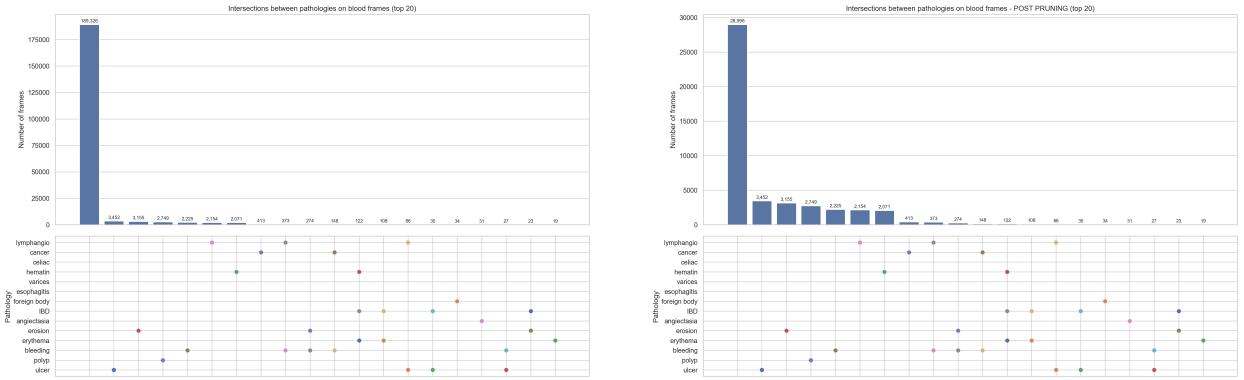


(a) Before pruning



(b) After pruning

Figure 17: Blood-only frame counts per patient before (top) and after (bottom) pruning. Only the 25 patients that underwent effective pruning are shown.



(a) Before pruning

(b) After pruning

Figure 18: GALAR blood co-occurrence patterns before (left) and after (right) pruning. Each bar shows the frequency of frames where the Blood label (category 008) appears alone or in combination with other pathological labels.

tificial training distribution that no longer reflects the epidemiological structure of screening data. The chosen ranges therefore represent the narrowest feasible gap between superclasses that still maintains the correct ordinal ranking, bringing the classes to comparable orders of magnitude while avoiding the extremes of both uncorrected imbalance and unrealistic symmetry. Importantly, internal subclass imbalance is deliberately tolerated when it reflects clinical reality—for instance, the heterogeneous frequency of individual pathologies within the Other superclass is preserved rather than artificially leveled, since it mirrors the genuine epidemiological distribution encountered in screening practice.

5 Dataset Overview After Curation

This chapter provided a dataset-level snapshot of the empirical pool after harmonization and early curation (Chapter 3), with the explicit goal of characterizing the label space and the internal structural differences across the three sources before the embedding-based redundancy reduction and the dedicated Blood mitigation described in Chapter 4. The analysis was intentionally conducted at the level of global category counts rather than patient-wise stratification: for the purposes of motivating the downstream pruning and splitting design, what matters at this stage is the aggregate geometry of the label distribution and the presence (or absence) of systematic co-occurrence patterns.

A key methodological caveat must be stated upfront. GALAR was a multi-label dataset, and the category totals reported for GALAR reflected multi-label counting: a single frame may contribute to multiple pathological categories. Consequently, dataset-level category totals in GALAR should not be interpreted as mutually exclusive partitions of the same frame pool, and the sum of category counts can exceed the number of frames. This was not a bookkeeping artifact, but the core structural feature that made harmonization and downstream bias-mitigation non-trivial: co-occurrence patterns, particularly those involving Blood, are precisely the mechanism through which shortcut learning and label-semantic leakage can emerge if reductions are applied naively. By contrast, Kvasir-Capsule and CrohnIPI are predominantly single-label in the configurations used here, and their category totals more closely approximate exclusive partitions of their respective pools.

5.1 GALAR: Multi-Label Combinations and Long-Tailed Co-Occurrence Structure

GALAR constitutes the structural backbone of the merged dataset because it combines scale with multi-label annotation richness. Figure 19 summarizes the distribution of the most frequent label combinations using an UpSet-style visualization—a set intersection chart in which each bar represents a distinct label combination and a dot matrix below encodes which labels are present in that combination.

The plot highlights two dominant properties that guide the methodological choices in subsequent chapters. First, the distribution is strongly long-tailed: a small set of combinations accounts for a substantial fraction of the observed label mass, while many combinations appear with relatively low counts. This asymmetry is expected in capsule endoscopy, where a large fraction of frames are visually unremarkable, and pathological findings occur sparsely and often cluster temporally around lesion segments. In multi-label settings, however, the tail is not merely a prevalence phenomenon; it also reflects annotation overlap patterns, where a small number of visually salient categories can repeatedly co-occur with clinically distinct lesions.

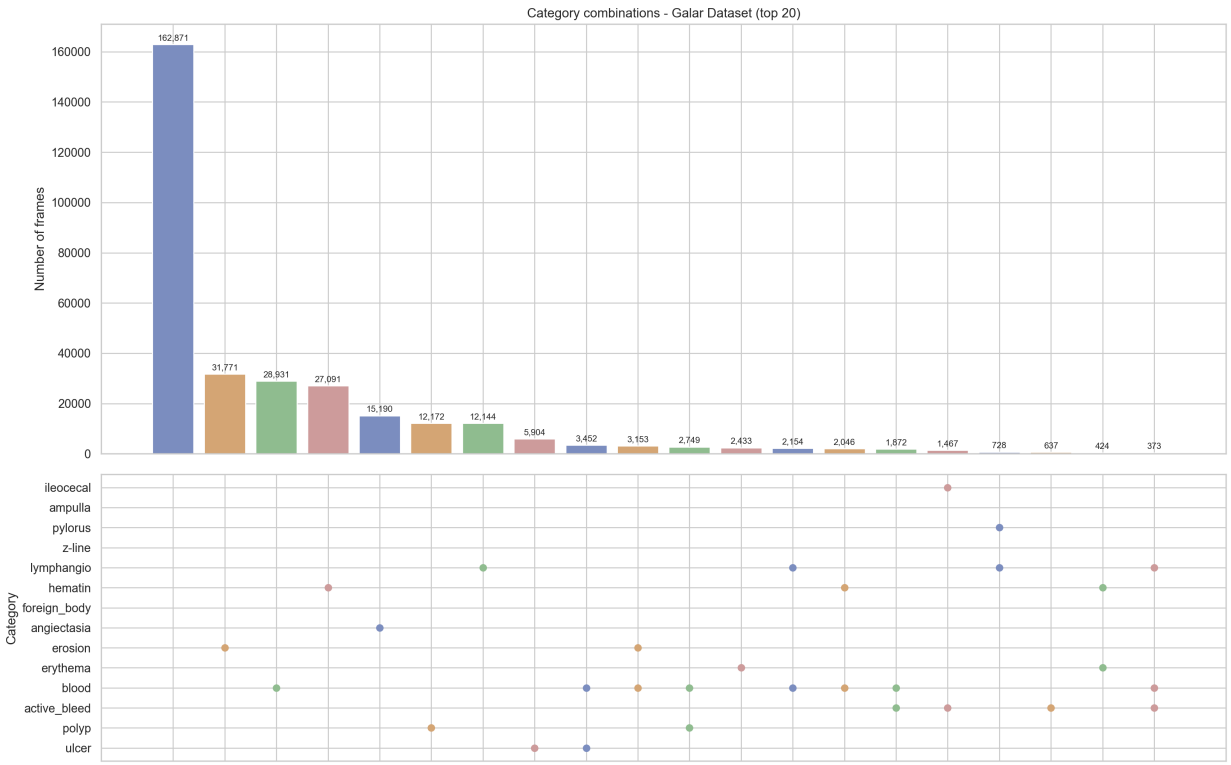


Figure 19: GALAR curated dataset composition: each bar represents a distinct multi-label combination, with the dot matrix below indicating which pathological categories are present.

Second, the UpSet structure makes the multi-label nature operationally visible. The combinations are not limited to single pathology labels but include mixed patterns in which specific categories—most importantly Blood and Crohn-related lesions—intersect non-negligibly. This observation is the bridge between Chapter 3's priority rule ($C > O > N$) and Chapter 4's dedicated Blood reduction: the same category that dominates the "Other" superclass in raw counts also becomes the dominant intersecting label with Crohn findings, meaning that any reduction strategy must explicitly preserve Crohn-co-occurrence frames while targeting blood-only redundancy. The visual summary therefore motivates treating Blood not as a generic "Other" label but as a structurally special subclass whose dominance and co-occurrence patterns can systematically bias learning dynamics if not controlled.

5.2 Kvasir-Capsule: Predominantly Single-Label Structure with Anatomy Enrichment

While GALAR supplies multi-label complexity and scale, Kvasir-Capsule contributes a large Normal pool and additional non-Crohn diversity under a substantially simpler label topology. Figure 20 shows the top category distribution for Kvasir in the same UpSet-style format.

The resulting structure is predominantly single-label: each bar typically corresponds to a single category rather than a multi-label combination. In practical terms, this means Kvasir plays a complementary role: instead of enriching co-occurrence structure, it broadens the negative and non-Crohn diversity and introduces a set of anatomy landmarks (e.g., pylorus, ileocecal valve, ampulla) that are absent or differently represented in the other sources.

From the standpoint of the three-superclass mapping, Kvasir includes clinically relevant Crohn proxies such as ulcer (854 frames) and erosion (592), but at lower absolute prevalence compared to its Normal pool (34,338). The dataset also contains blood-related labels (blood_fresh: 506, blood_hematin: 12) and several non-Crohn pathologies such as polyp, lymphangiectasia, and foreign body, each contributing additional "Other" heterogeneity. Kvasir also includes a `reduced_view` category, which—unlike the quality labels removed from GALAR during early curation (Section 3.3)—is retained and mapped to the Normal superclass under the harmonization scheme. Importantly, because Kvasir's structure is largely single-label in this configuration, these categories do not create the same multi-label co-occurrence risks observed in GALAR; rather, their primary value is coverage: they reduce the risk that the model conflates Crohn with a narrow "pathology vocabulary" learned only from GALAR, and they inject anatomy and quality variability that strengthens robustness under screening conditions.

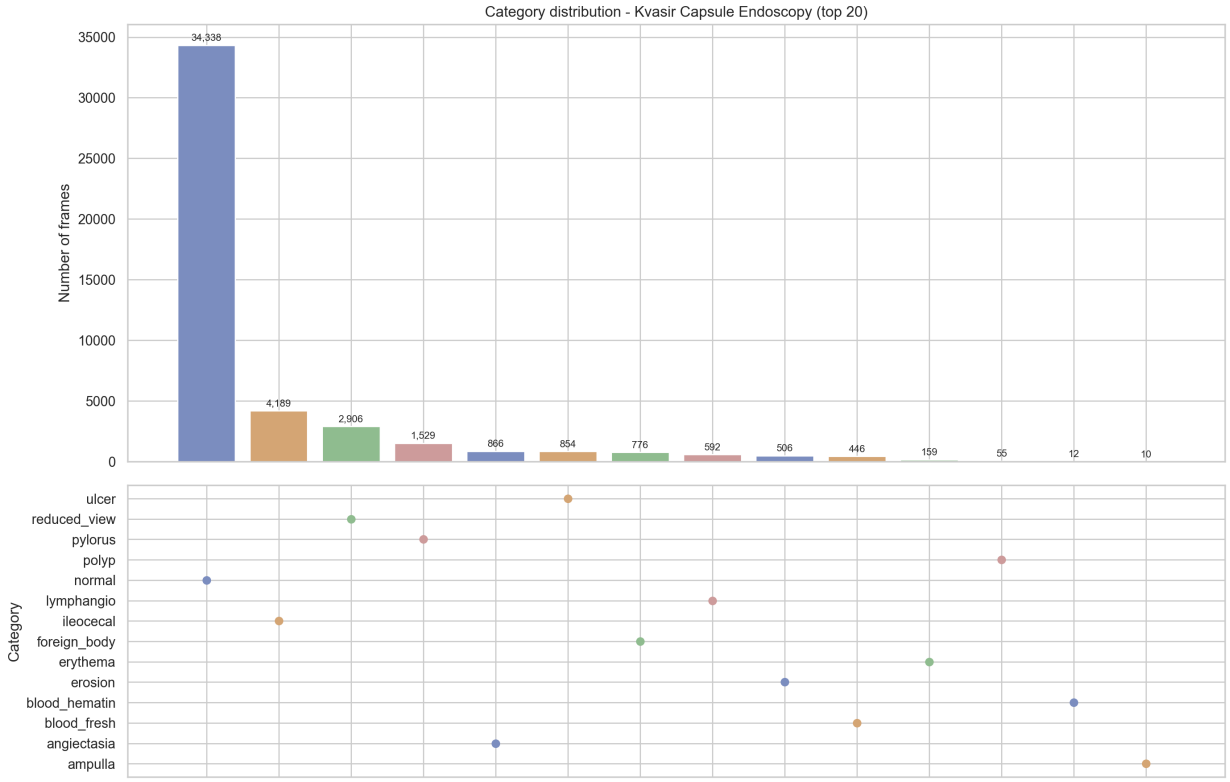


Figure 20: Category distribution of Kvasir Capsule Endoscopy (top 20 combinations). Kvasir is predominantly single-label: each bar corresponds to one category, with Normal accounting for the vast majority of frames (34,338). The dataset also contributes anatomical landmarks (pylorus, ileocecal, ampulla) and non-Crohn pathologies (polyp, lymphangio, foreign_body).

5.3 CrohnIPI: Crohn-Enriched Evidence with Ulcer Severity Granularity

CrohnIPI is structurally distinct from both GALAR and Kvasir. It is a smaller, Crohn-enriched dataset that provides granular ulcer severity grading, which is valuable for clinical interpretability even if the downstream task is ultimately cast as a three-superclass screening problem. Figure 21 shows the full category distribution.

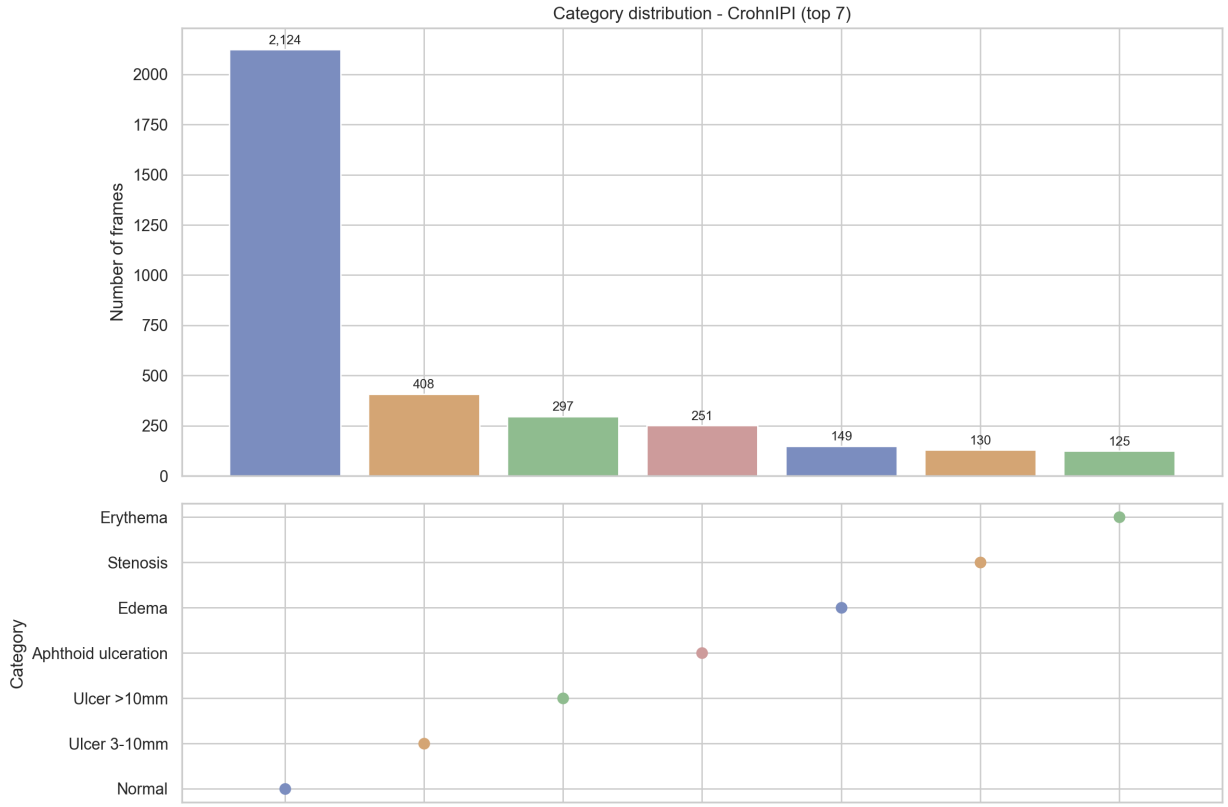


Figure 21: Category distribution of CrohnIPI (all 7 categories). CrohnIPI provides granular ulcer grading: Ulcer 3–10 mm (408), Ulcer >10 mm (297), and Aphthoid ulceration (251). It also includes Edema (149), Stenosis (130), and Erythema (125). Normal frames (2,124) serve as the negative class. All labels are single-label.

Unlike GALAR, CrohnIPI does not contribute multi-label co-occurrence structure; and unlike Kvasir, it is not primarily an anatomy-enrichment dataset. Instead, its primary contribution is to strengthen the minority Crohn signal through expert-grade annotations that explicitly separate ulcer severities (Ulcer 3–10 mm and Ulcer >10 mm) and include Aphthoid ulceration as an additional Crohn-relevant category. Aphthoid ulceration, being a recognized form of mucosal ulceration, is mapped to the Crohn superclass under the same harmonization rule that assigns ulcer and erosion labels to C. In the present pool, Normal frames (2,124) provide the negative class, while Crohn-related categories appear with moderate but meaningful counts. Additional categories such as Edema, Stenosis, and Erythema are present and

are mapped into the "Other" superclass under the harmonization scheme, since the screening objective is to separate Crohn-related lesion proxies (ulceration and erosion patterns) from non-Crohn abnormalities and normal mucosa. Under this mapping, the CrohnIPI superclass totals are: $C = 956$ (Aphthoid ulceration 251 + Ulcer 3–10 mm 408 + Ulcer >10 mm 297), $O = 404$ (Edema 149 + Stenosis 130 + Erythema 125), and $N = 2{,}124$, for a total of 3,484 frames. A further 14 frames were present in the image archive but absent from the original annotation CSV; these were assigned to the Test split as a fallback and are excluded from the CrohnIPI allocation counts.

Crucially, CrohnIPI lacks patient or video identifiers in the form required for strict group-wise splitting, which prevents it from being treated symmetrically to GALAR and Kvasir in the splitting optimization. This structural limitation explains the two-step protocol introduced in Chapter 6: CrohnIPI acts as a calibration reservoir that can be allocated at frame level after the group-wise assignment of the other sources has fixed the leakage-safe split skeleton.

5.4 Implications for Downstream Pruning and Split Design

Taken together, these three datasets define a merged pool with complementary structural properties. GALAR supplies scale and multi-label co-occurrence richness, but also introduces the strongest risk of dominance-driven shortcuts due to categories such as Blood and the inherent redundancy of densely sampled normal sequences. Kvasir contributes additional Normal volume, anatomy landmarks, and a diverse non-Crohn pathology vocabulary in a predominantly single-label regime, thereby broadening coverage while minimizing co-occurrence confounds. CrohnIPI strengthens the Crohn minority class with severity-graded ulcer annotations but must be handled under relaxed allocation due to missing grouping identifiers.

These complementary profiles motivate the subsequent methodological steps. The redundancy reduction protocol (Chapter 4) is concentrated where redundancy is structurally induced and where label imbalance is most severe, while the splitting strategy (Chapter 6) is explicitly designed to enforce group integrity where possible (GALAR patients, Kvasir videos) and to treat CrohnIPI as a controlled calibration component. Consequently, the dataset overview serves as the structural justification for why the pipeline must combine (i) embedding-guided pruning, (ii) targeted subclass mitigation, and (iii) constrained optimization for splitting under Crohn-priority constraints.

6 Patient-wise Splitting as a Combinatorial Optimization Problem

6.1 Problem Formulation as MDMWNPP

The constraints identified in the previous chapters — strict patient-wise separation to prevent data leakage, severe class imbalance with Crohn as a low-prevalence minority (Section 3.4), and the heterogeneous grouping structure across datasets (Section 5.4) — cannot be satisfied simultaneously by conventional random or stratified splitting. This motivated the formulation of the splitting task as a formal optimization problem, where all requirements are encoded as explicit objectives or constraints rather than addressed through informal heuristics.

Assigning patients to train/validation/test while balancing multiple labels can be cast as a multidimensional multi-way number partitioning problem (MDMWNPP), NP-hard in general [25]. Each indivisible group i is represented by a 3D count vector:

$$\mathbf{x}_i = (C_i, O_i, N_i)$$

The objective is to partition these vectors into splits $S = \{\text{tr}, \text{va}, \text{te}\}$ so that each split matches target constraints. Long-tailed and multi-label medical datasets require careful stratification strategies to prevent minority-class collapse during training. Recent work on weighted stratification in multi-label contrastive learning has shown that explicitly modelling comorbidity scores and class frequency can improve minority-class representation without disrupting label co-occurrence structure [26].

Connections to combinatorial optimization literature.

MDMWNPP generalizes classical number partitioning (NP-hard) to the simultaneous balancing of multiple numerical dimensions across more than two bins. The problem admits several complementary interpretations that situate the present formulation within well-studied algorithmic frameworks:

From a *vector scheduling* perspective, the splits can be viewed as "machines" and the superclass counts as "resource loads" to be balanced; this connects to the multidimensional packing and scheduling results of Chekuri and Khanna [27], who establish approximation guarantees for vector scheduling under capacity constraints. Alternatively, the formulation admits a *vector bin packing* interpretation, where each split acts as a "bin" with soft vectorial capacity targets and the objective minimizes deviations from these capacities. Finally, the treatment of all balance requirements as soft targets, penalized through L_1 deviations from ideal values, aligns with the *goal-programming* paradigm [28], where multiple potentially conflicting objectives are reconciled by minimizing weighted deviation variables.

From an algorithmic standpoint, the greedy baseline used for comparison (Section 6.3) is conceptually related to list scheduling [29], the classical online heuristic that assigns each job to the least-loaded machine. More recent work on optimal multi-way number partitioning [30] provides exact algorithms and tighter bounds for the one-dimensional case and motivates the use of exact solvers when the number of items is moderate, as in the present setting (98 indivisible units in Step 1).

6.2 Multi-Criteria Objective Function

Let $A_{s,d}$ denote the number of frames of superclass $d \in D = \{C, O, N\}$ assigned to split $s \in S$. The composite objective function combines three terms, each addressing a distinct methodological priority.

The first term enforces intra-split superclass balancing for the Train and Validation splits. For $s \in \{\text{tr}, \text{va}\}$, deviations from perfect equality among the three superclasses are penalised using an L_1 formulation over all pairwise differences:

$$L_{\text{bal}}(s) = |A_{s,C} - A_{s,O}| + |A_{s,C} - A_{s,N}| + |A_{s,O} - A_{s,N}|$$

This term ensures that the learning stage operates on a balanced class distribution, preventing gradient domination by the majority superclass and enabling model selection that is not confounded by prevalence artefacts. The use of absolute deviations rather than squared differences reflects a goal-programming perspective in which near-equality is preferred but exact equality is not required, given the indivisible nature of the patient and video groups.

The second term governs the distribution of Crohn frames across all three splits. Given target proportions $r = (0.60,\ 0.20,\ 0.20)$, the desired Crohn allocation for each split is $T_s = r_s \cdot A_C^{\text{tot}}$, where A_C^{tot} is the total number of Crohn frames available at the current optimisation stage. The corresponding penalty is:

$$L_{\text{crohn}} = \sum_{s \in S} |A_{s,C} - T_s|$$

This component receives a higher relative weight in the composite objective, reflecting the clinical priority of concentrating sufficient Crohn evidence in Train and Validation to support sensitivity-oriented modelling, while reserving a representative portion for unbiased evaluation in Test.

The third term expresses a soft preference for overall split sizes approximating a 70/15/15 ratio. Letting T_s^{split} denote the target total frame count for split s, the split-size penalty is:

$$L_{\text{split}} = \sum_{s \in S} \left| \sum_{d \in D} A_{s,d} - T_s^{\text{split}} \right|$$

This term is deliberately assigned a lower weight than the preceding two, since enforcing exact cardinality ratios can render the problem infeasible when combined with strict Crohn balancing and hard group integrity constraints.

The three components are combined into a single weighted objective:

$$L = w_{\text{bal}}\big(L_{\text{bal}}(\text{tr}) + L_{\text{bal}}(\text{va})\big) + w_{\text{crohn}} L_{\text{crohn}} + w_{\text{split}} L_{\text{split}}$$

The configuration adopted throughout this work uses $w_{\text{bal}} = 1.0$, $w_{\text{crohn}} = 3.0$, and $w_{\text{split}} = 0.3$, with w_{split} activated only in the second optimisation step. The choice of $w_{\text{crohn}} > w_{\text{bal}}$ encodes the structural observation that the Crohn superclass is the dominant limiting constraint of the partitioning problem: its scarcity constrains the maximum achievable balance in Train and Validation, and any deviation from the target Crohn distribution has disproportionate consequences for downstream sensitivity.

Split size is treated as soft to avoid infeasibility under strict Crohn balancing and hard group integrity. Enforcing near-perfect balancing for Train/Validation while respecting indivisible groups is already highly restrictive; a hard 70/15/15 constraint can admit no solution or can force clinically undesirable Crohn starvation in Train/Validation. In practice, the soft penalty ($w_{\text{split}} = 0.3$) steers total split sizes toward the desired ratio without rigidly constraining them, allowing the optimizer to prioritize the clinically more important balancing and Crohn-distribution objectives.

6.3 Exact Solving with CP-SAT

I solve assignment of indivisible groups (GALAR patients and Kvasir videos) using an exact CP-SAT approach. The hard constraint is each group belongs to exactly one split; the objective is the multi-criteria L_1 penalty above. OR-Tools CP-SAT implements a hybrid strategy combining lazy clause generation, linear relaxation, and SAT-based search [31]. Absolute-value terms $|A_{s,d} - A_{s,d'}|$ and $|A_{s,C} - T_s|$ are modelled natively via `AddAbsEquality`, avoiding big-M linearisations and preserving tightness of the relaxation. The solver configuration used throughout this work is: time limit 120 s, number of search workers set to auto-detect, random seed 42, and an objective scaling factor of 10^6 to maintain integer precision in the internal representation. This setup guarantees reproducible splits with explicit trade-offs, rather than ad-hoc randomization.

As a baseline, a greedy best-insertion heuristic is used: patients and videos are sorted by total frame count in descending order, and each unit is iteratively assigned to the split that minimizes the incremental increase in the composite loss L. This largest-first, best-fit strategy is a natural adaptation of list scheduling [29] to the multi-criteria setting: by placing the largest items first, the algorithm reduces the risk that late assignments cause large imbalances. The approach also connects to multi-way number partitioning heuristics [30], which employ similar ordering and insertion logic for balanced partitioning under cardinality

constraints. A quantitative comparison shows that exact solving is practically beneficial. On Step 1 (GALAR+Kvasir, 98 indivisible units), exact achieved loss 1,644.0 vs 26,972.0 for greedy (−93.9%). On Step 2 (adding CrohnIPI under loose allocation), exact achieved 121,798.8 vs 142,814.6 (−14.7%). The difference is not confined to the aggregate loss: the exact solver achieves Train balance deviation of approximately 0.34% versus 7.8% for greedy, and Validation balance deviation of 0.00% versus approximately 2.8%.

Metric	Exact	Greedy	Improvement
Loss Step 1	1,644	26,972	−93.9%
Loss Step 2	121,799	142,815	−14.7%
Train balance deviation	∼0.34%	∼7.8%	significant
Val balance deviation	0.00%	∼2.8%	significant

These results show that naive heuristics leave substantial performance on the table in constrained multi-objective splitting. The greedy strategy produces splits that appear superficially reasonable but carry residual imbalances that propagate into training dynamics and evaluation reliability. Figure 22 provides a multi-panel summary of the comparison. The comparable loss across both optimization steps (Panel A) shows that the addition of CrohnIPI improves the exact solution by −1.89% and the greedy baseline by −3.03%. Panel B reveals that the exact solver places the three superclass proportions on the 33.33% target line, while the greedy baseline leaves appreciable deviations that would directly affect gradient dynamics during training. Panel C illustrates the intentional asymmetry between balanced Train/Validation and naturally skewed Test under the exact solution. Panel F provides a device-level provenance breakdown, showing that the exact solution distributes capsule systems (PillCam SB3, SB2, COLON2, Olympus E10) across splits without concentrating any single device in a particular split, reducing the risk of device-specific confounders in evaluation.

6.4 Two-Step Protocol

Step 1 assigns all 55 GALAR patients and 43 Kvasir videos at group level, using the binary variables $x_{i,s}$ and the hard integrity constraint $\sum_s x_{i,s} = 1$. In this step only the balancing and Crohn-distribution terms are active ($w_{\text{split}} = 0$).

Step 2 introduces CrohnIPI frames under a relaxed, frame-level allocation scheme. Because CrohnIPI lacks patient or video grouping metadata, its frames can be distributed freely across splits. Integer variables $y_{s,d} \geq 0$ represent the number of CrohnIPI frames of superclass d assigned to split s, subject to the availability constraint

$$\sum_{s \in S} y_{s,d} = F_d \qquad \forall\, d \in D,$$

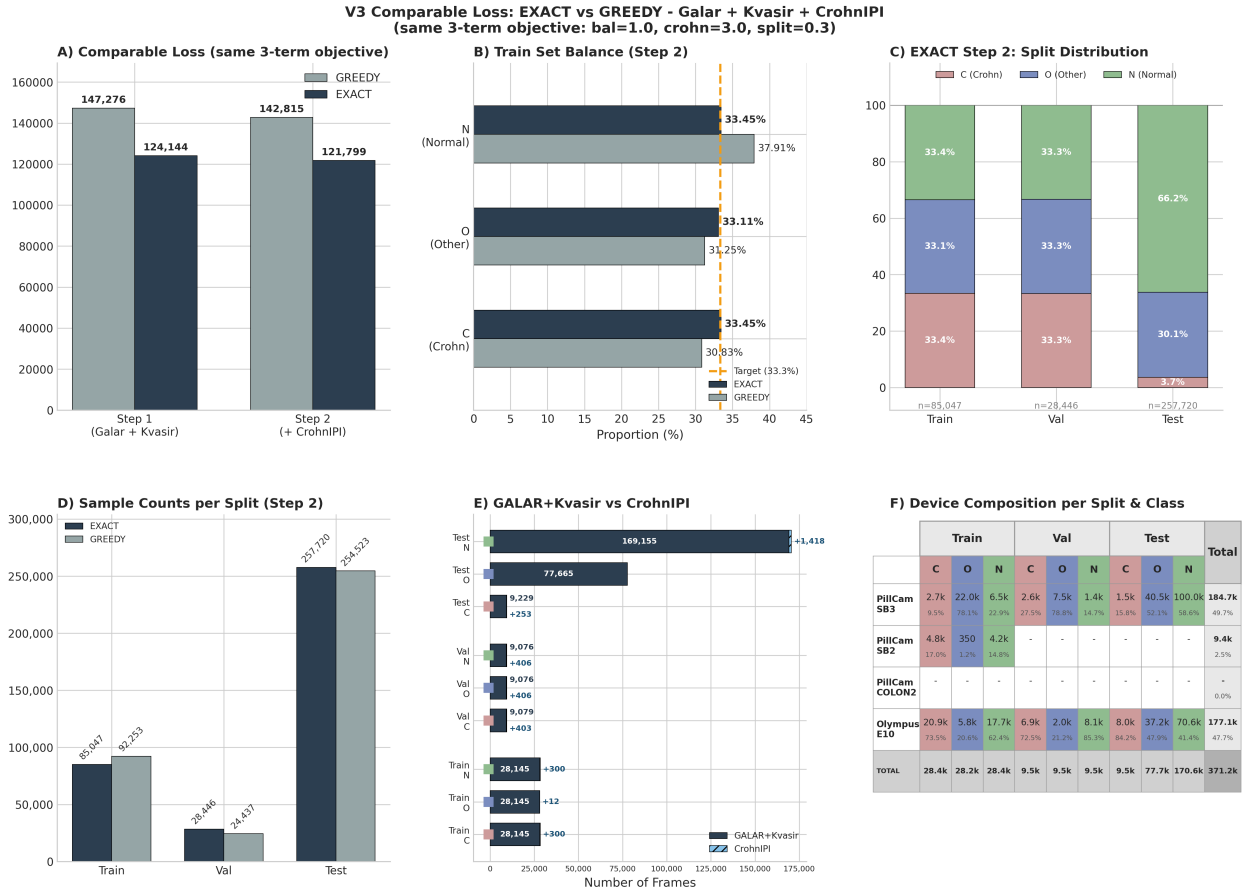


Figure 22: Multi-panel summary of the comparable-loss splitting optimization (exact vs greedy), evaluated under the same three-term objective function (balance w=1, Crohn distribution w=3, split-size w=0.3). Panel A: comparable loss across optimization steps. Panel B: Train set superclass balance. Panel C: per-split superclass composition under the exact solution. Panel D: absolute frame counts per split. Panel E: relative contribution of GALAR+Kvasir versus CrohnIPI. Panel F: device-level provenance breakdown per split and superclass.

where F_d is the total number of CrohnIPI frames of superclass d. The final per-split counts become $A^{(2)}_{s,d} = A^{(1)}_{s,d} + y_{s,d}$, and the split-size penalty L_{split} is activated ($w_{\text{split}} = 0.3$) to softly steer the overall distribution toward the 70/15/15 target. All Step 1 assignments are frozen; Step 2 can only improve or maintain the composite loss.

6.5 Target Coherence Across Optimization Steps

A subtle but important design choice concerns the Crohn distribution targets T_s used in each optimization step. In Step 1, the targets are computed from the total Crohn frames available in GALAR and Kvasir alone:

$$T^{(1)}_s = r_s \cdot \sum_i a_{i,C}, \qquad r = (0.60,\ 0.20,\ 0.20).$$

In Step 2, once CrohnIPI frames are introduced, the targets are *recalculated* on the basis of the new Crohn total:

$$T^{(2)}_s = r_s \cdot \left(\sum_i a_{i,C} + F_C \right).$$

This recalculation is essential. If Step 2 were to reuse the Step 1 targets, the addition of CrohnIPI Crohn frames would *structurally* increase the deviation terms $|A_{s,C} - T_s|$, penalizing the very improvement the second step is designed to achieve. Under stale targets, the optimizer would face a built-in penalty for incorporating new Crohn evidence—an incoherence that could lead to suboptimal or counterintuitive allocations.

With coherent targets, Step 2 strictly enlarges the feasible region: every Step 1 solution remains feasible (by setting all CrohnIPI allocations to zero), and any improvement from CrohnIPI allocation reduces the loss monotonically. This guarantee holds by construction and was verified empirically: the Step 2 loss never exceeded the Step 1 loss in any configuration tested. All split assignments are serialized in JSON format for full reproducibility.

6.6 Structural Consequences of Exact Optimization

The 93.9% reduction in composite loss achieved by the exact solver over the greedy baseline is not a marginal numerical improvement but reflects a qualitative difference in structural balance. The greedy heuristic, despite its largest-first ordering, accumulates imbalances early in the assignment sequence: once a large patient group has been irrevocably placed, subsequent assignments cannot compensate for the resulting skew. This early-assignment lock-in effect compounds across iterations, leaving substantial residual imbalance that propagates directly into the statistical structure of the resulting splits.

The intermediate Step 1 results (before CrohnIPI calibration) illustrate the solver's effectiveness. The exact solution assigns 9 GALAR patients and 6 Kvasir videos to Train, 6 patients and 4 videos to Validation, and 40 patients and 33 videos to Test, yielding the following superclass distribution across 367,715 frames:

Split	Total	C	O	N	%C	%O
Train	84,435	28,145	28,145	28,145	33.33	33.33
Validation	27,231	9,079	9,076	9,076	33.34	33.33
Test	256,049	9,229	77,665	169,155	3.60	30.33

Train achieves perfect three-way balance (33.33% per superclass), and Validation is within 0.01 percentage points of perfect equality. Step 2 calibration then distributes CrohnIPI frames to refine the solution:

Split	C	O	N	Total
Train	300	12	300	612
Validation	403	392	132	927
Test	253	0	1,692	1,945

After this calibration, the Crohn distribution target of 60/20/20 across splits is achieved exactly: 28,445 Crohn frames in Train (60.0%), 9,482 in Validation (20.0%), and 9,482 in Test (20.0%). In the exact solution, Train and Validation achieve near-perfect superclass balancing, while Test preserves the naturally skewed prevalence structure. This asymmetry is a deliberate design choice rather than a limitation of the solver. Balanced Train and Validation splits stabilize the optimization dynamics of the learning stage—gradient magnitudes across superclasses remain comparable, and model selection via validation metrics is not confounded by prevalence artifacts. Maintaining realistic skew in Test, by contrast, ensures that final evaluation reflects screening-like conditions rather than an artificially symmetric distribution. The practical consequence is that performance metrics computed on Test can be interpreted as estimates of operational behaviour under plausible clinical prevalence, rather than as artefacts of a balanced evaluation set that would never arise in deployment.

Treating patient-wise splitting as a formal optimization problem therefore materially alters the statistical geometry of the dataset within which learning occurs. The difference between the exact and greedy solutions is not confined to the objective value: it determines whether downstream evaluation reflects clinically meaningful generalization or structural artefacts of imbalance and leakage. It should be noted that the current formulation does not enforce a minimum number of patients per split, nor does it stratify covariates such as severity, pathology subtypes, or demographics; for larger-scale settings, metaheuristics or multi-objective Pareto formulations may improve scalability and enable explicit characterization of the trade-offs between competing objectives.

6.7 Physical Construction and Validation of the First Dataset

After computing the optimized assignments, the final dataset is materialized as three WebDataset tar archives (Train, Validation, Test) from a combined base archive. WebDataset is a tar-based data format for machine learning that stores each sample as a group of files sharing the same basename with different extensions (e.g., image and JSON metadata); its streaming design avoids full decompression and is natively compatible with PyTorch data loaders. A two-pass strategy is adopted to guarantee deterministic CrohnIPI allocation before streaming redistribution:

1. Load the optimized group-level assignments for GALAR and Kvasir, and the frame-level allocation for CrohnIPI.
2. First pass: collect CrohnIPI frames, compute superclass labels, and group by superclass.
3. Allocate CrohnIPI frames: for each superclass $\{C, O, N\}$, assign the first n_{train} frames to Train, the next n_{val} to Validation, and the remainder to Test, adapting to the observed export counts.
4. Second pass: stream through all samples and write each to the appropriate split archive, mapping by patient or video identifier (GALAR and Kvasir) or by frame identifier (CrohnIPI).

The resulting dataset contains 371,194 frames (9 fewer than the nominal 371,203), distributed as follows. The source provenance per split is reported in the second table.

Split	Total	Crohn	Other	Normal
Train	85,047	28,641	27,961	28,445
Validation	28,158	9,532	9,144	9,482
Test	257,989	13,148	74,264	170,577
Total	371,194	51,321	111,369	208,504

Split	GALAR	Kvasir	CrohnIPI
Train	82,318	2,117	612
Validation	25,290	1,941	927
Test	212,859	43,171	1,945

The total archive size is approximately 48 GB (Train 12.4 GB, Validation 4.2 GB, Test 31.6 GB). Two minor discrepancies were observed: 9 frames are missing relative to the

nominal count, likely due to entries lacking JSON metadata, and 14 frames with an empty source field were assigned to Test as a fallback and are excluded from the CrohnIPI provenance count. These discrepancies are marginal and do not affect the structural balancing guarantees. Additionally, the physical dataset build adapted the CrohnIPI frame-level allocation to the observed archive metadata, resulting in minor deviations from the optimizer targets. The actual per-split counts in the final archives differ from the optimizer targets reported above due to the frame-level redistribution at build time. The realized superclass distribution is:

Split	Crohn	Other	Normal	Total
Train	28,758	31,859	30,538	91,155
Validation	10,120	9,950	10,179	30,249
Test	8,522	73,481	167,787	249,790
Total	47,400	115,290	208,504	371,194

All downstream experiments (Sections 8–11) operate on these realized counts.

Six formal validation checks were verified on the constructed dataset:

- Hard patient constraint: each GALAR patient is assigned to exactly one split (55 unique patients).
- Hard video constraint: each Kvasir video is assigned to exactly one split (43 unique videos).
- CrohnIPI allocation completeness: per-split allocations sum to the total CrohnIPI frames per superclass.
- Train balancing: superclass proportions within 3.5 percentage points of perfect equality (31.5%/34.9%/33.5% for C/O/N).
- Validation balancing: superclass proportions within 0.7 percentage points of perfect equality (33.5%/32.9%/33.6%).
- Crohn distribution across splits: 60.7%/21.3%/18.0% (Train/Val/Test), close to the 60/20/20 optimizer target.

7 Feature Extraction Strategy and Data Augmentation

7.1 DINOv2 Embedding Extraction

This chapter formalizes the representation-learning choices adopted in this thesis and motivates them in relation to a screening-oriented objective. In capsule endoscopy, the central challenge is not the classifier design alone; rather, it lies in obtaining representations that remain stable across heterogeneous acquisition conditions, label co-occurrence patterns, and strong class imbalance, while still enabling computational efficiency and reproducible evaluation. For these reasons, the pipeline is designed around a representation-first approach: each frame is mapped into a compact embedding space by a pre-trained self-supervised encoder, and downstream learning is performed by low-capacity classifiers (linear models, shallow MLPs) operating on pre-computed embeddings. This separation improves experimental controllability and reproducibility of ablations, and allows dataset engineering and evaluation choices to be studied without entangling them with end-to-end optimization dynamics. At the same time, it remains consistent with the broader clinical motivation of workload reduction and triage in SBCE, where robust discrimination under realistic heterogeneity is central and efficiency constraints play a key role in determining the value of an AI-assisted workflow.

The encoder selected for this work is DINOv2 ViT-B/14 (`facebook/dinov2-base`), a Vision Transformer with patch size 14 that produces 768-dimensional embeddings [13]. DINOv2 is pre-trained using a self-supervised objective on LVD-142M, a curated dataset of 142 million images spanning diverse visual domains. The choice is motivated by two properties. First, DINOv2 representations achieve strong linear-probe performance across a wide range of downstream tasks, suggesting that the learned features are general-purpose and transferable to domains not explicitly represented during pre-training. This is particularly relevant for medical imaging pipelines where labeled data may be limited or unevenly distributed across institutions and cohorts. Second, the ViT-B/14 variant offers a trade-off between inference throughput, memory footprint, and representation capacity. It is expressive enough to capture fine-grained visual patterns relevant in SBCE—such as subtle mucosal texture differences—while still allowing inference on a mid-range GPU (NVIDIA GTX 1650, 4 GB VRAM) within acceptable time. Within this thesis, the encoder is treated as a stable representation provider rather than a component to be optimized end-to-end. This design choice is methodological: by keeping the feature extractor fixed, the downstream stages can be interpreted more directly, and performance differences can be attributed with higher confidence to dataset construction, balancing, augmentation, and classifier design rather than to shifts in the internal representation induced by fine-tuning. Figure 23 illustrates the architecture of a Vision Transformer, the backbone on which DINOv2 builds its feature extraction process: the input image is divided into fixed-size patches, each patch is linearly projected into a token embedding, and the sequence of tokens—augmented with positional encodings

and a learnable [CLS] token—is processed through a stack of Transformer encoder blocks. The final [CLS] representation serves as the global image embedding used by all downstream classifiers in this thesis.

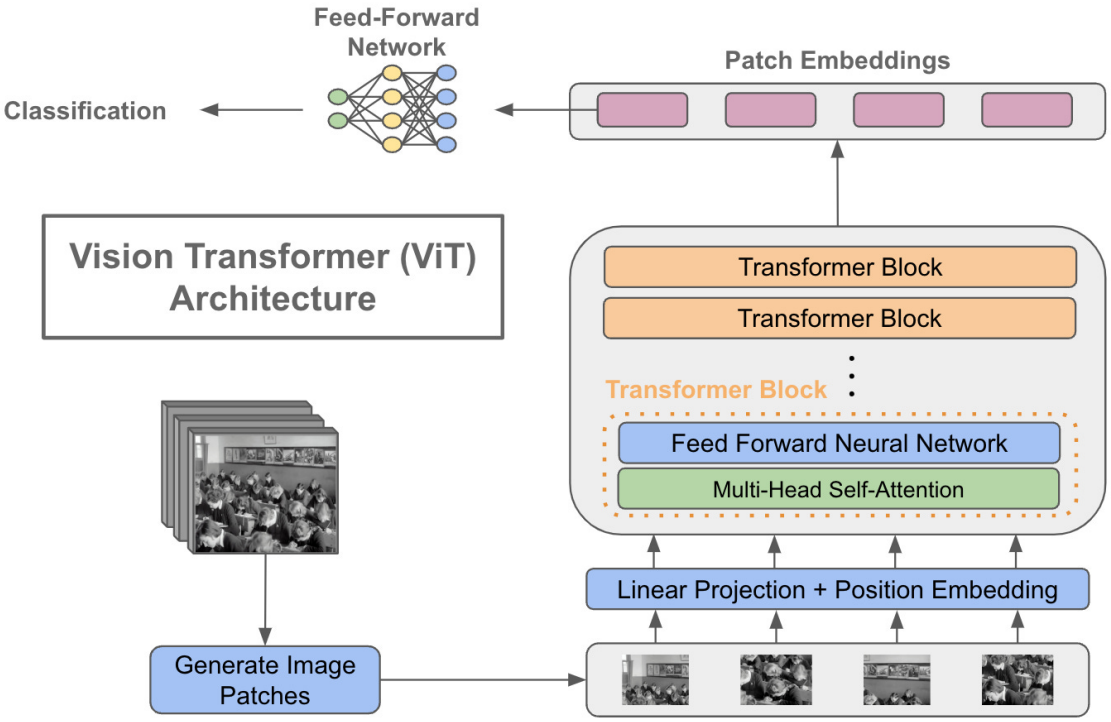


Figure 23: Architecture of the Vision Transformer (ViT). The input image is split into fixed-size patches, linearly embedded, and processed by a sequence of Transformer encoder layers. The output of the [CLS] token provides the global image representation that DINOv2 uses as its feature extraction mechanism.

Rather than adopting end-to-end fine-tuning, a two-stage pipeline is implemented: embeddings are extracted once with the frozen DINOv2 encoder, and parametrically small classification heads (linear models, shallow MLPs) are then trained on the pre-computed embeddings. This separation yields several advantages. Embedding extraction is a one-time cost; the same 768-dimensional vectors are reused across all classifier experiments without updating the encoder. Training a classification head on frozen embeddings is computationally inexpensive, enabling systematic exploration of hyperparameters, loss functions, and class-weighting strategies without re-running the encoder. The approach also avoids catastrophic forgetting of the general-purpose representation, since encoder weights are never updated. The main trade-off is that the encoder cannot adapt its internal representations to the specific appearance statistics of capsule endoscopy frames. If frozen features do not capture the discriminative cues required for fine-grained distinctions (for example, subtle erosions versus normal mucosa), this limitation may not be compensated by head tuning alone. Whether this ceiling matters depends on the empirical results in later chapters. Importantly, the viability of frozen-feature pipelines for SBCE has been independently supported in the literature. Varam et al. [32], for example, report strong performance on Kvasir-Capsule

classification using transfer learning from Vision Transformer representations and simple downstream models, indicating that pre-trained embeddings can carry sufficient discriminative information for VCE pathology classification even without domain-specific fine-tuning.

Each video frame is converted into a single 768-dimensional embedding through a deterministic extraction pipeline.

Preprocessing.

Frames are loaded from disk (JPEG or PNG) and converted to RGB. The `AutoImageProcessor` provided by the Hugging Face `transformers` library is used to apply the same preprocessing adopted during DINOv2 pre-training, including resizing to 224×224 pixels and channel-wise normalization using ImageNet mean and standard deviation. No additional preprocessing (such as histogram equalization, cropping, or color jittering) is applied, in order to maintain consistency and to avoid introducing uncontrolled domain-specific heuristics at the feature stage.

Forward pass and pooling.

The preprocessed tensor is passed through the frozen ViT-B/14 encoder. From the `last_hidden_state` output, the CLS token (position 0) is discarded and only the $16 \times 16 = 256$ patch tokens are retained. The final embedding is obtained by mean pooling over the patch-token dimension, yielding a single vector with 768 components. In preliminary experiments, this mean-pooled representation yielded higher and less variable validation Macro-F1 across classifier configurations than the CLS token alone. The resulting embedding is L2-normalized to unit length to ensure that downstream classifiers operate on a consistent scale and that similarity-based operations remain well-conditioned.

Precision and hardware.

Two extraction configurations were used over the course of the project. Initial runs operated in full FP32 precision with a fixed batch size of 32 on an NVIDIA GTX 1650 (4 GB VRAM). Later runs adopted FP16 mixed-precision inference (`torch.amp.autocast`) with auto-tuned batch sizes in the range 8–96 (typically around 64), which reduced extraction time substantially without measurable differences in downstream classification metrics. To avoid half-precision rounding effects during downstream training, stored embeddings are cast back to `float32`.

7.2 Offline Data Augmentation

Although Crohn findings are set to represent roughly one third of the training set by construction (as described in Chapter 6), in absolute terms they remain a minority of all frames and exhibit limited visual diversity because a substantial fraction originates from a relatively small number of patients. As Figure 24 illustrates, the limited visual diversity within each patient motivates offline augmentation to increase variability in the training set. To mitigate this limitation while preserving a clean evaluation protocol, offline augmentation is applied exclusively to the Crohn class within the training split. Validation and test sets are never augmented, so that evaluation metrics reflect the unmodified data distribution.

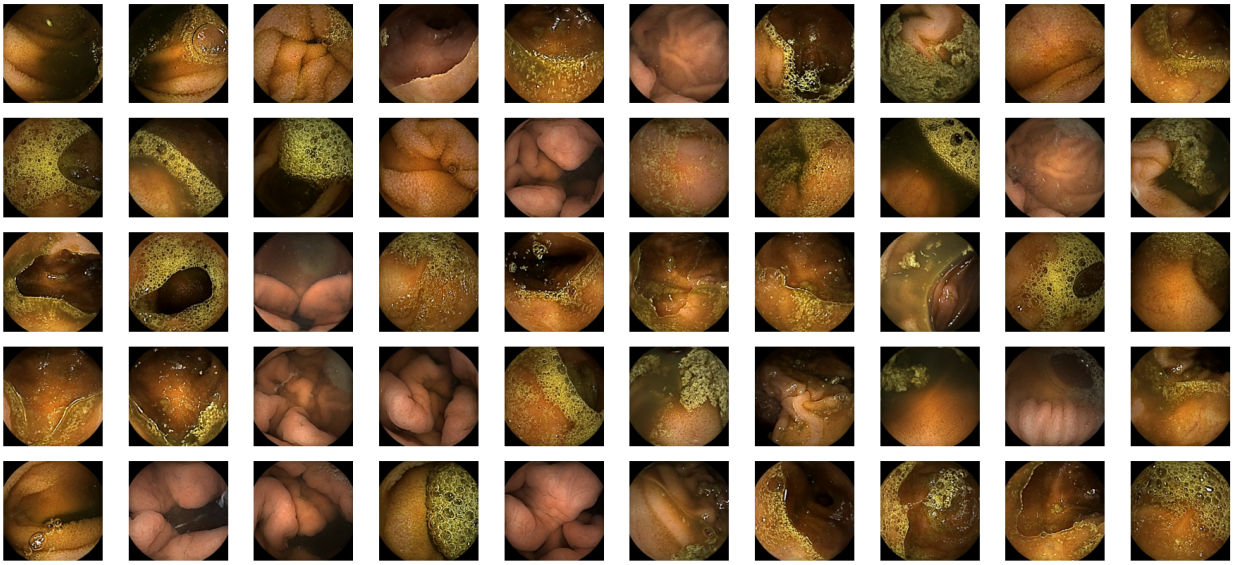


Figure 24: Representative sample of Crohn frames from a single patient (Patient 11) after pruning.

Augmentation strategy.

Augmentation is performed at the image level prior to embedding extraction. Each source frame is transformed to generate a new image instance, and the augmented image is then passed through the frozen DINOv2 encoder to produce a corresponding embedding. While more computationally expensive than perturbing embeddings directly, this design ensures that augmented representations remain valid points on the encoder's output manifold rather than arbitrary vector-space modifications.

Transforms.

The transforms are intentionally conservative and focus on geometric perturbations that reflect plausible variability in capsule imagery, which has no canonical orientation. Three

transforms are applied stochastically: horizontal flip (probability 0.5), vertical flip (probability 0.5), and rotation within $\pm 12^\circ$ (probability 0.3). If no transform is selected by the random draws, one flip is forced (horizontal or vertical with equal probability), ensuring that every augmented sample differs from its original. All random seeds are fixed to 42 to guarantee reproducibility.

This augmentation policy is supported by precedent in recent WCE work. Habe et al. [33] report a capsule endoscopy detection setting in which gastrointestinal frames are augmented using flips, 90° rotations, and small-angle rotations in the $\pm 12^\circ$ range, among other transformations, to improve robustness under realistic orientation variability. In the context of this thesis, adopting the same family of transforms provides an external methodological anchor: the perturbations are not arbitrary computer-vision heuristics but correspond to invariances considered reasonable in WCE analysis. The conservative rotation range is also deliberate: while small rotations are plausible and common in uncontrolled capsule motion, larger rotations can introduce boundary artifacts and interpolation effects that may distort clinically meaningful fine-grained texture cues.

Multiplier configurations.

Two augmentation intensities were tested:

- **x2**: one augmented version per Crohn frame, doubling the Crohn count from 28,758 to 57,516 frames. The augmented training set totals 119,913 frames with Crohn prevalence of approximately 48%.
- **x3**: two augmented versions per Crohn frame, tripling the Crohn count to 86,274 frames. The augmented training set totals 148,671 frames with Crohn prevalence of approximately 58%.

Augmented embeddings are stored in separate TAR archives rather than merged into the original training archive. This separation allows each experiment to include augmentation at a chosen intensity by concatenating archives at data-loading time. Each augmented sample retains metadata linking it to its source frame and recording the transforms applied.

The augmentation strategy is aligned with the broader screening and efficiency motivations discussed in the capsule endoscopy literature on frame reduction and reading-time reduction. Work such as Morera et al. [34] and Oh et al. [35] emphasizes that clinically meaningful gains arise when AI systems can reduce physician workload without compromising diagnostic integrity, which in turn depends on robust performance under uncontrolled frame variability. The conservative augmentation policy adopted here is intended to contribute to that robustness while preserving interpretability and avoiding distortions that would be clinically implausible.

8 Experiments with Linear Classifiers

8.1 Experimental Design

This chapter established a baseline for Crohn screening on frozen DINOv2 embeddings using three linear classifiers, each implementing a different loss function: Linear SVM (squared hinge loss with L2 regularization), Logistic Regression (cross-entropy loss with L2 regularization), and Ridge Classifier (least-squares loss with L2 regularization). The objective was twofold: to quantify what performance level was achievable with linear models on the pre-extracted 768-dimensional embeddings, and to verify whether the recall–precision trade-off for the Crohn class depended on the specific loss function or was a structural property of linear decision boundaries in this embedding space.

All three classifiers shared the same pipeline. The input embeddings were L2-normalized and fed to the multiclass classification head. Linear SVM and Ridge Classifier used a one-versus-rest (OvR) decomposition, while Logistic Regression used a multinomial (softmax) formulation. Model selection was performed via grid search over the regularization hyperparameter—$C \in \{0.001, 0.01, 0.1, 1, 10, 100\}$ for SVM and Logistic Regression, $\alpha \in \{0.1, 1, 10, 100, 1000\}$ for Ridge—using Macro-F1 on the validation set as the selection criterion. Two training configurations were evaluated for each classifier: a baseline without sample weighting (Exp0), and a patient-balanced configuration (Exp1) where each frame i belonging to patient p received weight $w_i = 1/n_{\text{frames}}(p)$, ensuring approximately equal per-patient contribution to the loss. This weighting reflected the fact that frames from the same patient are not independent observations—they depict the same mucosal surface and the same lesions—and that the clinically relevant unit of analysis is the patient, not the individual frame. All random seeds were fixed at 42 for reproducibility.

The linear classifier experiments operated on a multi-source embedding dataset combining GALAR, Kvasir-Capsule and CrohnIPI, totalling 371,194 frames grouped into 100 unique patient-level units (55 GALAR patients, 43 Kvasir videos, CrohnIPI as a single pseudo-group, and 1 unknown). Because CrohnIPI frames were allocated at frame level rather than group level (Section 6), CrohnIPI appears in all three splits; the group counts in the table below therefore reflect per-split appearances rather than unique groups. The dataset was split patient-wise with balancing prioritized for Train and Validation. The Test split was allowed to remain closer to natural prevalence so that final evaluation reflected realistic screening distribution rather than an artificially balanced scenario. The Crohn prevalence in the training and validation sets was approximately 33% (by construction), while in the test set it dropped to approximately 3.4%, with Other at 29.4% and Normal at 67.2%, reflecting the realistic clinical distribution.

Split	Frames	Groups
Train	91,155	14
Validation	30,249	7
Test	249,790	81

The per-class distribution across splits is shown below, highlighting the severe imbalance in the test set:

Split	Crohn	Other	Normal	Total
Train	28,758	31,859	30,538	91,155
Validation	10,120	9,950	10,179	30,249
Test	8,522	73,481	167,787	249,790

8.2 Evaluation Protocol

Each classifier was evaluated under multiple post-processing strategies to systematically test whether the recall–precision trade-off could be broken by better decision rules rather than better models.

The baseline evaluation used argmax prediction: the class with the highest decision score (or probability) was selected. In a screening context, the costs of misclassification are inherently asymmetric: missing a Crohn lesion (false negative) risks delayed diagnosis and disease progression, whereas a false positive only incurs an additional review by the clinician. This asymmetry motivated privileging recall over precision. To push Crohn recall beyond the argmax operating point, I applied a Crohn-first threshold rule: if the Crohn score exceeded a threshold τ, the frame was classified as Crohn regardless of the scores for Other and Normal; otherwise the prediction defaulted to the argmax over the remaining two classes. For each target recall level $r \in \{0.90, 0.95, 0.99\}$, the minimum threshold achieving Crohn recall $\geq r$ on the validation set was selected, choosing the threshold with the highest precision among all valid candidates.

For classifiers that did not natively produce calibrated probabilities (Linear SVM, Ridge Classifier), I applied Platt scaling via sigmoid calibration fitted on the validation set. Logistic Regression produced calibrated probabilities natively. The threshold tuning was then repeated on the calibrated probabilities to test whether better-calibrated scores improved the recall–precision trade-off.

As a third strategy, I applied Bayesian prior correction to the calibrated probabilities. The training set had approximately equal class proportions ($\pi_{\text{train}} \approx [0.33, 0.33, 0.33]$), but the expected clinical prevalence was $\pi_{\text{target}} = [0.03, 0.17, 0.80]$ for Crohn, Other, and Normal

respectively. These target priors were derived from the class proportions observed in the full GALAR dataset before balancing [18], used as an empirical proxy for the expected clinical prevalence; they were specified *a priori* and were not estimated from the test set, so no information leakage occurred. The discrepancy between the balanced training distribution and the skewed deployment distribution constituted a *prior shift*, and the correction aimed to realign the posteriors accordingly. The corrected posterior is [36]:

$$p_{\text{target}}(k \mid x) \propto p_{\text{train}}(k \mid x) \cdot \frac{\pi_{\text{target},k}}{\pi_{\text{train},k}},$$

with renormalization. This shifted the decision boundary to account for the realistic prevalence, effectively reducing the Crohn prior by a factor of $\sim$11$\times$. The threshold tuning was then applied to these corrected probabilities.

8.3 Results

The grid search revealed a marked interaction between patient weighting and regularization. Without sample weights, all three classifiers preferred strong regularization (SVM: $C = 0.01$, LogReg: $C = 0.01$, Ridge: $\alpha = 100$). With patient-balanced weights, the optimal regularization reversed direction: SVM and LogReg preferred $C = 100$ (minimal regularization), while Ridge preferred $\alpha = 0.1$. This reversal occurred because patient weighting substantially altered the effective loss landscape: patients with few frames received large per-frame weights, creating high-gradient regions that required more model capacity to fit. For Ridge, patient weighting was particularly destabilizing: at $\alpha \geq 1$, the model degenerated to predicting a single class, losing all Crohn discrimination.

The following table reports argmax test-set performance for all three classifiers under both weighting configurations:

Classifier	Weights	Macro-F1	PR-AUC$_{\text{C}}$	Recall$_{\text{C}}$	Precision$_{\text{C}}$
Linear SVM	None	0.474	0.210	0.403	0.155
Linear SVM	Patient	0.504	0.207	0.384	0.199
LogReg	None	0.450	0.177	0.392	0.159
LogReg	Patient	0.489	0.163	0.321	0.215
Ridge	None	0.461	0.196	0.408	0.148
Ridge	Patient	0.483	0.127	0.283	0.194

Patient weighting consistently improved Macro-F1 and Crohn precision at the cost of Crohn recall: the model became more conservative in its Crohn predictions, producing fewer but more reliable positive predictions. Linear SVM with patient weights achieved the best overall Macro-F1 (0.504) and was used as the reference linear model throughout the thesis. All three

classifiers exhibited a common pattern: Crohn precision under argmax ranged from 0.15 to 0.22, while Crohn PR-AUC ranged from 0.13 to 0.21, indicating poor ranking quality for the minority class regardless of the loss function. Figure 25 shows the grid search, Precision–Recall curve, and confusion matrices for the unweighted SVM baseline.

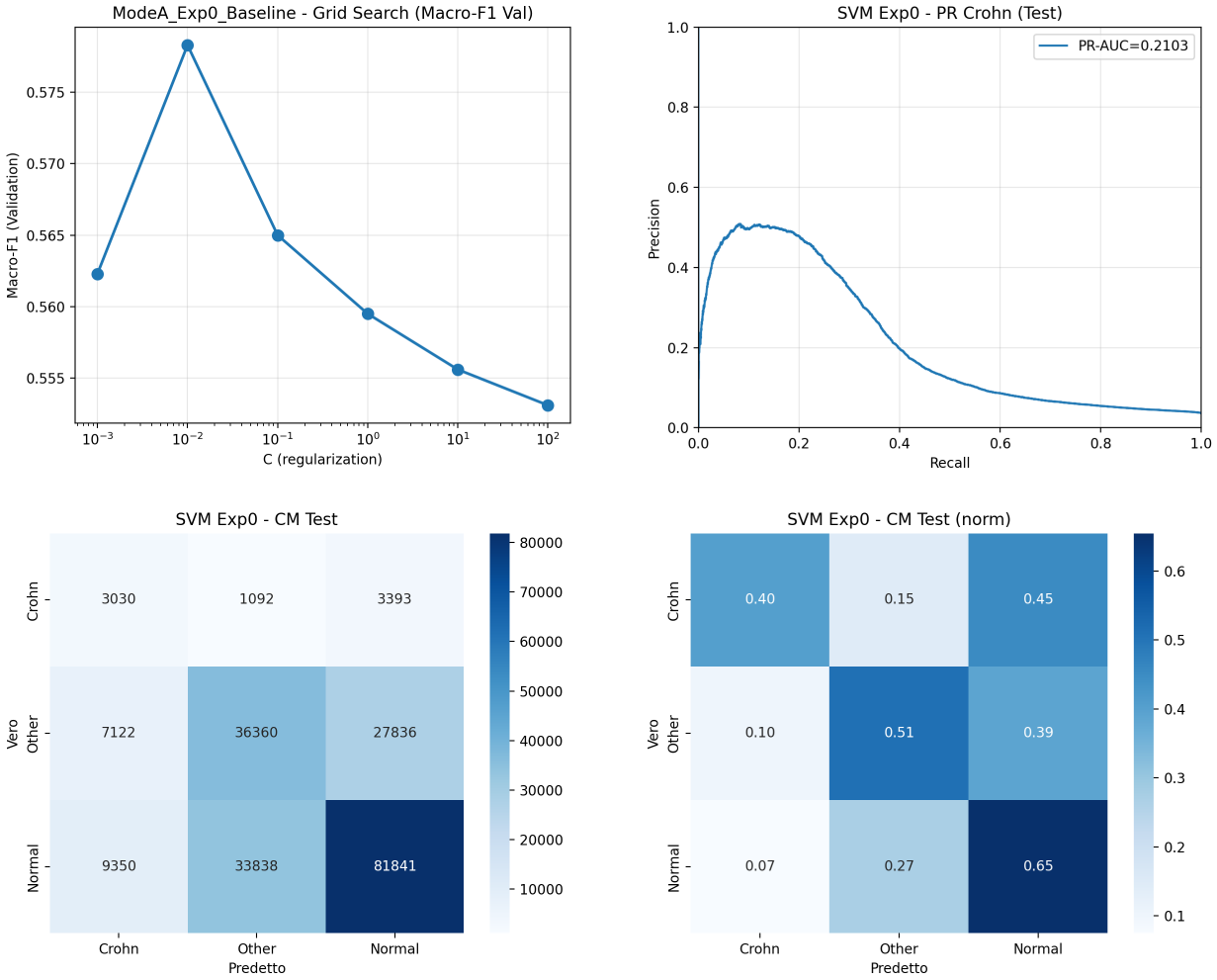


Figure 25: Linear SVM baseline (no sample weighting). Top left: grid search over C with Macro-F1 on the validation set. Top right: Precision–Recall curve for the Crohn class on the test set. Bottom left: confusion matrix (absolute counts). Bottom right: confusion matrix (row-normalized).

To test whether the recall–precision trade-off could be improved by better post-processing, I applied Crohn-first threshold tuning at r_{95} across all classifiers and all scoring modes (raw decision scores, calibrated probabilities, prior-corrected probabilities):

Classifier	Scoring	Recall_C	Precision_C	Macro-F1
Linear SVM	Decision scores	0.954	0.040	0.110
Linear SVM	Calibrated	0.942	0.041	0.109
Linear SVM	Prior-corrected	0.951	0.041	0.109
LogReg	Probabilities	0.950	0.044	0.164
LogReg	Prior-corrected	0.976	0.043	0.136
Ridge	Decision scores	0.966	0.043	0.137
Ridge	Calibrated	0.947	0.045	0.148
Ridge	Prior-corrected	0.947	0.044	0.121

At the r_{95} operating point, Crohn precision fell to 4.0–4.5% across all eight configurations. Neither calibration nor Bayesian prior correction provided any meaningful improvement: the precision values were within 0.5 percentage points of each other. The invariance of this result across three loss functions (hinge, cross-entropy, least-squares) and three scoring methods (raw, calibrated, prior-corrected) demonstrated that the trade-off was not an artifact of a particular model or post-processing choice but a structural property of linear decision boundaries in the DINOv2 embedding space under the given class prevalence. Notably, the prior correction with asymmetric target priors caused a near-complete collapse of the Other class: the scaling factor for Normal ($\sim$2.4$\times$) dominated that for Other ($\sim$0.5$\times$), driving Other recall toward zero and effectively reducing the three-class problem to a de facto binary Crohn-vs-Normal discrimination. This showed that prior correction was not merely ineffective but actively harmful in this three-class setting, as it eliminated the intermediate class rather than improving minority-class discrimination.

I additionally tested offline data augmentation (horizontal flip and rotation, applied at the image level before encoding) with Logistic Regression. Augmenting each training sample 2$\times$ or 3$\times$ improved Crohn PR-AUC from 0.177 to 0.276 (a 56% relative gain) and increased Crohn recall under argmax from 0.392 to 0.563. Between the two augmentation levels, 2$\times$ represented the recommended compromise: the more aggressive 3$\times$ augmentation shifted the operating thresholds, and Crohn recall at the r_{95} point dropped from 0.976 to 0.929, indicating diminishing returns and increased sensitivity to threshold selection. However, the threshold-tuning trade-off remained unchanged: at r_{95}, precision was still approximately 4.5%, confirming that augmentation improved ranking quality but did not resolve the structural precision collapse at high recall. At the patient level, sensitivity remained 1.0 across all augmentation variants—every Crohn patient was correctly identified as positive at r_{95}—confirming that frame-level metrics were the discriminating dimension while patient-level detection was already saturated for linear models. Figure 26 presents the corresponding diagnostics for the patient-weighted SVM, which achieved the best Macro-F1 among all linear configurations.

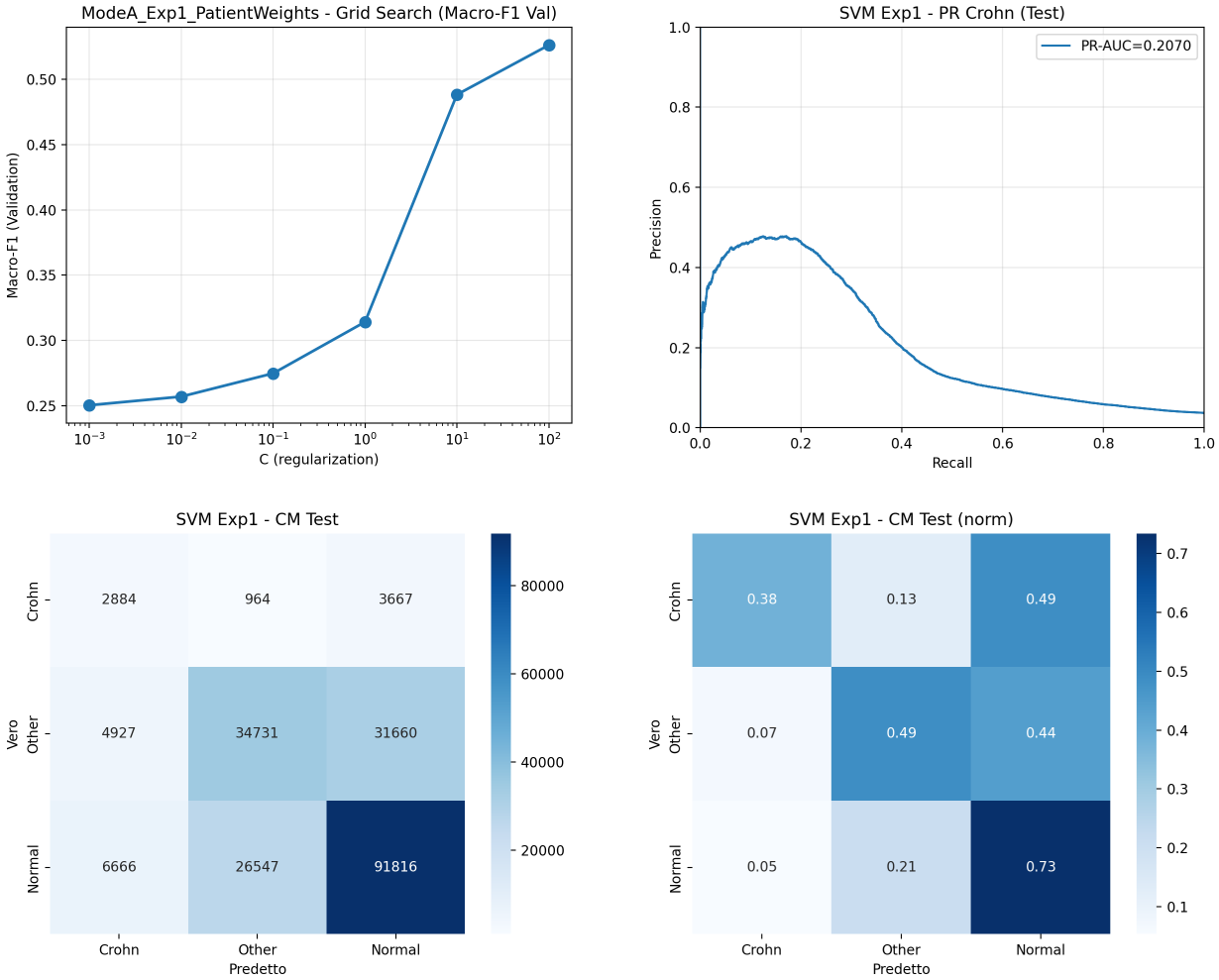


Figure 26: Linear SVM with patient-balanced weighting (best linear model, Macro-F1 = 0.504). Top left: grid search over C with Macro-F1 on the validation set. Top right: Precision–Recall curve for the Crohn class on the test set. Bottom left: confusion matrix (absolute counts). Bottom right: confusion matrix (row-normalized).

8.4 Structural Limits of Linear Classifiers

The linear classifier experiments established that frozen DINOv2 embeddings could achieve moderate frame-level screening performance (Macro-F1 $\sim$0.50, Crohn PR-AUC $\sim$0.21 for the best model), but encountered a persistent recall–precision trade-off for the Crohn class. Pushing Crohn recall above 90% invariably caused precision to collapse to approximately 4–5%, regardless of the classifier family, the loss function, the weighting scheme, the calibration strategy, or the prior correction applied. This invariance across all linear approaches strongly suggested that the DINOv2 embedding space was not linearly separable for the Crohn class with sufficient margin: Crohn frames were interleaved with Other and Normal frames in regions of the embedding space that no hyperplane could separate with sufficient margin. Among the three loss functions, the hinge loss (Linear SVM) achieved the highest Crohn PR-AUC (0.21), followed by cross-entropy (Logistic Regression, 0.16) and least-squares (Ridge, 0.13). This ordering was consistent with the loss geometry: the hinge loss focuses on samples near the decision boundary, producing better ranking in the critical region; cross-entropy penalizes errors proportionally to model confidence; and least-squares treats all residuals uniformly, making it least suited to discriminating rare classes where the score distribution matters most. The implication was that overcoming this barrier required introducing non-linearity in the classification head, which is the subject of the next chapter.

9 Experiments with Non-linear Classifiers

The objective of this chapter was to test whether the introduction of controlled non-linearity could overcome the recall–precision barrier identified in the linear classifier experiments, while preserving the frozen-encoder paradigm. I replaced the linear classification head with a shallow multi-layer perceptron (MLP), maintaining the same dataset, splits, evaluation protocol, and Crohn-first decision rule used in the linear classifier experiments to enable direct comparison. The working hypothesis was that a non-linear head could exploit discriminative patterns in the embedding space that linear classifiers could not access, improving Crohn separation without requiring fine-tuning of the encoder.

9.1 MLP Architecture Design

I evaluated two MLP configurations of increasing depth, both operating on the 768-dimensional DINOv2 ViT-B/14 embeddings:

Component	MLP-1	MLP-2
Input normalization	LayerNorm(768)	LayerNorm(768)
Hidden layers	$768 \to 512$	$768 \to 512 \to 128$
Activation	GELU	GELU
Dropout	0.3	0.3 (each layer)
Output	$512 \to 3$	$128 \to 3$
Parameters	$\sim$393,000	$\sim$459,000

Each design choice was motivated by established practice. LayerNorm on the input stabilized training by normalizing the pre-extracted embeddings, reducing internal covariate shift. GELU (Gaussian Error Linear Unit) provides a smooth approximation of ReLU with superior gradient properties and is the standard non-linearity in modern transformer architectures such as ViT. Dropout at rate 0.3 provided regularization against overfitting while balancing capacity and generalization. The hidden dimension of 512 effected a moderate reduction from the 768-dimensional input, providing sufficient capacity without overparameterization. Both architectures were deliberately shallow (one or two hidden layers), as the goal was to introduce non-linearity in the simplest possible manner without requiring advanced training techniques such as skip connections or learning rate warmup.

9.2 Training Configuration

Training used cross-entropy loss with patient-wise sample weighting, identical to the linear classifier experiments. Each sample i belonging to patient p received weight $w_i = 1/n_{\text{frames}}(p)$, preventing patients with many frames from dominating the gradient and ensuring approximately equal per-patient contribution. The optimizer was AdamW with learning rate 10^{-3}, weight decay 10^{-4}, and default momentum parameters $(\beta_1, \beta_2) = (0.9, 0.999)$. Batch size was 512.

Model selection used early stopping on validation Macro-F1 with patience 8 and a maximum of 50 epochs. The best validation checkpoint was retained for subsequent evaluation. All random seeds were fixed (42) for reproducibility. Training completed in approximately 3 minutes on a consumer GPU (NVIDIA GTX 1650, 4 GB VRAM), confirming that the frozen-encoder approach kept computational cost negligible even on modest hardware.

9.3 Post-Hoc Calibration and Threshold Tuning

After training, I applied temperature scaling [37] to calibrate the predicted probabilities without modifying argmax predictions. A scalar parameter $T > 0$ was introduced such that the calibrated probabilities were:

$$p_{\text{cal}}(k \mid x) = \text{softmax}\big(\mathbf{z}(x)/T\big)_k,$$

where $\mathbf{z}(x)$ denotes the logit vector. The temperature T was optimized by minimizing cross-entropy on the validation set using L-BFGS, with bounds $T \in [0.05, 50]$ for numerical stability. A temperature $T > 1$ indicates that the model was overconfident (logits too extreme), while $T < 1$ indicates underconfidence.

The calibrated probabilities were then used for Crohn-first threshold tuning, identical to the protocol described in the previous chapter. The decision rule assigned class Crohn whenever $p_{\text{cal}}(\text{Crohn} \mid x) \geq \tau$; otherwise the prediction defaulted to the argmax over Other and Normal. For each target recall level $r \in \{0.90, 0.95, 0.99\}$, I found the minimum threshold τ on the validation set that achieved Crohn recall $\geq r$, selecting the threshold with the highest precision among all valid candidates. These thresholds were then applied to the test set to measure the realized recall–precision trade-off under realistic prevalence.

9.4 Results: Baseline MLP

Figure 27 shows the training dynamics for both MLP architectures. Both architectures converged at epoch 6 for their best validation Macro-F1, with early stopping triggered at epoch 14:

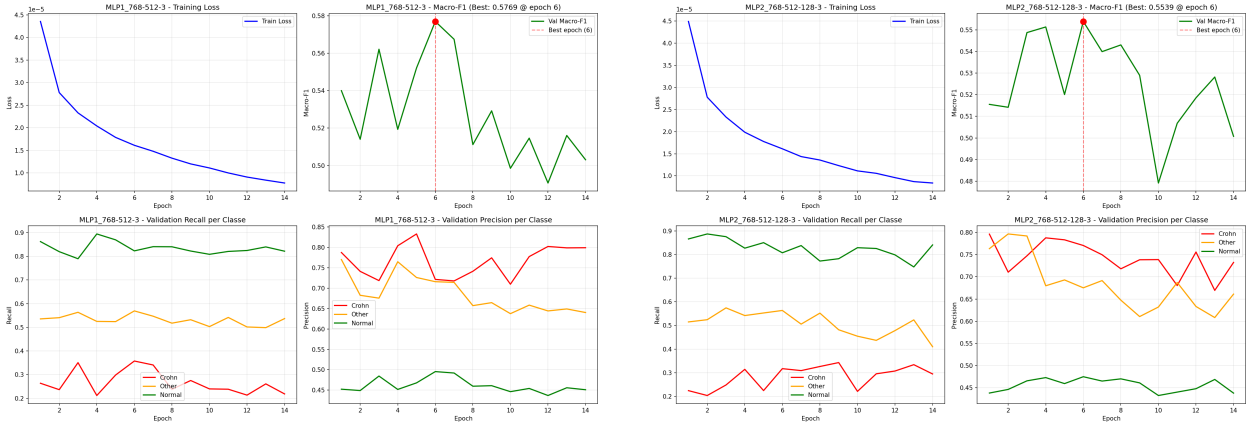


Figure 27: Training curves for MLP-1 (left) and MLP-2 (right). Each panel shows training loss, validation Macro-F1, and per-class recall and precision across epochs.

Model	Best epoch	Val Macro-F1	Early stop
MLP-1 (768-512-3)	6	0.5769	14
MLP-2 (768-512-128-3)	6	0.5539	14

The simpler MLP-1 outperformed MLP-2, indicating that moderate non-linearity was sufficient to exploit the discriminative structure of the embedding space. The additional capacity of MLP-2 provided no benefit and may have introduced overfitting. All subsequent results used MLP-1. The following table compares the selected MLP-1 against the linear classifiers on the test set under argmax prediction (no threshold tuning):

Model	Macro-F1	PR-AUC$_C$	Recall$_C$	Precision$_C$
LogReg (linear)	0.47	0.16	0.28	0.18
Linear SVM (linear)	0.50	0.21	0.38	0.20
Ridge (linear)	0.48	0.13	0.28	0.19
MLP-1 (non-linear)	0.53	0.27	0.44	0.25

Relative to the best linear model (Linear SVM), the MLP achieved improvements of +6% in Macro-F1, +29% in Crohn PR-AUC, +16% in Crohn recall, and +25% in Crohn precision. Notably, both recall and precision for the Crohn class increased simultaneously (recall: $0.38 \rightarrow 0.44$, precision: $0.20 \rightarrow 0.25$). This joint improvement was unattainable with linear classifiers, which invariably exhibited a trade-off between the two metrics. The substantial PR-AUC gain confirmed that the MLP achieved a better ranking of Crohn frames relative to non-Crohn frames across all operating points, not merely at a single threshold. The optimal temperature was $T = 4.27$, indicating significant overconfidence in the raw logits: the network produced probability estimates that were far more extreme than warranted by

its actual discriminative ability. As expected, temperature scaling did not alter argmax predictions but produced calibrated probabilities suitable for threshold-based screening decisions. Figure 28 presents the per-class Precision–Recall curves and the test-set confusion matrix: the Crohn AP of 0.269 represented a 7.3× improvement over the prevalence baseline (0.037), while 43.9% of Crohn frames were correctly identified and the main error mode was misclassification as Normal (43.4%).

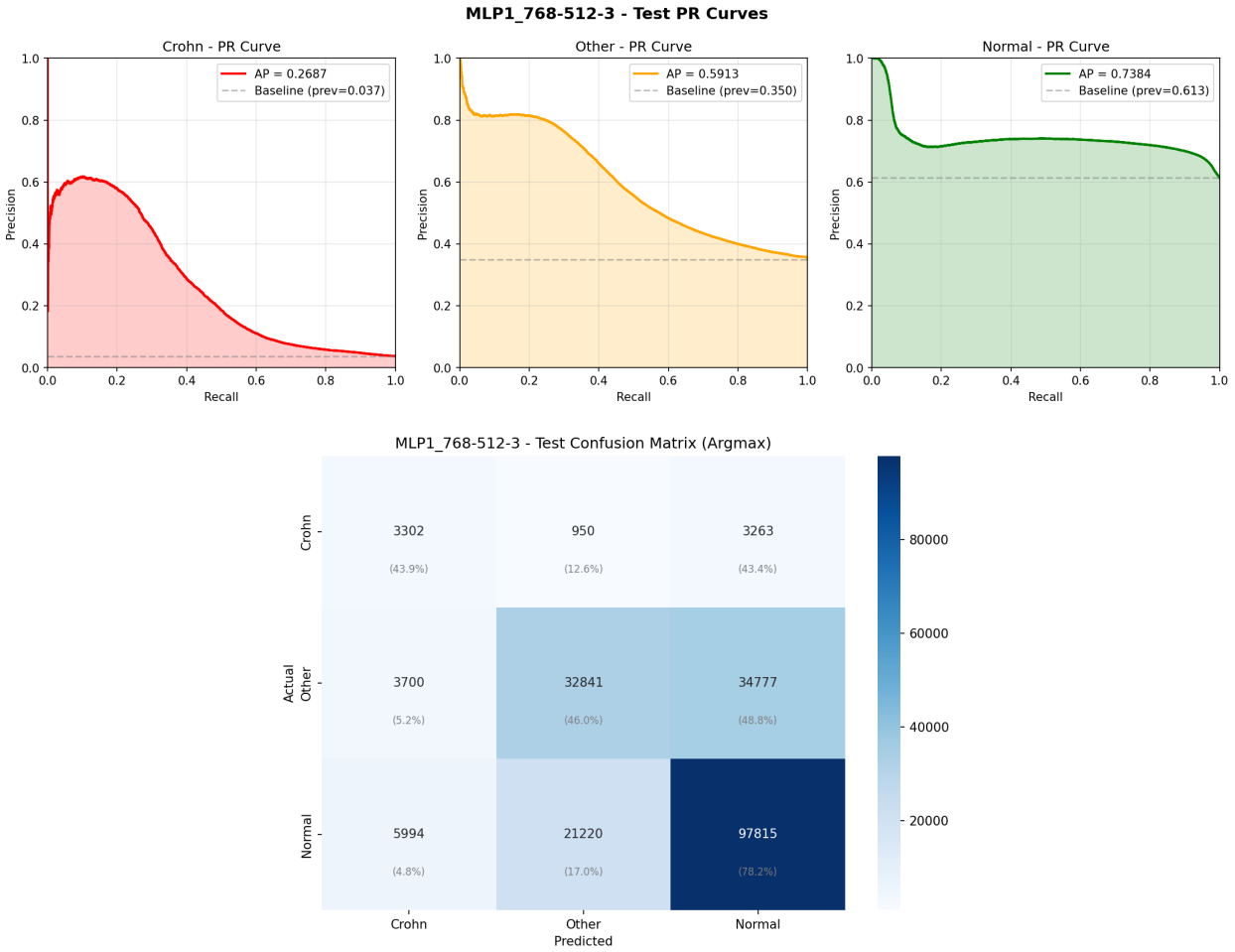


Figure 28: Top: per-class Precision–Recall curves on the test set for MLP-1, with average precision (AP) and prevalence baselines. Bottom: test-set confusion matrix under argmax prediction.

The gap between validation and test metrics was substantial and structurally driven. On the validation set, the MLP achieved $\text{Recall}_C = 0.357$, $\text{Precision}_C = 0.721$, and $\text{PR-AUC}_C = 0.657$; on the test set, these dropped to 0.439, 0.254, and 0.269, respectively. The apparent paradox—higher recall on test but much lower precision—reflected the distributional asymmetry between the two splits: the validation set was near-balanced by construction (∼33% per superclass), while the test set preserved realistic prevalence (∼3.4% Crohn, ∼29% Other, ∼67% Normal). Under this ∼9× prevalence gap, the false-positive rate that produced acceptable precision on validation translated into overwhelming false-positive contamination on

test. This distributional mismatch became the dominant factor in all subsequent threshold-based analyses.

The following table reports the Crohn-first threshold tuning results at three recall targets. Thresholds were optimized on the calibrated validation probabilities and applied to the test set.

Target	Threshold	Test Recall_{C}	Test $\text{Precision}_{\text{C}}$	Test Macro-F1	$\%\text{Pred}_{\text{C}}$
r_{90}	0.1342	0.816	0.060	0.349	50.5%
r_{95}	0.1077	0.876	0.052	0.298	62.3%
r_{99}	0.0727	0.935	0.043	0.189	80.5%

At the r_{95} operating point, the MLP predicted Crohn on approximately 62% of test frames, compared with $\sim$80% for the Linear SVM at a comparable recall target. This yielded higher precision (0.052 vs. 0.040) and substantially better Macro-F1 (0.30 vs. $\sim$0.15), because the MLP did not collapse the Other and Normal classes as aggressively when lowering the Crohn threshold. Nevertheless, the precision at high recall remained impractically low in absolute terms—approximately one true positive for every 19 false positives at r_{95}—a consequence of the low Crohn prevalence ($\sim$3.4%) in the realistic test distribution. Figure 29 visualizes the threshold analysis: the score distributions of Crohn-positive and non-Crohn frames exhibited substantial overlap, and at the r_{95} threshold 87.6% of Crohn frames were recovered at the cost of massive false-positive contamination from Other and Normal classes.

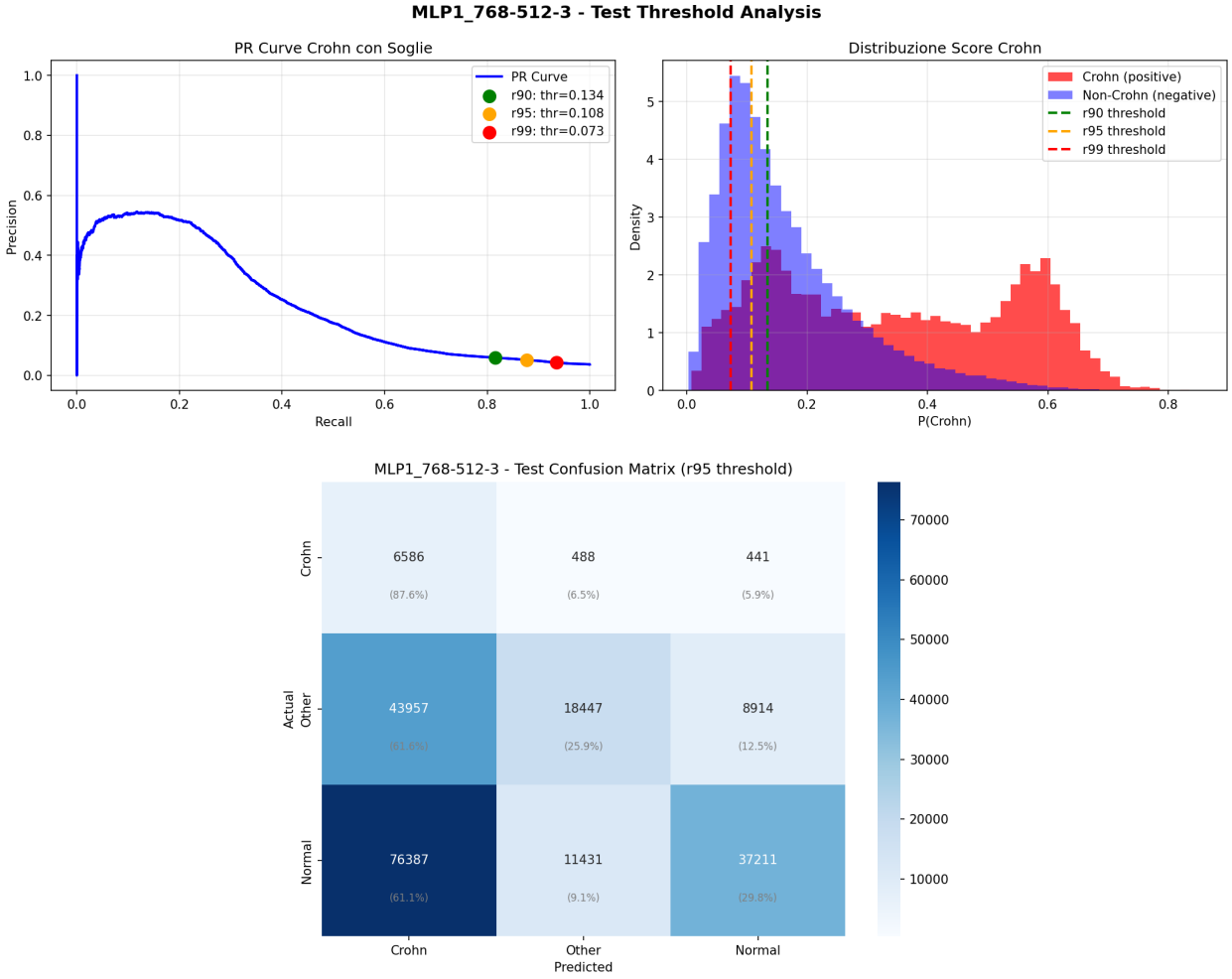


Figure 29: Top: Crohn threshold analysis on the test set. Left panel shows the Precision–Recall curve with the r_{90}, r_{95}, and r_{99} operating points. Right panel shows the score distribution of Crohn-positive and non-Crohn frames with the corresponding thresholds. Bottom: test-set confusion matrix at the r_{95} threshold.

9.5 Iterative Hyperparameter Exploration

The baseline MLP demonstrated that even a single hidden layer on frozen DINOv2 embeddings improved over all linear classifiers. A natural question was whether further gains could be obtained through more complex architectures, alternative loss functions, or better-aligned optimization objectives. I conducted a series of iterative Optuna-based campaigns, each informed by the failures and insights of the previous one. This subsection traces the experimental progression from initial search to systematic diagnostic analysis, reporting both positive and negative results to document the reasoning process that guided the investigation.

The first hyperparameter search used a composite optimization objective designed to balance overall classification quality with Crohn-specific ranking:

$$\mathcal{O} = 0.55 \cdot \text{Macro-F1} + 0.45 \cdot \text{AP}_{\text{Crohn}} - 0.10 \cdot \text{FracPred}_{\text{Crohn}},$$

where the penalty term discourages the optimizer from selecting models that trivially predict every frame as Crohn. A total of 60 trials explored architectures with 1–3 hidden layers, hidden dimensions in $[64, 1024]$, and standard regularization options. The best trial selected a single-hidden-layer network ($768 \to 256 \to 3$), achieving Val Macro-F1 of 0.585 and Test Macro-F1 of 0.520 with temperature $T = 4.86$. The improvement over the baseline was within the expected variance of the validation metric (Val $+0.008$, Test -0.010), and the smaller hidden dimension (256 vs. 512) suggested that the search had found a marginally different local optimum rather than a qualitatively better architecture. The main takeaway was that the bottleneck was unlikely to be architectural, motivating a shift toward alternative objectives and evaluation strategies.

Wide Search with Prior Weighting

This campaign expanded the search space substantially relative to the initial search and introduced a new evaluation strategy. The search space included 1–5 hidden layers, four architectural patterns (pyramid, reverse pyramid, bottleneck, uniform), five activation functions (ReLU, GELU, LeakyReLU, SiLU, Mish), Focal loss with tunable $\gamma \in [1, 5]$, label smoothing $\in [0, 0.2]$, three class weighting modes (none, inverse frequency, square-root inverse), learning rate schedulers (plateau, cosine), and optional residual connections. A total of 80 trials were evaluated using TPE sampling, of which 51 (64%) were pruned early by a median pruner.

The principal innovation of this campaign was prior weighting: during validation evaluation, sample weights were applied to approximate the realistic test prevalence on the balanced validation set, using target priors $(0.037, 0.350, 0.613)$ estimated from the GALAR source distribution for the three classes. The optimization objective was Crohn precision at recall ≥ 0.90 on this reweighted validation (PAR@0.9). The rationale was that the balanced validation set overrepresented Crohn by a factor of $\sim 9\times$, and that reweighting during evaluation would produce models whose threshold-based performance better predicts test-set behavior.

The best trial (#69 out of 80, objective value 0.072) selected a reverse-pyramid architecture with a 64-unit bottleneck (768 → 64 → 1280 → 1024 → 1024 → 3, approximately 2.9M parameters), trained with Focal loss ($\gamma = 3.28$), label smoothing (0.11), square-root inverse class weighting, BatchNorm, residual connections, and a plateau learning rate scheduler. The optimal temperature was $T = 1.47$, substantially lower than the baseline's $T = 4.27$, indicating that the Focal loss already produced less overconfident logits. Notably, the model peaked at epoch 1 and was stopped at epoch 9, suggesting immediate overfitting. As Figure 30 shows, the best validation Macro-F1 (0.555) occurred at epoch 1 and all subsequent training degraded the model; the Crohn recall remained flat at ∼0.30 throughout and the Macro-F1 oscillated erratically without recovery, confirming that the architecture was degenerate rather than undertrained.

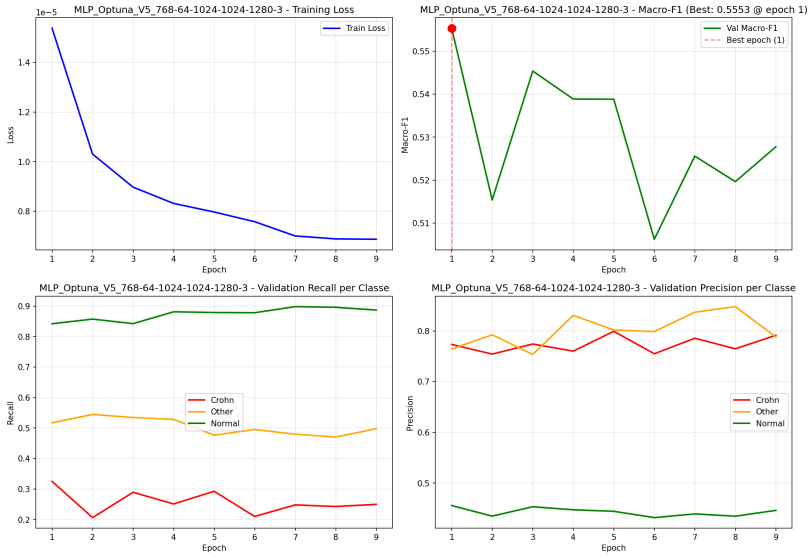


Figure 30: Training curves for the wide-search architecture (768 → 64 → 1280 → 1024 → 1024 → 3). The best validation Macro-F1 occurs at epoch 1; all subsequent training degrades the model.

Despite the elaborate design, the wide-search architecture performed worse than both the baseline MLP and the initial-search model on every test-set metric:

Model	Macro-F1	PR-AUC$_C$	Recall$_C$	Precision$_C$
Baseline MLP	0.531	0.269	0.439	0.254
Initial search	0.520	0.242	0.381	0.296
Wide search	0.509	0.170	0.401	0.216

The Crohn PR-AUC dropped by 37% relative to the baseline, from 0.269 to 0.170. The validation-to-test PR-AUC gap was severe: 0.656 → 0.170, a 74% collapse. At the r_{95} threshold, test precision fell from 0.052 (baseline) to 0.044, while the fraction of frames

predicted as Crohn increased from ∼62% to ∼77%. Three factors explain this failure, each offering a lesson that informed subsequent experiments.

First, the extreme bottleneck (768 → 64) discarded discriminative information that subsequent expansion layers could not recover. The DINOv2 embedding encodes fine-grained visual features across all 768 dimensions, and compressing them by a factor of 12 irreversibly destroyed class-relevant structure. Second, prior weighting simulated class prevalence but not the conditional distribution $P(\mathbf{x} \mid \text{class})$, which differed between the balanced validation and realistic test sets; the model therefore optimized for a simulated distribution that did not match the true test domain. Third, the enlarged search space—Focal loss, class weighting, residual connections, scheduler, five activations—introduced degrees of freedom that the TPE sampler exploited to overfit to validation artifacts rather than discover genuinely better solutions. This result demonstrated that expanding the search space without constraining it appropriately led to worse results, not better ones.

Standardized Multi-Objective Campaign

The failure of the wide search motivated a fundamental redesign. Rather than searching for a single best model, this campaign was designed as a diagnostic tool to answer a specific question: was the performance limit architectural (solvable with better hyperparameters) or structural (inherent to the data and formulation)?

The search space was deliberately constrained relative to the wide search. Two architecture presets were defined:

Parameter	Robust	Capacity
Max hidden layers	4	6
Hidden dimensions	$[64, 1024]$	$[64, 2048]$
Max parameters	∼6M	∼16M
Normalization	none, LayerNorm	none, LayerNorm, BatchNorm
Optimizer	AdamW only	AdamW, SGD

The Robust preset was designed to resist overfitting on the small balanced validation set by limiting capacity and excluding BatchNorm (which can memorize batch statistics on small datasets). The Capacity preset tested whether increased model capacity could help when combined with additional regularization options. Both presets retained patient-wise sample weighting and Crohn-first decision logic.

Nine studies were executed, crossing the two presets with four objective families. Each objective captured a different facet of the clinical screening problem:

- **$\mathbf{AP_{Crohn}}$** (ranking): $\max\big(\mathrm{AP}_{\mathrm{Crohn}} + 0.05 \cdot \text{Macro-F1}_{\mathrm{argmax}}\big)$. Tests whether the embeddings support a useful ranking of Crohn frames independently of threshold choice.
- **PAR@recall** (precision at recall target): given a recall floor $r \in \{0.90, 0.95\}$, maximize precision subject to $\mathrm{Recall}_C \geq r$, with a workload regularizer to penalize solutions that classify nearly all frames as Crohn.
- **min_workload@recall**: given $r \geq 0.90$, minimize the fraction of frames flagged for review. This directly operationalizes the clinical constraint: achieve a sensitivity floor with the smallest possible workload.
- **max_recall@workload**: given a workload budget $w \in \{0.05, 0.10, 0.20\}$, maximize Crohn recall. This inverts the previous formulation: fix the operational cost and maximize diagnostic yield.

Each study ran 15 trials with TPE sampling and median pruning, for a total of 135 trials across the campaign. All studies used prior-weighted evaluation with target priors $(0.037, 0.350, 0.613)$.

Study	Preset	Objective	Target
robust__ap	Robust	$\mathrm{AP}_{\mathrm{Crohn}}$ (ranking)	—
robust__par_r90	Robust	Precision at Recall	$r \geq 0.90$
robust__par_r95	Robust	Precision at Recall	$r \geq 0.95$
robust__minwl_r90	Robust	Min workload at Recall	$r \geq 0.90$
robust__wl_w05	Robust	Max Recall at workload	$w = 0.05$
robust__wl_w10	Robust	Max Recall at workload	$w = 0.10$
robust__wl_w20	Robust	Max Recall at workload	$w = 0.20$
cap__ap	Capacity	$\mathrm{AP}_{\mathrm{Crohn}}$ (ranking)	—
cap__par_r90	Capacity	Precision at Recall	$r \geq 0.90$

A key methodological innovation introduced in this campaign was aligned retraining. In previous campaigns, Optuna selected the best hyperparameters by optimizing a specific objective (e.g., "maximize Recall at workload $\leq 10\%$"), but the final model was retrained with early stopping on Macro-F1—a different criterion. This mismatch meant the retrained model might stop at a different epoch than the one Optuna evaluated, distorting the conclusions. In the initial aligned retraining procedure, the early stopping criterion during retraining was set to match the Optuna objective exactly, ensuring that the final model was optimized for the same criterion that drove architecture selection.

The first aligned retrain on the best architecture from this campaign ($768 \to 128 \to 576 \to 960 \to 3$) produced a revealing result:

Split	Recall_{C}@w=0.10	$\text{Precision}_{\text{C}}$	Realized workload
Validation	0.495	0.183	10.0% (by construction)
Test	0.540	0.109	18.2%

Figure 31 shows the training dynamics for this architecture: the validation $\text{Recall}_{\text{Crohn}}$ at 10% workload plateaued at approximately 0.49 after epoch 2 and never exceeded 0.50, despite continued decrease in training loss. This ceiling persisted across all nine campaign studies and was consistent with a structural performance limit.

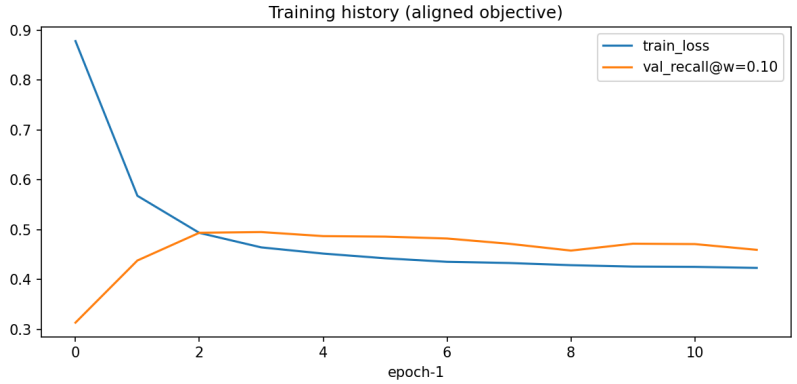


Figure 31: Training curve for the aligned retraining of the best architecture from the multi-objective campaign (768 → 128 → 576 → 960 → 3). The validation $\text{Recall}_{\text{Crohn}}$ at 10% workload plateaus after epoch 2 despite continued decrease in training loss.

The threshold tuned to achieve exactly 10% workload on validation produced 18.2% workload on test—nearly double the target. This workload drift was the first quantitative evidence that the validation-to-test transfer problem was not merely a metric-level issue (PR-AUC dropping) but an operational one: a threshold calibrated for clinical deployment would systematically overshoot the intended workload budget on unseen patients.

9.6 Systematic Retraining and Procedural Alignment

During the analysis of the campaign results, I discovered a procedural inconsistency in the initial aligned retraining pipeline: the weighted loss normalization differed between search and retrain phases ($\sum w_i \cdot \ell_i$ vs. $\sum w_i \cdot \ell_i / \sum w_i$), and the workload threshold computation used inconsistent tie-handling. While the conceptual alignment was correct, these implementation mismatches could cause the retrain to select a different epoch than intended, potentially distorting the conclusions about workload drift.

I corrected these issues in a revised version of the pipeline: (1) the weighted loss was standardized to $\sum(w_i \cdot \ell_i)/\sum(w_i)$, (2) the workload threshold computation was made consistent

between search and retrain, and (3) checkpoint saving was aligned with the selection criterion. All nine campaign studies were retrained with the corrected pipeline.

The retrained models confirmed the workload drift pattern observed with the initial pipeline. For the representative study robust_wl_w10 (architecture $768 \to 576 \to 384 \to 1088 \to 3$, Focal loss with $\gamma = 4.57$, LeakyReLU, LayerNorm, residual connections):

Split	Recall_C@w=0.10	Precision_C	AP_C	Realized workload
Validation	0.488	0.181	0.190	10.0%
Test	0.584	0.118	0.145	16.9%

The correction did not eliminate the drift: the test workload of 16.9% still substantially exceeds the validation target of 10%. This result was important because it ruled out an implementation error as the source of the drift: the discrepancy was not caused by a pipeline defect but by a genuine distributional difference between the validation and test patient populations. The validation set (6 patients, balanced prevalence ∼33% Crohn) and the test set (80 patients, realistic prevalence ∼3.4%) differed in both patient-level characteristics and class distributions, and no threshold calibrated on the former could reliably transfer to the latter. Figure 32 visualizes this workload drift: the threshold calibrated to select exactly 10% of frames on the validation set results in 16.9% workload on the test set, a direct consequence of the distributional mismatch between the balanced validation and realistic-prevalence test sets.

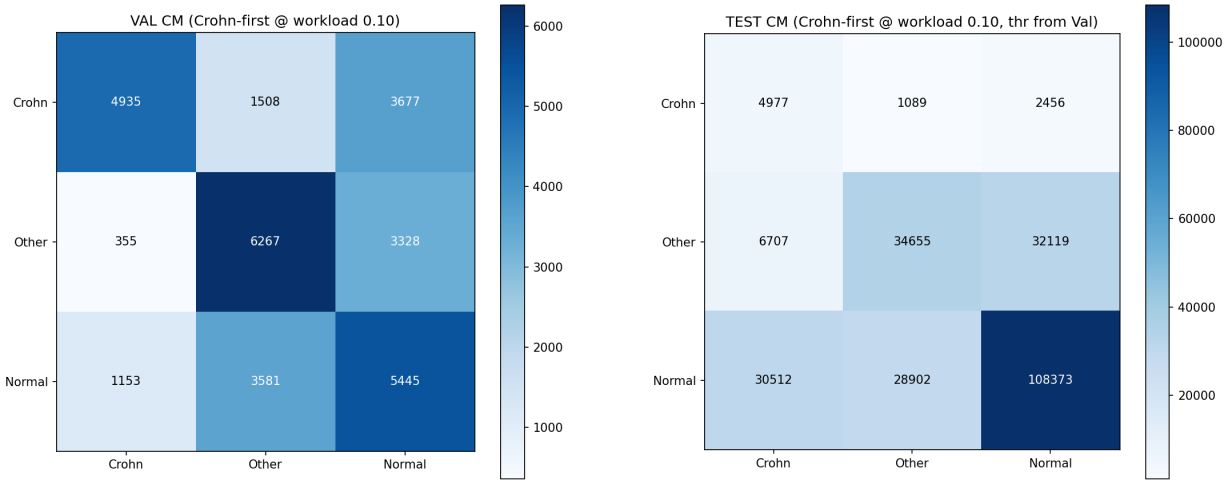


Figure 32: Confusion matrices at the 10% workload operating point for the aligned retrained model. Left: validation set, where the threshold is calibrated to select exactly 10% of frames. Right: test set, where the same threshold results in 16.9% workload.

Across all nine retrained studies, test AP_{Crohn} ranged from 0.105 to 0.190, consistently below the baseline MLP's 0.269. The Robust and Capacity presets produced similar test-set performance, and no objective family (ranking, precision-at-recall, workload-bounded) achieved

a clear advantage over the others. These convergent results supported the conclusion that the performance ceiling was structural rather than architectural: regardless of how the optimization problem was formulated, the test-set outcome remained bounded by the same factors.

9.7 Cross-Validation Analysis

The recurring workload drift and the small size of the validation set (6 patients) raised a fundamental question: was any conclusion drawn from a single hold-out split trustworthy? To quantify the validation instability, I conducted a 10-fold patient-grouped cross-validation over the Train+Validation pool (19 patients total, split as 15 train and 4–5 validation per fold). The architecture was fixed to a single hidden layer ($768 \to 512 \to 3$) with GELU, LayerNorm, dropout 0.2, and Focal loss ($\gamma = 2.0$) with square-root inverse class weighting. The primary metric was $\text{Recall}_{\text{Crohn}}$ at 10% workload with prior-weighted evaluation.

Fold	n_{train}	n_{val}	Best epoch	Recall@w=0.10	Threshold
1	72,303	49,101	1	0.497	0.490
2	102,381	19,023	6	0.421	0.138
3	85,028	36,376	8	0.325	0.618
4	98,918	22,486	2	0.725	0.206
5	90,131	31,273	10	0.424	0.585
6	81,005	40,399	4	0.428	0.531
7	103,509	17,895	1	0.583	0.279
8	87,553	33,851	4	0.529	0.625
9	106,682	14,722	1	0.486	0.136
10	76,930	44,474	1	0.166	0.648
Mean ± std			3 (median)	0.458 ± 0.149	—

Three patterns emerged from this table. First, the inter-fold variance was extreme: Fold 4 achieved Recall@w=0.10 of 0.725 while Fold 10 achieved only 0.166, a 4.4× difference driven entirely by which patients happened to fall in the validation set. The standard deviation of 0.149 on the primary metric was large enough to render any single-split ranking of models or hyperparameters unreliable. Second, the optimal epoch varied widely (1 to 10, median 3), indicating that the model overfit rapidly and that the stopping point depended more on the validation composition than on the training dynamics. Third, the threshold required to achieve 10% workload on each fold's validation set ranged from 0.136 to 0.648—a nearly 5× range—demonstrating that the threshold was not a transferable quantity but a fold-specific artifact of the patient composition.

To mitigate the instability of per-fold thresholds, I aggregated out-of-fold (OOF) predictions: each frame in the Train+Validation pool received a prediction from the fold in which it served

as validation data, and the workload threshold was computed on these pooled predictions. This OOF strategy produced a single threshold ($\tau = 0.509$ in the prior-weighted space) estimated from all 19 patients rather than from 4–5.

Applying this OOF threshold to the held-out test set yielded:

Evaluation	Precision_C	Recall_C	Realized workload
OOF (Train+Val pool)	0.119	0.321	10.0% (by construction)
Test (OOF threshold)	0.212	0.402	7.0%

These metrics were computed in the prior-weighted space; in raw (unweighted) space, the OOF pool yielded $\text{Precision}_C = 0.666$ and $\text{Recall}_C = 0.206$, reflecting the balanced class distribution of the training pool versus the realistic test prevalence. The OOF threshold achieved 7.0% workload on test rather than the target 10%, a drift in the opposite direction compared to the single-split experiments (which drifted upward to ∼17–18%). This asymmetry arose because the OOF threshold was estimated on a balanced patient pool (where Crohn prevalence was ∼33%), while the test set had realistic prevalence (∼3.4%); the threshold that captured 10% of frames in the high-prevalence pool was too conservative for the low-prevalence test domain. The retrained model on the full Train+Validation pool achieved Test Macro-F1 of 0.505 and $\text{Recall}_{\text{Crohn}}$ of 0.490 under argmax, consistent with previous results and confirming that the architecture was not the limiting factor.

The cross-validation analysis revealed three structural problems that could not be solved by hyperparameter optimization alone:

1. The training pool contained only 13–15 patients per fold, of which few carried Crohn pathology. Patient-level diversity was insufficient to learn decision boundaries that generalized across the full patient population.

2. Threshold calibration was inherently fragile because the validation prevalence (∼33% Crohn in the balanced split) differed from the test prevalence (∼3.4%). No single threshold could simultaneously achieve a target workload on both distributions, and the direction of the drift (upward or downward) depended on which patients composed the validation set.

3. The frame-level formulation treated each frame independently, ignoring temporal context and patient-level structure. This prevented the model from learning that consecutive frames of the same lesion should receive correlated predictions, and forced it to rely entirely on per-frame visual features that might not be sufficiently discriminative at the individual frame level.

9.8 Summary and Identified Structural Limits

This chapter followed an iterative experimental trajectory from a baseline MLP through progressively broader hyperparameter searches to systematic diagnostic analysis. The baseline MLP ($768 \to 512 \to 3$, $\sim$393,000 parameters, trained in under 3 minutes) broke the recall–precision trade-off inherent to linear classifiers, simultaneously improving both metrics for the Crohn class: recall from 0.38 to 0.44 and precision from 0.20 to 0.25 relative to the best linear model. This result was obtained with minimal computational overhead, confirming that frozen-encoder non-linear probing was a viable and efficient approach.

The subsequent hyperparameter exploration revealed a consistent pattern: increased model complexity failed to improve test-set generalization, and in some cases actively degraded it. The initial Optuna search found a marginally different architecture with no meaningful gain. The wide search with prior weighting produced a severe PR-AUC collapse (-37%), teaching that unconstrained search spaces and prevalence simulation without conditional distribution matching led to overfitting. The multi-objective campaign, designed as a diagnostic rather than an optimization tool, confirmed that neither the architecture preset (Robust vs. Capacity) nor the objective function family (ranking, precision-at-recall, workload-bounded) meaningfully changed the test-set performance ceiling.

The aligned retraining experiments isolated the workload drift phenomenon: thresholds calibrated for 10% workload on validation consistently produced 17–18% workload on test, and this drift persisted after correcting procedural inconsistencies, confirming its structural nature. The cross-validation analysis provided the final diagnostic: with a standard deviation of $\sim$0.15 on the primary metric across folds and a $4.4\times$ range between the best and worst folds, single-split conclusions were unreliable. These findings collectively established that the performance limit was structural—driven by low Crohn prevalence, insufficient patient diversity, and the frame-level formulation—and motivated the paradigm shift to pathology-level prediction with workload-aware evaluation pursued in the next chapter.

10 Pathology-level Classification and Workload-based Triage

10.1 Motivation: From Superclasses to Pathology-Level Prediction

The preceding chapters explored frame-level classification using three aggregated superclasses: Crohn (categories 003, 004, 005), Other (categories 001, 002, 006–013), and Normal. Despite the improvement obtained by introducing non-linearity, three structural issues remained unresolved. First, the validation instability documented in the MLP experiments (standard deviation $\sim$0.15 on the primary metric under repeated patient-grouped cross-validation) rendered conclusions drawn from a single validation split unreliable. Second, thresholds calibrated on validation drifted substantially when transferred to the test set, consistent with patient-level domain shift and insufficient representativeness of the training and validation pools. Third, the superclass Other aggregated pathologies with extremely heterogeneous visual patterns—vascular lesions (angiectasia, lymphangiectasia), mucosal abnormalities (erythema, edema), structural anomalies (stenosis, polyps), and acute conditions (bleeding, hematin)—introducing high intra-class variance in the embedding space and impeding the learning of robust decision boundaries.

These observations motivated a change of approach: abandoning the aggregated superclass formulation in favor of pathology-level (microclass) prediction. The hypothesis was that classifying individual pathologies directly should reduce intra-class variance, produce more compact and separable clusters in the embedding space, and yield more clinically interpretable predictions. Pathology-level probabilities could always be aggregated post-hoc into superclass scores ($P(\text{Crohn}) = P(003) + P(004) + P(005)$) for comparison with previous steps. Additionally, I shifted the primary evaluation from threshold-dependent frame-level metrics to a workload-aware triage metric that directly modeled the clinical objective of reducing reading workload while maintaining sensitivity.

10.2 Dataset Enrichment

The validation instability documented in the MLP experiments—standard deviation of approximately 0.15 on the primary metric under repeated patient-grouped evaluation—indicated that conclusions drawn from a single validation split were unreliable. A root cause was the small number of patients contributing Crohn frames: too few patients meant that each validation fold was dominated by the idiosyncratic visual characteristics of one or two individuals, producing high variance in threshold calibration and metric estimates. To address this, I identified additional patients with a high density of Crohn lesions to increase the diversity of Crohn-positive cases available for training and validation.

The superclass experiments (Sections 8–9) used 55 GALAR patients. For the pathology-level task, two of those patients (36 and 37) were replaced by patients 23 and 64, selected from the GALAR source for their higher concentration of Crohn lesion frames; the total remained 55. Both patients were processed through the identical curation pipeline described in the preceding chapters: embedding extraction with the same frozen DINOv2 ViT-B/14 encoder, followed by the same pruning methodology—Normal frames reduced to 20% via MiniBatchKMeans clustering ($K = 600$), blood-only frames reduced to 15% ($K = 400$, minimum 50 frames preserved), and all Crohn and Other pathology frames preserved in their entirety. The patient-wise splitting optimization was then recomputed with the same formulation and solver to incorporate the new patients into the existing split structure.

	Normal		Blood-only		Crohn/Other	
Patient	Before	After	Before	After	(preserved)	Total after
23	10,912	2,182	422	63	873	3,118
64	8,504	1,701	387	58	701	2,460

After enrichment, the dataset contained 55 patients from the GALAR source with complete embeddings, all pruned and curated following a single consistent methodology. The enriched dataset was characterized at the pathology level in Figures 33–36, providing the distributional context for the classification experiments that followed. Figure 33 shows that erosion accounted for 76.1% of all Crohn frames, making it the dominant lesion type. Figure 34 breaks down ulceration into superficial and deep categories and reports the number of patients contributing to each class. Figure 35 presents the Crohn lesion distribution across splits and dataset sources, confirming that GALAR contributed the large majority of both erosion and ulceration frames. Figure 36 highlights the concentration of Crohn frames in a small number of patients, a structural property that limits the effective diversity of the training pool.

10.3 Triage Metric: Recall@10% Workload per Patient

In the clinical workflow, a capsule endoscopy examination generates thousands of frames per patient. The practical value of an AI triage system lies not only in frame-level accuracy but in its ability to reduce the number of frames the clinician must review while maintaining the ability to detect lesions. This workload-reduction paradigm is well-established in AI-assisted capsule endoscopy reading, where deep learning systems are used to filter normal frames and prioritize abnormal ones for clinical review [34], [35]. I therefore adopted a workload-based triage metric as the primary evaluation criterion for pathology-level classification.

For each patient p with N_p frames, the model produces per-frame probabilities. A Crohn-positivity score is computed as $s_i = \hat{P}_i(\text{Ulcer}) + \hat{P}_i(\text{Erosion})$, and the top-$k_p$ most suspicious

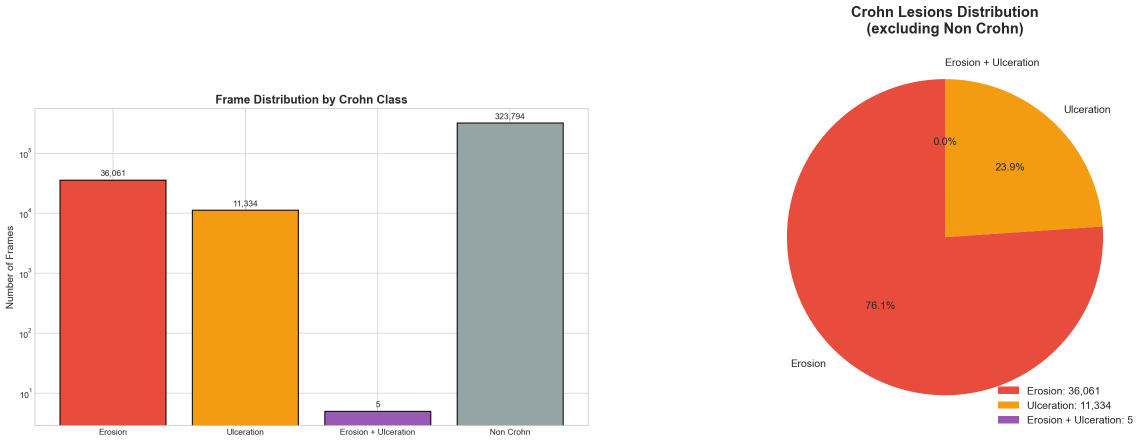


Figure 33: Left: frame distribution across Crohn lesion classes (Erosion, Ulceration, co-occurring) and non-Crohn frames, on a logarithmic scale. Right: relative composition of Crohn lesions only, showing that erosion accounts for 76.1% of all Crohn frames.

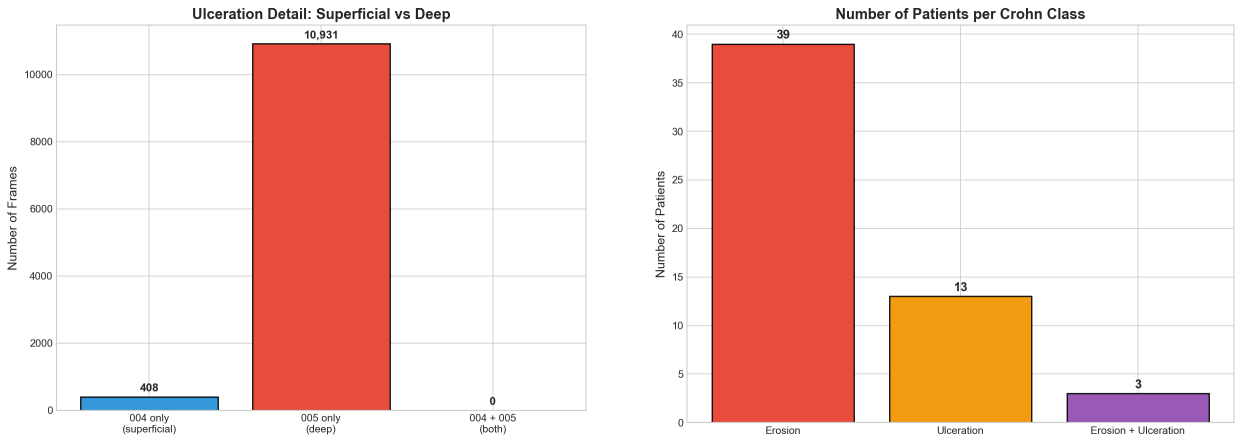


Figure 34: Left: breakdown of ulceration frames into superficial (category 004, 408 frames) and deep (category 005, 10,931 frames). Right: number of patients contributing frames to each Crohn lesion class.

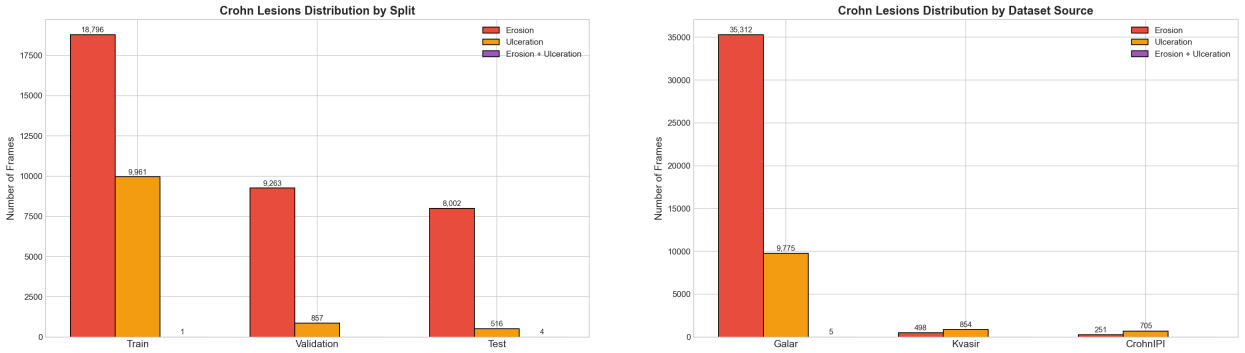


Figure 35: Left: Crohn lesion distribution across Train, Validation, and Test splits. Right: Crohn lesion distribution by dataset source (GALAR, Kvasir, CrohnIPI), showing that GALAR contributes the large majority of both erosion and ulceration frames.

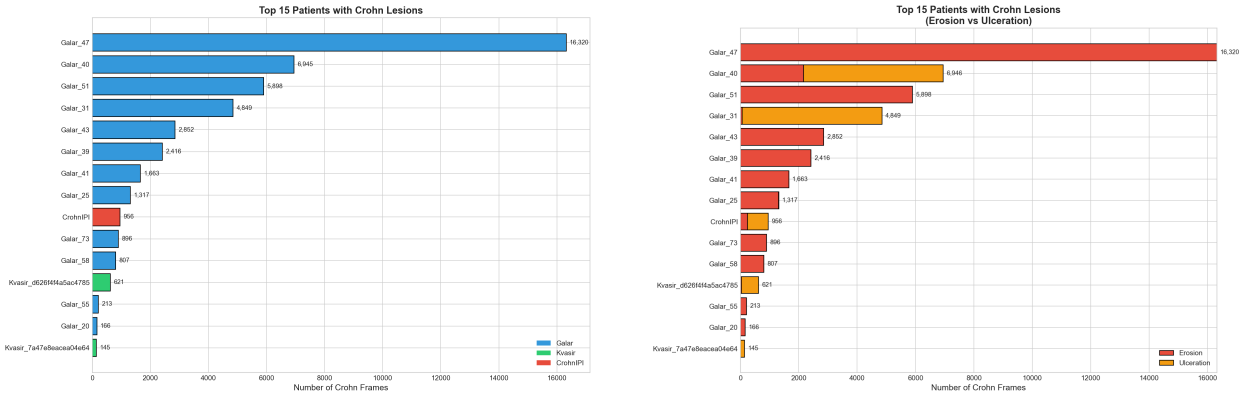


Figure 36: Top 15 patients by number of Crohn frames. Left: total count colored by dataset source. Right: stacked breakdown into erosion and ulceration, highlighting the concentration of Crohn frames in a small number of patients.

frames are selected, where $k_p = \lceil 0.10 \cdot N_p \rceil$ (10% of frames). The per-patient recall is:

$$\text{Recall}_p = \frac{\text{TP}_p}{\text{Pos}_p},$$

where TP_p is the number of truly positive frames (Ulcer or Erosion) within the top-k_p and Pos_p is the total number of positive frames for patient p. The denominator includes all positives in the patient, not only those within the selected subset, ensuring that the metric correctly penalizes false negatives left outside the top 10%.

Two aggregation modes were used: macro-patient (mean of Recall_p over positive patients) and micro-patient ($\sum \text{TP}_p / \sum \text{Pos}_p$). At 10% workload, a random baseline would capture approximately 10% of lesions; a Recall@10% of 0.35 therefore represents a 3.5× improvement over random selection. The metric cannot be inflated by predicting all frames as positive, since the budget is fixed at 10% per patient.

10.4 Crohn Three-Class Classification

Class Mapping and Dataset Structure

I defined three classes for Crohn-focused pathology-level classification:

Class	Contents	GALAR codes
Ulceration	Deep and superficial ulcers	004, 005
Erosion	Aphthoid erosions	003
Rest	Everything else (normal, other pathologies)	All other

Frames carrying both erosion and ulceration annotations (3–5 in the entire dataset) were assigned to the Ulceration class by priority; their negligible count made this choice immaterial.

The frame-level distribution was heavily imbalanced: Rest comprised 323,794 frames (87.2%), Erosion 36,061 (9.7%), and Ulceration 11,334 (3.1%). More critically, the imbalance operated at the *patient level*: ulceration was present in only 13 patients, with approximately 84% of ulcerative frames concentrated in two patients. This extreme patient-level concentration was the primary driver of the split-dependent instability observed throughout these experiments.

A new patient-workload-aware split optimizer was developed to address the instability revealed in the MLP experiments. The optimizer enforced minimum counts of ulcer-positive and erosion-positive patients in each split, prevented all dominant patients from being assigned to the same split (anti-dominance constraint), and targeted balanced pathology distributions. Patients were allocated approximately 60/20/20 across Train, Validation, and Test.

The assignment of enrichment patients (23, 64) to specific splits varied across experiments depending on the optimizer configuration.

Experimental Design and Results

Seven experiments were conducted, varying the split strategy and hyperparameter configuration (fixed defaults versus Optuna-tuned). All experiments used an MLP head on frozen DINOv2 embeddings with patient-wise sample weighting and AdamW optimization. The fixed-parameter configuration used hidden dimensions (768, 512, 256) with GELU activation, dropout 0.3, and batch size 512.

Experiment	R@10% macro	P@10% micro	PR-AUC	Macro-F1
opt_oldobj (fixed)	0.193	0.087	0.182	0.433
opt_newobj (fixed)	0.226	0.530	0.653	0.449
manual_A (fixed)	0.227	0.708	0.688	0.311
manual_B (fixed)	0.157	0.787	0.762	0.394
manual_C stress (fixed)	0.122	0.856	0.818	0.219
opt_workload_v1 (fixed)	0.315	0.400	0.495	0.374
suggested_split (Optuna)	0.172	0.538	0.604	0.389

The workload-aware optimizer (opt_workload_v1) achieved the highest Recall@10% macro (0.315) among all experiments, demonstrating that the split design was the single most consequential variable. Manual splits with very small test sets (3–4 patients) exhibited high precision and PR-AUC, but these values were unreliable given the very small evaluation pool. The Optuna-tuned experiment on the same split achieved higher global PR-AUC (0.604 vs. 0.495) but *lower* Recall@10% (0.172 vs. 0.315), revealing that Optuna optimization on a single validation split selected parameters that did not transfer reliably to the test domain.

Across all seven experiments, the variance attributable to the split assignment was substantially larger than the variance attributable to hyperparameter choice. With ulceration concentrated in very few patients, moving one or two key patients between splits radically changed all metrics. This made the split a more important methodological decision than the choice of learning rate, architecture depth, or optimization strategy—a finding with direct implications for experimental design in patient-scarce medical imaging settings.

10.5 Binary Erosion-Versus-Rest Triage

To isolate the triage signal for the most frequent Crohn lesion type, I reformulated the problem as a binary classification: Erosion (aphthoid erosions, code 003) versus Rest (everything

else, including ulcerations, other pathologies, and normal frames). This eliminated the confounding influence of the rare ulceration class and directly evaluated triage utility for the lesion type most amenable to workload reduction.

The experimental design followed a $2 \times 2 \times 2$ factorial structure:

Factor	Levels
Split	Recommended (15 test patients) vs. Optimized (17 test patients)
Hyperparameters	Fixed defaults vs. Optuna-tuned (40 trials)
Overlap	No overlap (erosion pure) vs. Overlap (erosion+ulcer as positive)

To mitigate the dominance of patients with many frames, I applied patient-aware subsampling at each epoch: a maximum of 1,024 frames per GALAR patient and 512 per Kvasir patient, with a guaranteed minimum of 32 positive frames and a target positive fraction of 25%.

Experiment	R@10%	P@10%	PR-AUC	ROC-AUC	F1@0.5
rec / fixed / no-overlap	0.272	0.523	0.503	0.784	0.215
opt / fixed / no-overlap	0.357	0.276	0.413	0.770	0.225
rec / optuna / no-overlap	0.276	0.496	0.467	0.750	0.337
opt / optuna / no-overlap	0.304	0.283	0.452	0.793	0.263
rec / fixed / overlap	0.271	0.522	0.501	0.782	0.214
opt / fixed / overlap	0.313	0.283	0.360	0.727	0.373
rec / optuna / overlap	0.276	0.495	0.467	0.750	0.338
opt / optuna / overlap	0.281	0.282	0.455	0.794	0.281

The factorial structure permitted systematic attribution of performance variation to each factor. Split (dominant factor). The optimized split outperformed the recommended split on Recall@10% in the majority of comparisons: +0.086 for fixed/no-overlap and +0.028 for Optuna/no-overlap. The optimized split contained more positive patients in the test set (12 vs. 10) and enforced anti-dominance constraints that improved representativeness. However, the recommended split preserved higher precision ($\sim$0.50 vs. $\sim$0.28) and global PR-AUC, reflecting the trade-off between sensitivity and specificity inherent to the split choice.

Hyperparameters (secondary factor). Fixed default parameters often outperformed Optuna-tuned parameters on the test set: for instance, opt/fixed/no-overlap achieved R@10% = 0.357 versus opt/optuna/no-overlap at 0.304 (-0.053). This confirmed that Optuna optimization on a single validation split tended to overfit, selecting configurations that maximized the validation metric but did not transfer to the test domain. The validation-to-test gap was smallest for the optimized split with fixed parameters ($0.385 \rightarrow 0.357$, gap = 0.028), suggesting that this combination produced the most stable training.

Overlap (negligible factor). The distinction between no-overlap and overlap modes had no meaningful impact on any metric. The co-occurrence of erosion and ulceration involved only 3–5 frames across the entire dataset, and the small differences observed were attributable to training noise. This finding simplified future experimental design: the overlap mode could be dropped as a factor.

The best model across all pathology-level experiments—opt/fixed/no-overlap, binary erosion triage—achieved a Recall@10% macro of 0.357. Reviewing only the 10% most suspicious frames per patient captured 35.7% of erosions, a 3.6× improvement over the random baseline of 10%. However, 64.3% of erosions were still missed, and only 27.6% of the frames flagged as suspicious actually contained erosions (P@10% micro = 0.276). The ROC-AUC of 0.770 indicated moderate discriminative ability, well below the >0.90 typically expected of clinically reliable diagnostic systems.

These results were not yet clinically deployable as an autonomous triage system. The performance was structurally limited by four factors. First, the DINOv2 embeddings were generic visual representations not specialized for mucosal lesion detection; the subtle visual signatures of aphthoid erosions (small, superficial, with blurred margins) are difficult to capture without domain-specific feature learning. Second, the small patient pool (erosion present in ∼39 patients, ulceration in 13) meant that each patient contributed approximately 10% to the macro-averaged metric, producing inherently high variance. Third, inter-patient heterogeneity in acquisition conditions (luminosity, bowel preparation quality, capsule transit speed) consumed part of the model's capacity for domain adaptation rather than lesion recognition. Fourth, even for expert gastroenterologists, inter-observer agreement on erosions is lower than on frank ulcerations, placing a fundamental upper bound on achievable automated performance. Despite the modest absolute performance, the experimental findings were robust and informative. The systematic factorial design confirmed that the split mattered more than hyperparameters, that Optuna on a single split was fragile, that the binary formulation outperformed three-class for erosion triage, and that the overlap distinction was irrelevant. These findings provided a rigorous baseline and clear direction for future improvement.

11 Zero-Shot Evaluation with a Medical Vision-Language Model

The preceding chapters established a pipeline in which frozen DINOv2 embeddings were classified by lightweight heads—linear classifiers (Section 8), a shallow MLP (Section 9), and pathology-level binary models (Section 10). All approaches relied on supervised training over the curated dataset. A natural question arises: *can a general-purpose medical vision-language model, applied zero-shot without any task-specific training, match or exceed these results?* If so, the added value of the supervised pipeline would need to be reconsidered; if not, the comparison quantifies the benefit of task-specific feature extraction and domain-aware dataset construction.

This chapter evaluates **MedGemma 4B** [38], a 4-billion-parameter medical vision-language model from the Gemma family, on the same three-class SBCE classification task used throughout the thesis. MedGemma was deployed zero-shot—without fine-tuning, few-shot prompting, or any exposure to the training data—to provide a relevant open medical foundation model baseline for zero-shot evaluation on this domain.

11.1 Model Selection Rationale

One of the reviewer comments on the published paper (Appendix C) suggested expanding the set of vision-language models evaluated for SBCE classification. Among the models identified as promising candidates but not yet tested was MedGemma, Google's open collection of medical vision-language models built on the Gemma 3 architecture [38]. Google positions MedGemma as its most capable open model family for health AI development, reporting that it significantly exceeds the performance of similar-sized generative models across 22 datasets spanning four medical imaging modalities—radiology, dermatology, histopathology, and ophthalmology. Independent evaluation confirms this standing: in a zero-shot comparison on six medical image classification tasks, MedGemma 4B achieved a mean accuracy of 80.37%, outperforming GPT-4 (69.58%) by over ten percentage points [39]. On the gastrointestinal subsets of the MedFrameQA multi-image clinical reasoning benchmark [40], MedGemma 27B achieved 55.56% accuracy on small intestine images—outperforming all evaluated models outside the OpenAI and Gemini families—and 37.98% on large intestine images, trailing only the OpenAI, Gemini, and Anthropic Claude families. Since all three of these proprietary families were already evaluated in the published study (Appendix C), MedGemma represented a strong open-source candidate for extending the benchmark beyond proprietary models, with demonstrated competence specifically in the gastrointestinal imaging domain. Furthermore, to the best of our knowledge, no prior work has evaluated MedGemma on small-bowel capsule endoscopy frames, making this comparison novel in addition to being directly motivated by the peer-review process.

11.2 Experimental Setup

MedGemma 4B—formally `medgemma-4b-it` (v1), hereafter referred to simply as MedGemma 4B—is an instruction-tuned multimodal model released through Google's Vertex AI Model Garden, designed for medical image understanding and clinical text generation. The model accepts an image and a text prompt, and returns a free-form text response. It was deployed as a dedicated online endpoint on Vertex AI with the following inference configuration:

Parameter	Value
Model	MedGemma 4B (multimodal, v1)
Inference mode	Zero-shot
Temperature	0.0 (deterministic)
Max output tokens	2,000
Image source	Google Cloud Storage bucket

The max output tokens parameter was set to 2,000 after preliminary tests with 500 tokens revealed a PARSE_ERROR rate of ∼8%: the model often produced verbose free-text output before the JSON answer, which increased parsing failures when the token budget was too small. Increasing to 2,000 tokens reduced the PARSE_ERROR rate to 0.6%.

Each frame was classified independently with no shared context between predictions. The prompt was designed to be directly comparable with the three-class task used in Sections 8–10.

The system prompt was: *"You are an expert gastroenterologist analyzing Small-Bowel Capsule Endoscopy (SBCE) frames."*

The user prompt, accompanied by the frame image, was:

> Classify this capsule endoscopy frame into exactly ONE of three classes:
>
> 1. "Crohn" — aphthoid erosions, superficial or deep ulcerations (mucosal breaks typical of Crohn's disease)
>
> 2. "Other" — non-Crohn pathological findings: blood, active bleeding, hematin, angiectasia, lymphangiectasia, erythematous patches, edema, stenosis, polyps, foreign bodies
>
> 3. "Normal" — healthy mucosa, anatomical landmarks (pylorus, ileocecal valve, z-line, ampulla of Vater, esophagus, stomach, small intestine, colon), or imaging artifacts (bubbles, dirt, reduced view)
>
> Do NOT explain your reasoning. Respond ONLY with a JSON object in this exact format, nothing else: {`"class": "<your_choice>"`}

Class definitions mirror the annotation guidelines used in dataset construction, ensuring semantic alignment with ground truth labels. The placeholder `<your_choice>` avoids biasing toward any specific class. The "Do NOT explain" instruction aims to minimize chain-of-thought output, though MedGemma may still produce free-text output before the JSON response.

Prompt variant tested and discarded.

In Sections 8–10, multi-label frames are resolved with a Crohn-first priority rule. To align MedGemma's behavior with this convention, a variant prompt was tested that included the instruction: *"If the frame contains findings from both 'Crohn' and 'Other' classes, classify it as 'Crohn' (Crohn takes priority)."* Preliminary evaluation on 50 frames showed that this clause drastically worsened performance (Macro-F1 dropped from 0.267 to 0.160, accuracy from 36% to 24%), as the model interpreted the priority instruction as a general bias toward predicting Crohn (52% of all predictions). A separate test reordering the class options in the prompt (Normal → Other → Crohn) showed no significant effect (Macro-F1 remained 0.267). The final prompt therefore uses the original class order without the priority clause, evaluating the model under the prompt configuration that performed best in preliminary testing.

11.3 Representative Sample Construction

Running MedGemma on all 249,790 test set frames via online endpoint would require approximately 260 hours at the observed throughput of ∼0.27 img/s. To make evaluation feasible, a representative sample of 4,000 frames was constructed using multi-dimensional stratified sampling.

The sampling algorithm was *proportional stratified sampling*, preserving the full test set's joint distribution over multiple dimensions by proportional allocation within nested strata.

The procedure was as follows:

1. **Primary stratification.** All 249,790 frames were grouped into strata defined by the Cartesian product of (superclass, patient ID, capsule system, source dataset), yielding 166 non-empty strata.
2. **Proportional target allocation.** Each stratum received a target count:

$$\text{target}_i = \text{round}\left(\frac{n_i}{249{,}790} \times 4{,}000\right),$$

with a floor of 1 to guarantee every stratum was represented. After rounding, targets were adjusted to sum exactly to 4,000.

3. **Category-pattern sub-stratification.** Within each primary stratum, frames were further grouped by their category intersection pattern—the sorted set of fine-grained annotation IDs present on the frame. Only the 27 most frequent patterns (covering 95% of the full test set) were tracked individually; rarer patterns were pooled.

4. **Random sampling.** Within each group, the allocated number of frames was drawn uniformly at random without replacement (seed = 42).

Dimension	Max deviation
Superclass	0.32%
Patient (81/81 represented)	0.07%
Source dataset	0.14%
Capsule system	0.15%
Category pattern (top 30)	0.18%

All deviations between the sample and full test set marginal distributions were below 0.35%, confirming excellent representativeness. The sample preserved the realistic test set prevalence: Normal 66.9%, Other 29.5%, Crohn 3.6%. Since MedGemma was evaluated on this representative subsample rather than on the full test set, the comparison with previously reported full-test metrics from Sections 8–9 should be interpreted as approximate. A strict apples-to-apples comparison would require re-evaluating the supervised baselines on the same 4,000-frame subset.

11.4 Results

All 4,000 frames were processed via the online endpoint, with images read directly from the Google Cloud Storage bucket. Of these, 3,976 returned valid predictions. The remaining 24 frames (0.6%) produced responses that could not be parsed into one of the three target classes (PARSE_ERROR) and were excluded from all subsequent metric computations; they did not contribute to accuracy, F1 scores, or the confusion matrix. While this exclusion does not materially affect the overall results, it represents an inherent operational limitation of the generative inference setting: unlike supervised classifiers, which always produce a valid class prediction, a VLM can fail to return a parseable answer. Total inference time was approximately 4 hours.

Overall performance.

Model	Macro-F1	PR-AUC$_C$	Recall$_C$	Precision$_C$
LogReg (linear)	0.47	0.16	0.28	0.18
Linear SVM	0.50	0.21	0.38	0.20
Ridge	0.48	0.13	0.28	0.19
MLP-1 (non-linear)	**0.53**	**0.27**	0.44	**0.25**
MedGemma 4B (zero-shot)	0.33	—	**0.52**	0.06

MedGemma achieved the highest Crohn recall (0.524) among all models evaluated in this thesis, but at the cost of clinically impractical precision (0.062). Its Macro-F1 of 0.334 was substantially below even the simplest linear classifier (LogReg, 0.47), indicating that the overall classification quality was poor.

Per-class metrics.

Class	Precision	Recall	F1-Score	Support
Crohn	0.062	0.524	0.110	145
Other	0.337	0.293	0.314	1,173
Normal	0.734	0.476	0.577	2,658

Figure 37 compares per-class Precision, Recall, and F1 between MedGemma and the MLP baseline from Section 9. The MLP outperforms MedGemma on every metric for every class, with the sole exception of Crohn Recall—where MedGemma's higher value (0.52 vs. 0.44) is offset by very low precision (0.06 vs. 0.25).

Confusion matrix.

Figure 38 shows the confusion matrix under argmax prediction, in the same format used for the MLP baseline in Section 9.

The row-normalized confusion matrix reveals three key patterns:

1. **Crohn row (52.4% / 20.0% / 27.6%).** The model correctly identifies slightly more than half of Crohn frames, but disperses the rest roughly equally between Other and Normal.
2. **Other row (35.0% / 29.3% / 35.7%).** Predictions are distributed almost uniformly

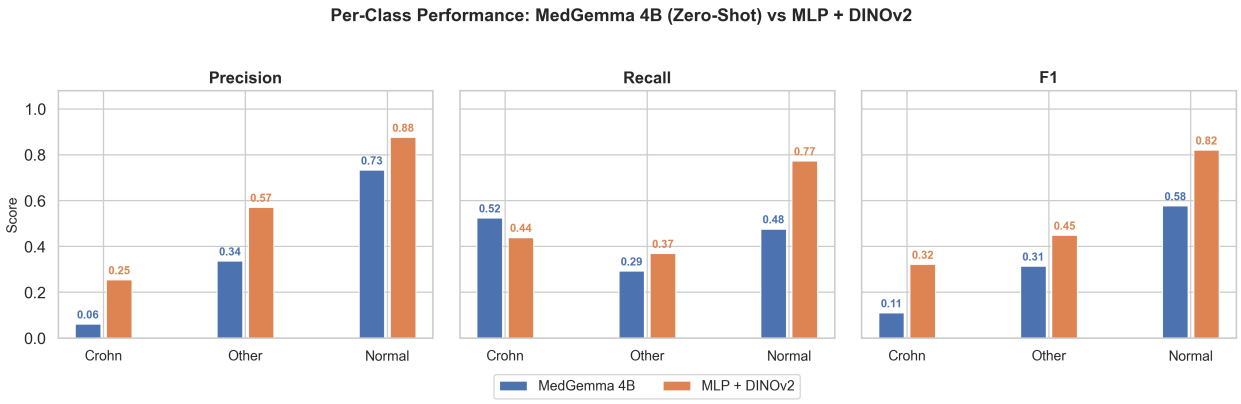


Figure 37: Per-class Precision, Recall, and F1 comparison between MedGemma 4B (zero-shot) and the MLP + DINOv2 baseline (Section 9). MedGemma exceeds the MLP only on Crohn Recall, at the cost of near-zero Crohn Precision.

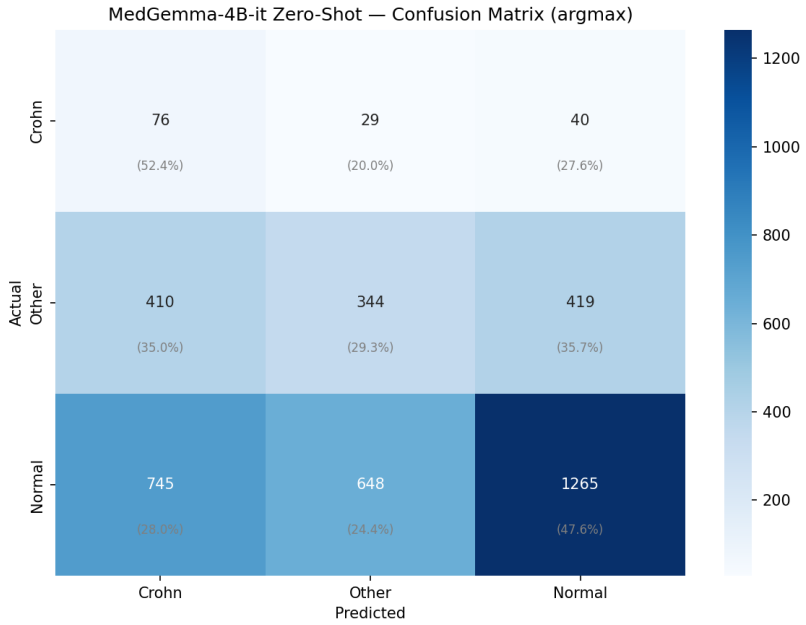


Figure 38: MedGemma 4B zero-shot confusion matrix on the 4,000-frame representative sample. Raw counts with row-normalized percentages in gray. Rows represent ground truth, columns represent predictions.

across the three classes, indicating that MedGemma has *no discriminative ability for the Other class*—it effectively guesses at random.

3. **Normal row (28.0% / 24.4% / 47.6%).** Only 47.6% of Normal frames are correctly classified. The remaining 52.4% are split between Crohn (28.0%) and Other (24.4%), revealing a strong tendency to over-interpret healthy mucosa as pathological.

11.5 Analysis

Systematic Crohn over-prediction.

The most striking finding is the magnitude of Crohn over-prediction. MedGemma predicted Crohn for 31.0% of all frames, despite a ground truth prevalence of only 3.6%—an over-prediction factor of 8.5×. Of the 1,231 Crohn predictions, only 76 were true positives (PPV = 6.2%), meaning that 15 out of every 16 Crohn predictions were false alarms. The resulting Crohn specificity was 69.9% (false positive rate = 30.1%). The false positives originated predominantly from Normal frames (64.5%) rather than Other (35.5%), suggesting that the model interprets normal mucosal features—folds, reflections, villi patterns—as pathological findings. Figure 39 visualizes this distortion: while the ground truth distribution is heavily skewed toward Normal (66.9%), MedGemma's predicted distribution is far more uniform, with Crohn inflated from 3.6% to 31.0% of all predictions.

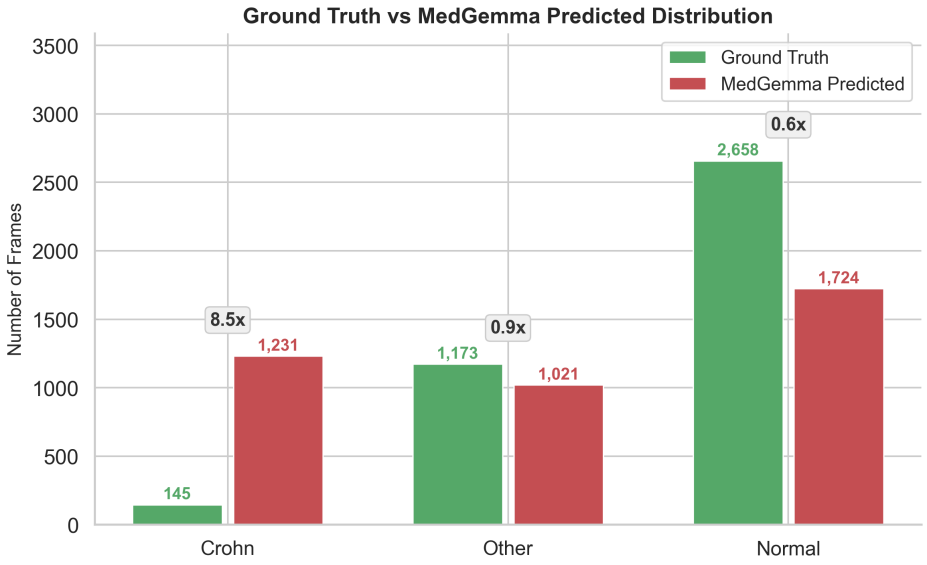


Figure 39: Ground truth vs. MedGemma predicted class distribution on the 3,976 valid frames. The ratio labels indicate over-prediction (8.5× for Crohn) and under-prediction (0.6× for Normal). MedGemma dramatically overestimates Crohn prevalence while underestimating Normal frames.

Prompt-induced prior bias.

The systematic Crohn over-prediction is consistent with a prompt-induced prior bias: by explicitly naming "Crohn" as one of the target classes, the instruction may predispose the model to favour that label independently of the visual evidence. This interpretation is supported by the preliminary prompt-engineering experiment (Section 11.2), in which adding a Crohn-priority clause raised the Crohn prediction rate to approximately 52% of all frames.

Even the final prompt, which includes no such clause, produces an 8.5× over-prediction ratio, suggesting that the class name alone is sufficient to shift the model's output distribution. Recent work on medical VLMs has shown that these models can maintain plausible outputs even when visual grounding is absent or misleading, relying on textual cues and learned statistical priors rather than fine-grained image discrimination. The present results are consistent with this pattern, although a definitive attribution would require ablation experiments beyond the scope of this thesis.

Source dataset dependence.

Performance varied substantially across source datasets:

Source	Frames	Accuracy	Macro-F1
Galar	3,239	38.6%	0.322
Kvasir-Capsule	705	58.3%	0.314
CrohnIPI	31	41.9%	0.318

Kvasir frames achieved higher accuracy (58.3% vs. 38.6% for Galar), driven almost entirely by better Normal recall (60.1% vs. 43.6%). One possible explanation is that Kvasir images are lower-resolution JPEGs with a more "canonical" endoscopic appearance, plausibly closer to the type of data MedGemma encountered during pre-training. Galar frames, which are higher-resolution PNGs from Olympus and PillCam systems, contain finer textural details that the model may misinterpret as pathological findings.

Inter-patient variability.

Crohn recall varied dramatically across patients: Patient 41 achieved 93% (25/27), Patient 25 achieved 90% (19/21), but Patient 43—the largest Crohn patient with 46 frames—reached only 20% (9/46). This extreme variability mirrors the inter-fold variance observed in the MLP cross-validation experiments (Section 9), reinforcing the structural finding that patient diversity is a binding constraint in SBCE classification.

Comparison with threshold-tuned MLP.

A particularly informative comparison is between MedGemma's argmax predictions and the MLP at the r_{95} operating point (Section 9), both of which aim to maximize Crohn detection:

Model	Recall_C	$\text{Precision}_\text{C}$	%Frames flagged
MedGemma (argmax)	0.524	0.062	31.0% (as Crohn)
MLP-1 at r_{95}	0.876	0.052	62.3% (as Crohn)
MLP-1 at r_{90}	0.816	0.060	50.5% (as Crohn)

Even at comparable precision levels ($\sim$0.06), the threshold-tuned MLP achieves substantially higher Crohn recall (0.816–0.876 vs. 0.524), though at the cost of flagging more frames. MedGemma flags 31% of frames as Crohn but only captures 52.4% of actual Crohn frames, while the MLP at r_{90} flags 50.5% but captures 81.6%. The supervised approach is therefore more efficient: per flagged frame, it captures more true Crohn cases.

11.6 Summary

On the 4,000-frame representative sample, MedGemma 4B produced a Macro-F1 of 0.334, below the MLP baseline (0.531) and all linear classifiers from Section 8. The model over-predicted Crohn by a factor of 8.5× relative to the true prevalence, generating 1,155 false positives with a PPV of 6.2%. Its higher Crohn recall (0.524 vs. 0.439) came at a precision cost that would likely be impractical in a triage setting: 56.6% of frames would be flagged as pathological, requiring the clinician to review more than half of all frames while still missing nearly half of actual Crohn cases. These figures are based on a stratified subsample and should be taken as indicative rather than definitive.

These results suggest that, at least under zero-shot conditions, task-specific supervised learning on curated data can outperform a general-purpose medical VLM even when the supervised classifier is a lightweight MLP with fewer than 400,000 parameters. The advantage appears to reside not in architectural complexity but in the combination of domain-aware embeddings, rigorous dataset construction, and calibrated evaluation. The comparison also lends support to the methodological choices made throughout the thesis: the investment in dataset curation and patient-wise splitting produces classifiers that, on this benchmark, perform more reliably than a model orders of magnitude larger applied without domain adaptation.

12 Discussion

12.1 Key Findings and Methodological Takeaways

A central finding of this thesis is that dataset design was not an ancillary preprocessing step but a methodological contribution in its own right. Each intervention—label harmonization across three heterogeneous sources, embedding-guided redundancy reduction, Blood-class bias mitigation, and combinatorial patient-wise splitting—affected the conditions under which subsequent modelling decisions became possible, ultimately shaping class balance, redundancy, and the representativeness of the final splits. The resulting dataset was not simply smaller than the raw pool: it was structurally reconditioned to support screening-oriented modelling under realistic prevalence and hardware constraints.

The classification experiments progressed through two stages. Linear classifiers (Section 8) established a persistent recall–precision trade-off for the Crohn class that was invariant across all classifier families, loss functions, and calibration strategies. Offline data augmentation improved ranking quality (Crohn PR-AUC from 0.177 to 0.276, a 56% relative gain) but did not resolve the precision collapse at high recall ($\sim$4.5% at r_{95}), confirming that the trade-off was structural rather than a consequence of insufficient training data. Introducing a single hidden layer with approximately 393,000 parameters (Section 9) was sufficient to break this barrier, simultaneously improving Crohn recall (0.38 $\rightarrow$ 0.44) and precision (0.20 $\rightarrow$ 0.25). This confirmed that the DINOv2 embedding space contained non-linear discriminative structure that linear boundaries could not exploit.

However, more complex architectures and extensive hyperparameter searches (Optuna campaigns) did not improve upon this simple MLP baseline. This suggests that the remaining performance gap is driven by the structure of the problem—low test prevalence and frame-level independence—rather than by insufficient model capacity. In other words, the bottleneck lies in the data conditions, not in the classification head. The zero-shot MedGemma evaluation (Section 11) reinforced this conclusion from a different angle: a 4-billion-parameter medical VLM achieved lower Macro-F1 than the supervised MLP, with a Crohn over-prediction pattern consistent with prompt-induced prior bias rather than image-grounded discrimination.

The pathology-level experiments (Section 10) changed not only the model but also what was being measured and how. Shifting from aggregated superclasses to individual lesion types and from threshold-dependent frame metrics to a workload-aware triage metric (Recall@10% per patient) redefined the evaluation space: the binary erosion-versus-rest formulation outperformed the three-class formulation ($R@10\% = 0.357$ vs. 0.315) not because the classifier was better, but because the narrower task reduced intra-class confounding and made the triage signal measurable in the first place. In this reformulated setting, split assignment proved more consequential than hyperparameter variation: with ulceration concentrated in only 13 patients and approximately 84% of ulcerative frames originating from two individu-

als, moving a single key patient between splits altered all metrics substantially. The broader implication is that task formulation is itself a design variable—choosing *what to classify* can matter as much as choosing *how to classify it*.

Validation instability was a recurring theme. Cross-validation revealed a standard deviation of $\sim$0.15 on the primary metric across patient-grouped folds, and thresholds calibrated on validation consistently drifted when transferred to the test set (realized workload of 17–18% versus the 10% target). These observations point to patient-level domain shift that algorithmic improvements alone cannot resolve; larger and more representative patient pools remain a prerequisite for stable evaluation.

The workload-aware metric adopted in Section 10 offered a partial methodological response to this instability. By fixing the review budget at 10% of frames per patient, Recall@10% provided a clinically interpretable measure that was not artificially improved by trivial all-positive prediction strategies, and that remained meaningful even when threshold calibration proved unreliable.

Taken together, five messages emerge from the experimental programme:

1. **Dataset design is constitutive of model validity.** In SBCE Crohn screening, curation choices—label harmonization, redundancy reduction, Blood bias mitigation, patient-wise splitting—determine whether downstream metrics reflect genuine generalization or structural artefacts. This is a data-engineering contribution.

2. **Realistic evaluation conditions reveal higher task difficulty.** Under patient-wise, low-prevalence conditions, measured performance decreases substantially compared to more favorable setups (balanced prevalence, random splits). This is an evaluation-realism contribution: the resulting estimates are more conservative but also more informative about real-world deployment.

3. **Task formulation matters as much as model choice.** Reformulating the problem from three-class superclass classification to binary erosion-versus-rest triage changed what was measurable and improved triage quality, not through a better classifier but through a narrower, clinically grounded task definition.

4. **Workload-aware metrics are more appropriate for screening.** Recall@10% per patient directly models the clinical constraint of limited review time and is not artificially improved by trivial all-positive strategies, unlike threshold-dependent frame-level metrics.

5. **Large-scale medical pre-training does not substitute for task-specific curation.** The zero-shot MedGemma evaluation (Section 11) showed that a 4-billion-parameter medical VLM achieved lower Macro-F1 than a supervised MLP with fewer than 400,000 parameters. The model's systematic Crohn over-prediction—consistent with prompt-induced prior bias—suggests that general-purpose medical VLMs do not

yet offer a reliable shortcut for domain-specific screening tasks that require fine-grained visual discrimination.

12.2 Contextualization with the Existing Literature

Comparisons with prior SBCE studies must be interpreted cautiously, as published results are strongly shaped by differences in task formulation, class prevalence, split strategy, data source heterogeneity, and degree of model adaptation. The systematic review by Soffer et al. [10] identified these heterogeneities as pervasive in the WCE deep learning literature, noting that most studies relied exclusively on internal validation without external test sets. The more recent meta-analysis by Ali et al. [41], covering eight studies and 444 patients (2020–2024), reported pooled sensitivity of 94% and specificity of 97% for Crohn lesion detection, but explicitly noted that none of the included studies performed external validation. Rather than attempting a single ranking, the following paragraphs organize the comparison along three axes and highlight what each reveals about the present work.

Superclass classification (Sections 8–9) versus the classification-oriented literature.

Table 1 places the most relevant published studies alongside the results from the linear and non-linear classifier experiments.

The gap in absolute numbers is evident. What the comparison reveals, however, is how much of that gap can be attributed to differences in evaluation conditions rather than in modelling quality. The present work operates on a three-class task (Crohn / Other / Normal) that is inherently harder than the binary formulations used by most studies, because the Other class shares visual features with both Crohn and Normal. The test prevalence of 3.4% reflects realistic screening conditions, whereas most studies operate under balanced or enriched distributions that favour accuracy-based metrics. All cited studies fine-tune the feature extractor end-to-end, whereas the present work uses a frozen generic encoder (DINOv2 ViT-B/14) with a classification head of fewer than 400,000 parameters—a deliberate choice to isolate the contribution of the classification architecture, which also places a ceiling on achievable performance that fine-tuned systems do not face. Finally, patient-wise splitting across three sources and three capsule devices imposes a stricter generalization requirement than single-center random splits. Klang et al. [11] provided direct evidence of this effect: on the same dataset, their AUC dropped from 0.99 to 0.94 simply by switching from random to patient-level evaluation. This thesis contributes evidence on how performance changes when evaluation is shifted from favorable classification setups to realistic patient-wise screening conditions.

Study	Task	Dataset	Split	Encoder	Main result
Klang 2020 [11]	Binary (ulcer / normal)	Priv., 49 pt, 1 ctr	Rand. + Pt.	Fine-tuned	AUC 0.99 / 0.94
Aoki 2019 [42]	Binary (eros.+ulcer)	Priv., indep. test	?	Fine-tuned	AUC 0.958
Majtner 2021 [12]	5-class severity	Priv., 38 pt, 3 ctr	Patient	Fine-tuned	Sens 95.7%
Afonso 2022 [43]	Binary (ulcer / eros.)	Priv., 1 ctr	Random	Fine-tuned	Acc 95.6%
de Maissin 2021 [20]	Binary (path. / norm.)	CrohnIPI, 63 pt	Random	Fine-tuned	Acc 93.7%
Zhang 2024 [14]	11-class	Kvasir-Capsule	Random	Partial FT	Macro-F1 87.8%
This thesis (Sec. 8–9)	*3-class*	*3 src, 3 dev, 100 groups*	*Patient*	*Frozen*	*Macro-F1 0.53*

Table 1: Classification performance of Crohn-related capsule endoscopy studies. The Split and Encoder columns highlight the two key methodological differences: most studies use random splits and fine-tuned encoders, whereas this thesis uses patient-wise splitting and a frozen DINOv2 backbone. "?" = not specified; "Partial FT" = partial fine-tuning.

Pathology-level triage (Section 10) versus the workload-reduction literature.

Table 2 places the triage results alongside studies that address workload reduction in capsule endoscopy reading.

The best model in this thesis (binary erosion triage, $R@10\% = 0.357$) captured 3.6× more erosions than random selection within the same 10% frame budget. This falls short of the reading-time reductions reported in clinical deployment studies, which typically use end-to-end fine-tuned models on curated single-source datasets. However, the comparison also highlights a difference in evaluation philosophy: reading-time studies measure efficiency gains in real clinical workflows, whereas Recall@10% provides a model-intrinsic metric that quantifies triage quality independently of reader speed and interface design. The two approaches are complementary rather than competing, and the present work's contribution is to demonstrate that even frozen generic features can produce a non-trivial triage signal under rigorous patient-wise evaluation.

Methodological profile comparison.

Beyond raw performance, Table 3 compares the methodological conditions under which each study was evaluated.

Study	Approach	Evaluation	Main result
Aoki 2020 [4]	CNN pre-screening, single-centre	Reading time, sensitivity	−75% reading time, Sens 87%
Spada 2024 [5]	AI-assisted reading, multicentre prospective	Reading time, diagnostic yield	Significant time reduction
Oh 2024 [35]	Frame reduction	Final diagnosis preserved	Frame reduction without diagnostic loss
This thesis (Sec. 10)	*MLP on frozen DINOv2, patient-wise eval.*	*Recall@10% workload per patient*	*R@10% = 0.357 (3.6× random)*

Table 2: Workload-reduction studies in capsule endoscopy. Reading-time studies measure efficiency gains in clinical workflows; the present work uses a model-intrinsic, threshold-free metric that is independent of reader speed and interface design.

The present work is the only study that combines all five conditions. Each independently makes the evaluation more demanding; together, they produce a setting in which performance estimates are more conservative but also more representative of real-world screening difficulty. The point is not that stricter methodology compensates for lower numbers, but rather that the two are causally related: when evaluation is made more realistic, the measured difficulty of the task increases accordingly. This observation aligns with the broader concern raised by Soffer et al. [10] about methodological quality in WCE deep learning and suggests that headline performance figures in this domain should be interpreted with care.

The domain-specific foundation model EndoDINO [15], pre-trained on 3.5 billion endoscopy frames using DINOv2 self-supervised learning, achieved a Macro-F1 of 0.74 on 3-class Mayo endoscopic scoring with linear probing. While the task is not identical, this provides a useful reference for what domain-adapted frozen features can achieve on a comparable multi-class endoscopic classification problem. Closing the gap between generic and domain-specific encoders represents a promising direction for improving absolute performance while preserving the methodological rigor established in this thesis.

12.3 Limitations and Next Steps

Key limitations of the present work include the small patient pool that drove extreme split sensitivity (particularly for ulceration, concentrated in 13 patients), the absence of demographic stratification, the lack of explicit temporal modelling, the reliance on generic DINOv2 embeddings not specialized for mucosal lesion detection, and inter-observer agreement on aphthoid erosions that is limited even among expert gastroenterologists—placing a fundamental upper bound on achievable automated performance. The splitting formulation did

Study	*Patient-wise split*	*Multi-source data*	*Realistic prevalence*	*Multi-class task*	*Frozen encoder*
Klang 2020 [11]	Partial	–	–	–	–
Aoki 2019 [42]	?	–	✓	–	–
Majtner 2021 [12]	✓	✓	?	✓	–
Afonso 2022 [43]	–	–	–	–	–
de Maissin 2021 [20]	–	–	–	–	–
Zhang 2024 [14]	–	–	–	✓	Partial
Aoki 2020 [4]	?	–	✓	–	–
This thesis	✓	✓	✓	✓	✓

Table 3: Methodological profile of cited studies. ✓ = yes, – = no, ? = not specified, Partial = partially addressed.

not impose hard constraints on total split sizes or enforce minimum patient counts per split.

These limitations suggest four concrete directions for future work, ordered by expected impact:

1. **Expand the patient pool and redesign evaluation around patient diversity.** The most binding constraint identified across Sections 9–10 is patient scarcity: cross-validation variance ($\sim$0.15 std), split sensitivity, and threshold drift all trace back to too few Crohn-positive patients. The priority is to increase the number of patients with ulceration—currently concentrated in 13 individuals—and to incorporate additional centres and capsule systems. With a larger patient pool, patient-grouped cross-validation could serve as the primary evaluation protocol rather than a supplementary diagnostic.

2. **Move from frame-independent prediction to temporal or exam-level aggregation.** The current system treats each frame independently, ignoring that consecutive frames depict the same mucosal segment and that the clinically relevant unit is the examination, not the individual frame. Practical extensions include temporal smoothing over sliding windows, patient-level score aggregation, and exam-level triage ranking. These do not require large-scale architectural changes: even simple averaging over local windows could reduce the noise inherent in frame-level predictions.

3. **Replace generic frozen features with domain-adapted endoscopic representations.** The frozen DINOv2 ViT-B/14 encoder used throughout this thesis was pretrained on natural images, not on mucosal tissue. Domain-specific encoders such as

EndoDINO [15]—pre-trained on 3.5 billion endoscopy frames—have demonstrated substantially higher Macro-F1 on comparable multi-class endoscopic tasks (0.74 vs. 0.53). A systematic comparison between generic DINOv2, fine-tuned DINOv2, and domain-adapted encoders under the same patient-wise evaluation protocol established in this thesis would isolate the contribution of feature quality from the contributions of dataset design and evaluation rigour.

4. **Test the approach prospectively in a workflow-oriented setting.** The Recall@10% metric quantifies triage quality model-intrinsically, but does not measure whether the system actually reduces reading time or improves diagnostic yield in clinical practice. A prospective study integrating the triage pipeline into a capsule reading workflow would bridge this gap and validate whether the workload-aware evaluation framework translates into real clinical benefit.

13 Conclusion

This thesis investigated automated Crohn's disease screening in Small-Bowel Capsule Endoscopy through a data-centric lens. Rather than treating dataset preparation as a preliminary technical step, the work showed that data construction, patient-wise split design, and evaluation strategy are central determinants of what model performance actually means in this domain. In this sense, the main contribution of the thesis is methodological before it is architectural: it provides a more rigorous framework for studying Crohn screening under realistic conditions.

A first contribution lies in the construction of a unified SBCE pipeline from heterogeneous public sources. By harmonizing labels, reducing redundancy in the embedding space, mitigating Blood-related bias, and enforcing patient-wise splitting through constrained optimization, the thesis demonstrates that reliable experimentation in this setting requires explicit control of the data-generating conditions. This is especially important in SBCE, where lesions may be sparse and transient, videos are highly redundant, and apparent gains can easily be inflated by favorable split configurations or by overly simplified class formulations.

A second contribution is the clear identification of a structural limitation in frozen-feature classification. Linear models established a strong baseline, but also revealed a persistent Crohn recall–precision barrier that could not be resolved through loss changes or calibration alone. A shallow MLP improved over this baseline, confirming that the DINOv2 embedding space contains useful non-linear structure. At the same time, the limited gains obtained with more complex optimization campaigns indicate that the main bottleneck is not simply model capacity. Instead, performance is largely governed by realistic prevalence, scarcity of positive patients, and the weakness of frame-independent decision-making in a sequential visual domain.

A third contribution concerns evaluation itself. The pathology-level experiments showed that reformulating the task around workload-aware triage can be more informative than relying only on conventional frame-level metrics. In particular, Recall@10% per patient offered a more meaningful proxy for screening use, because it directly asked how many relevant lesions could be captured within a fixed review budget. Within this setting, the erosion-versus-rest formulation produced the strongest signal, while the ulceration class remained too fragile due to extreme concentration in very few patients. These experiments also led to one of the clearest findings of the thesis: in patient-scarce medical imaging tasks, split assignment may influence results more strongly than hyperparameter tuning.

Finally, the zero-shot comparison with MedGemma adds an important complementary lesson. Although large medical vision–language models are appealing because of their scale and flexibility, the experiments in this thesis suggest that scale alone does not guarantee robust behavior in fine-grained SBCE screening. The strong over-prediction of the Crohn class observed in the zero-shot setting indicates that prompt priors and shortcut-like behavior may distort the predicted class distribution, especially when the task requires subtle visual

discrimination under realistic prevalence. In this context, a lightweight supervised pipeline built on curated data proved more reliable than a much larger zero-shot model.

Overall, this thesis does not claim to have solved automated Crohn detection in SBCE. Its main value is different: it makes the difficulty of the problem visible under conditions that are much closer to realistic screening than many favorable benchmark-style setups. Many of the observed limitations—the fragility of split assignment, the persistence of the recall–precision barrier, the sensitivity to task formulation—are amplified by the intrinsic difficulty of the SBCE domain, where discriminative cues are fine-grained, transient, and often visually ambiguous even to expert readers. The work shows that trustworthy progress in this field depends not only on better models, but on better-controlled data, better patient-level evaluation, and task formulations that reflect operational constraints. Future advances will likely come from combining this methodological rigor with larger patient cohorts, domain-adapted endoscopic representations, and temporal or exam-level modelling strategies. In the meantime, the framework developed here offers a reproducible basis for studying Crohn screening in SBCE with more conservative, but more credible, expectations of performance.

References

[1] J. Torres, S. Mehandru, J.-F. Colombel, and L. Peyrin-Biroulet, "Crohn's disease," *The Lancet*, vol. 389, no. 10080, pp. 1741–1755, 2017. DOI: 10.1016/S0140-6736(16)31711-1

[2] M. Pennazio et al., "Small-bowel capsule endoscopy and device-assisted enteroscopy for diagnosis and treatment of small-bowel disorders: European Society of Gastrointestinal Endoscopy (ESGE) clinical guideline," *Endoscopy*, vol. 47, no. 4, pp. 352–376, 2015. DOI: 10.1055/s-0034-1391855

[3] Z. Ding et al., "Gastroenterologist-level identification of small-bowel diseases and normal variants by capsule endoscopy using a deep-learning model," *Gastroenterology*, vol. 157, no. 4, 1044–1054.e5, 2019. DOI: 10.1053/j.gastro.2019.06.025

[4] T. Aoki et al., "Clinical usefulness of a deep learning-based system as the first screening on small bowel capsule endoscopy reading," *Digestive Endoscopy*, vol. 32, no. 4, pp. 585–591, 2020. DOI: 10.1111/den.13562

[5] C. Spada et al., "Artificial intelligence-assisted capsule endoscopy reading in suspected small bowel bleeding: A multicentre prospective study," *Lancet Digital Health*, 2024. DOI: 10.1016/S2589-7500(24)00121-4

[6] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006. DOI: 10.1016/j.patrec.2005.10.010

[7] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, e0118432, 2015. DOI: 10.1371/journal.pone.0118432

[8] R. Leenhardt et al., "A guide for assessing the clinical relevance of findings in small bowel capsule endoscopy: Analysis of 8064 answers of international experts to an illustrated script questionnaire," *Clinical Research in Hepatology and Gastroenterology*, vol. 45, no. 6, p. 101 637, 2021. DOI: 10.1016/j.clinre.2021.101637

[9] D. Yablecovitch et al., "The Lewis score or the capsule endoscopy Crohn's disease activity index: Which one is better for the assessment of small bowel inflammation in established Crohn's disease?" *Therapeutic Advances in Gastroenterology*, vol. 11, p. 1756283X17747780, 2018. DOI: 10.1177/1756283X17747780

[10] S. Soffer et al., "Deep learning for wireless capsule endoscopy: A systematic review and meta-analysis," *Gastrointestinal Endoscopy*, vol. 92, no. 4, 831–839.e8, 2020. DOI: 10.1016/j.gie.2020.04.039

[11] E. Klang et al., "Deep learning algorithms for automated detection of Crohn's disease ulcers by video capsule endoscopy," *Gastrointestinal Endoscopy*, vol. 91, no. 3, 606–613.e2, 2020. DOI: 10.1016/j.gie.2019.11.012

[12] T. Majtner, J. B. Brodersen, J. Herp, J. Kjeldsen, M. L. Halling, and M. D. Jensen, "A deep learning framework for autonomous detection and classification of Crohn's disease lesions in the small bowel and colon with capsule endoscopy," *Endoscopy International Open*, vol. 9, no. 9, E1361–E1370, 2021. DOI: 10.1055/a-1507-4980

[13] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," 2024. arXiv: 2304.07193 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2304.07193

[14] B. Zhang et al., "Learning to adapt foundation model DINOv2 for capsule endoscopy diagnosis," 2024. arXiv: 2406.10508 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2406.10508

[15] P. Dermyer, A. Kalra, and M. Schwartz, "EndoDINO: A foundation model for GI endoscopy," 2025. DOI: 10.48550/arXiv.2501.05488 arXiv: 2501.05488.

[16] Z. Wang, C. Liu, L. Zhu, T. Wang, S. Zhang, and Q. Dou, "Improving foundation model for endoscopy video analysis via representation learning on long sequences," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 5, pp. 3526–3536, 2025. DOI: 10.1109/JBHI.2025.3532311

[17] S. Zhu et al., "Public imaging datasets of gastrointestinal endoscopy for artificial intelligence: A review," *Journal of Digital Imaging*, vol. 36, no. 6, pp. 2578–2601, 2023. DOI: 10.1007/s10278-023-00844-7

[18] M. Le Floch et al., "GALAR – a large multi-label video capsule endoscopy dataset," *Scientific Data*, vol. 12, p. 828, 2025. DOI: 10.1038/s41597-025-05112-7

[19] P. H. Smedsrud et al., "Kvasir-capsule, a video capsule endoscopy dataset," *Scientific Data*, vol. 8, p. 142, 2021. DOI: 10.1038/s41597-021-00920-z

[20] A. de Maissin et al., "Multi-expert annotation of Crohn's disease images of the small bowel for automatic detection using a convolutional recurrent attention neural network," *Endoscopy International Open*, vol. 9, E1136–E1144, 2021. DOI: 10.1055/a-1468-3964

[21] G. Wang, P. Wang, and B. Wei, "Multi-label local awareness and global co-occurrence priori learning improve chest x-ray classification," *Multimedia Systems*, vol. 30, 2024. DOI: 10.1007/s00530-024-01321-z

[22] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: Review of a decade of research," *Artificial Intelligence Review*, vol. 57, p. 273, 2024. DOI: 10.1007/s10462-024-10884-2

[23] T. J. Jaspers et al., "Robustness evaluation of deep neural networks for endoscopic image analysis: Insights and strategies," *Medical Image Analysis*, 2024. DOI: 10.1016/j.media.2024.103402

[24] W. Ye et al., *The clever hans mirage: A comprehensive survey on spurious correlations in machine learning*, 2025. arXiv: 2402.12715 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2402.12715

[25] A. F. Faria, S. R. de Souza, and E. M. de Sá, "A mixed-integer linear programming model to solve the Multidimensional Multi-Way Number Partitioning Problem," *Computers & Operations Research*, vol. 127, p. 105 133, 2021. DOI: 10.1016/j.cor.2020.105133

[26] Y.-C. Lin and Y.-S. Chen, "Weighted stratification in multi-label contrastive learning for long-tailed medical image classification," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, ser. Lecture Notes in Computer Science, vol. 15972, Springer, 2026, pp. 677–687. DOI: 10.1007/978-3-032-05169-1_65

[27] C. Chekuri and S. Khanna, "On multidimensional packing problems," *SIAM Journal on Computing*, vol. 33, no. 4, pp. 837–851, 2004. DOI: 10.1137/S0097539799356265

[28] A. Charnes and W. W. Cooper, "Goal programming and multiple objective optimizations: Part 1," *European Journal of Operational Research*, vol. 1, no. 1, pp. 39–54, 1977. DOI: 10.1016/0377-2217(77)90066-6

[29] R. L. Graham, "Bounds for certain multiprocessing anomalies," *Bell System Technical Journal*, vol. 45, no. 9, pp. 1563–1581, 1966. DOI: 10.1002/j.1538-7305.1966.tb01709.x

[30] E. L. Schreiber, R. E. Korf, and M. D. Moffitt, "Optimal multi-way number partitioning," *Journal of the ACM*, vol. 65, no. 4, 2018. DOI: 10.1145/3184400

[31] Google Optimization Team, *OR-Tools: CP-SAT solver*, Accessed: 2025, 2025. [Online]. Available: https://developers.google.com/optimization/cp/cp_solver

[32] D. Varam et al., "Wireless capsule endoscopy image classification: An explainable AI approach," *IEEE Access*, vol. 11, pp. 105 262–105 280, 2023. DOI: 10.1109/ACCESS.2023.3319068

[33] T. T. Habe, K. Haataja, and P. Toivanen, "Precision enhancement in wireless capsule endoscopy: A novel transformer-based approach for real-time video object detection," *Frontiers in Artificial Intelligence*, vol. 8, p. 1 529 814, 2025. DOI: 10.3389/frai.2025.1529814

[34] H. Morera et al., "Reduction of video capsule endoscopy reading times using deep learning with small data," *Artificial Intelligence in Medicine*, 2024. DOI: 10.1016/j.artmed.2024.102829

[35] D. J. Oh, Y. Hwang, S. H. Kim, J. H. Nam, M. K. Jung, and Y. J. Lim, "Reading of small bowel capsule endoscopy after frame reduction using an artificial intelligence algorithm," *Intestinal Research*, 2024. DOI: 10.5217/ir.2023.00109

[36] M. Saerens, P. Latinne, and C. Decaestecker, "Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure," *Neural Computation*, vol. 14, no. 1, pp. 21–41, 2002. DOI: 10.1162/089976602753284446

[37] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, vol. 70, PMLR, 2017, pp. 1321–1330.

[38] A. Sellergren et al., "MedGemma technical report," 2025. arXiv: 2507.05201 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2507.05201

[39] M. S. I. Prottasha and N. W. Rafi, "MedGemma vs GPT-4: Open-source and proprietary zero-shot medical disease classification from images," 2025. arXiv: 2512.23304 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2512.23304

[40] S. Yu et al., "MedFrameQA: A multi-image medical VQA benchmark for clinical reasoning," 2025. arXiv: 2505.16964 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2505.16964

[41] H. Ali et al., "Artificial intelligence-assisted capsule endoscopy for detecting lesions in Crohn's disease: A systematic review and meta-analysis," *Frontiers in Artificial Intelligence*, vol. 8, 2025. DOI: 10.3389/frai.2025.1531362

[42] T. Aoki et al., "Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network," *Gastrointestinal Endoscopy*, vol. 89, no. 2, 357–363.e2, 2019. DOI: 10.1016/j.gie.2018.10.027

[43] J. Afonso, M. Mascarenhas Saraiva, J. P. S. Ferreira, T. Ribeiro, H. Cardoso, and G. Macedo, "Automated detection of ulcers and erosions in capsule endoscopy images using a convolutional neural network," *Medical & Biological Engineering & Computing*, vol. 60, no. 3, pp. 719–725, 2022. DOI: 10.1007/s11517-021-02486-9

A Category Details and Superclass Mapping

The three source datasets adopted different annotation vocabularies: GALAR used an internal multi-label taxonomy, Kvasir-Capsule used medically verified single-label categories, and CrohnIPI used Crohn-specific severity grading. To enable joint modelling, all annotations were harmonized into a unified taxonomy of 28 categories, each assigned a numeric ID and mapped to one of three screening superclasses (Crohn, Other, Normal). The superclass assignment follows a clinical rationale: categories representing mucosal breaks characteristic of Crohn's disease map to **Crohn**; other pathological findings not specific to Crohn map to **Other**; and normal mucosa, anatomical landmarks, and image-quality descriptors map to **Normal**. For multi-label frames (GALAR only), the priority rule $C > O > N$ assigns the frame to the highest-priority superclass present.

Table 4 reports the unified mapping, while Table 5 shows how the original source-specific labels were mapped to unified category IDs.

ID	Category name	Original supercategory	Superclass
000	Normal	Pathology findings	Normal
001	Erythematous patch	Pathology findings	Other
002	Edema	Pathology findings	Other
003	Aphthoid erosion	Pathology findings	Crohn
004	Superficial ulceration	Pathology findings	Crohn
005	Deep ulceration	Pathology findings	Crohn
006	Stenosis	Pathology findings	Other
007	Polyp	Pathology findings	Other
008	Blood	Pathology findings	Other
009	Active bleeding	Pathology findings	Other
010	Hematin	Pathology findings	Other
011	Typical angiectasia	Pathology findings	Other
012	Lymphangiectasia	Pathology findings	Other
013	Foreign body	Pathology findings	Other
099	Other	Pathology findings	Other
100	Z-line	Anatomy	Normal
101	Pylorus	Anatomy	Normal
102	Ampulla of Vater	Anatomy	Normal
103	Ileocecal valve	Anatomy	Normal
104	Mouth	Anatomy	Normal
105	Esophagus	Anatomy	Normal
106	Stomach	Anatomy	Normal
107	Small intestine	Anatomy	Normal
108	Colon	Anatomy	Normal
200	Bubbles	Quality	Normal
201	Dirt	Quality	Normal
202	Reduced view	Quality	Normal
203	Good view	Quality	Normal

Table 4: Complete category-to-superclass mapping. Rows shaded in blue indicate categories assigned to the Crohn superclass. The schema is shared across all three source datasets; category availability varies by source.

ID	Unified name	GALAR label	Kvasir-Capsule label	CrohnIPI label
000	Normal	—	normal clean mucosa	Normal
001	Erythematous patch	erythema	erythema	Erythema
002	Edema	—	—	Edema
003	Aphthoid erosion	erosion	erosion	Aphthoid ulceration
004	Superficial ulceration	—	—	Ulceration 3–10 mm
005	Deep ulceration	ulcer	ulcer	Ulceration >10 mm
006	Stenosis	—	—	Stenosis
007	Polyp	polyp	polyp	—
008	Blood	blood	blood fresh	—
009	Active bleeding	active bleeding	—	—
010	Hematin	hematin	hematin	—
011	Typical angiectasia	angiectasia	angiectasia	—
012	Lymphangiectasia	lymphangioectasis	lymphangiectasia	—
013	Foreign body	foreign body	foreign body	—
099	Other	—	—	—
100	Z-line	z-line	—	—
101	Pylorus	pylorus	pylorus	—
102	Ampulla of Vater	ampulla of vater	ampulla of vater	—
103	Ileocecal valve	ileocecal valve	ileocecal valve	—
104	Mouth	mouth	—	—
105	Esophagus	esophagus	—	—
106	Stomach	stomach	—	—
107	Small intestine	small intestine	—	—
108	Colon	colon	—	—
200	Bubbles	bubbles	—	—
201	Dirt	dirt	—	—
202	Reduced view	reduced view	reduced mucosal view	—
203	Good view	good view	normal clean mucosa	—

Table 5: Cross-reference of original source-specific labels to unified category IDs. "—" indicates the category was not present in that dataset. Note that Kvasir-Capsule used "normal clean mucosa" for both Normal (000) and Good view (203); CrohnIPI distinguished ulceration severity by size (aphthoid, 3–10 mm, >10 mm), which was mapped to three separate Crohn categories. GALAR categories excluded from the unified schema (esophagitis, IBD, cancer, celiac, varices) were removed during early curation (Section 3.3).

B Published Paper

The following pages contain the full text of the workshop paper, presented at the *Multimodal Representation Learning for Healthcare* workshop held at EurIPS 2025 in Copenhagen, Denmark (December 6, 2025). The workshop focused on integrating diverse medical data modalities into interpretable patient representations, bringing together machine learning researchers, clinicians, and industry partners.

F. Felizzi, O. Riccomi, M. Ferramola, F. A. Causio, M. Del Medico, V. De Vita, L. De Mori, A. Piscitelli, P. E. Risuleo, B. Destro Castaniti, A. Cristiano, A. Longo, L. De Angelis, M. Vassalli, M. Di Pumpo. *Are Large Vision Language Models Truly Grounded in Medical Images? Evidence from Italian Clinical Visual Question Answering.* MMRL4H Workshop, EurIPS 2025.

This paper investigates visual grounding in frontier vision language models (Claude Sonnet 4.5, GPT-4o, GPT-5-mini, Gemini 2.0) on Italian medical visual question answering, revealing that most models maintain high accuracy even when correct images are replaced with blank placeholders.

Are Large Vision Language Models Truly Grounded in Medical Images? Evidence from Italian Clinical Visual Question Answering

Federico Felizzi[1,*], **Olivia Riccomi**[1], **Michele Ferramola**[2], **Francesco Andrea Causio**[3,1], Manuel Del Medico[3,1], Vittorio De Vita[3,1], Lorenzo De Mori[1,4], Alessandra Piscitelli[1,5], Pietro Eric Risuleo[3,1], Bianca Destro Castaniti[1,5], Antonio Cristiano[3,1], Alessia Longo[6], Luigi De Angelis[1,7], Mariapia Vassalli[1,5], Marcello Di Pumpo[3,1]

[1]SIIAM, Rome, Italy [2]NSBProject, Mantova, Italy
[3]Dept. of Life Sciences & Public Health, UCSC, Rome, Italy
[4]ASL RM 4, Bracciano, Italy [5]UCSC, Rome, Italy
[6]Univ. Paris Cité, France [7]Univ. of Pisa, Italy [*]Corresp. author: federico.felizzi@gmail.com

Abstract

Large vision language models (VLMs) have achieved impressive performance on medical visual question answering benchmarks, yet their reliance on visual information remains unclear. We investigate whether frontier VLMs demonstrate genuine visual grounding when answering Italian medical questions by testing four state-of-the-art models: Claude Sonnet 4.5, GPT-4o, GPT-5-mini, and Gemini 2.0 flash exp. Using 60 questions from the EuropeMedQA Italian dataset that explicitly require image interpretation, we substitute correct medical images with blank placeholders to test whether models truly integrate visual and textual information. Our results reveal striking variability in visual dependency: GPT-4o shows the strongest visual grounding with a 27.9pp accuracy drop (83.2% [74.6%, 91.7%] to 55.3% [44.1%, 66.6%]), while GPT-5-mini, Gemini, and Claude maintain high accuracy with modest drops of 8.5pp, 2.4pp, and 5.6pp respectively. Analysis of model-generated reasoning reveals confident explanations for fabricated visual interpretations across all models, suggesting varying degrees of reliance on textual shortcuts versus genuine visual analysis. These findings highlight critical differences in model robustness and the need for rigorous evaluation before clinical deployment.

1 Introduction

Recent advances in large vision language models have led to remarkable performance on medical benchmarks, with systems approaching or exceeding human expert performance on visual question answering tasks [1]. However, high benchmark scores may mask fundamental limitations in how these models process and integrate visual information with clinical reasoning [2, 3]. The medical AI community faces a critical question: do these models succeed through genuine multimodal understanding, or do they exploit spurious correlations and textual shortcuts? This question is particularly important for healthcare applications, where erroneous diagnoses based on faulty visual reasoning could have serious consequences. Building on recent work that exposed hidden fragilities in frontier models through systematic stress testing [1], we investigate visual grounding in medical question answering using Italian clinical cases. Our approach differs from prior work by (1) comparing multiple frontier VLMs, (2) focusing on a non-English medical dataset, and (3) employing a targeted visual substitution methodology that tests whether models truly rely on image content when rendering diagnostic judgments.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MMRL4H.

1.1 Contributions

- We present the first systematic comparison of four frontier VLMs (Claude Sonnet 4.5, GPT-4o, GPT-5-mini, and Gemini 2.0 flash exp - referred to as Gemini 2.0) on 60 Italian medical visual question answering cases requiring explicit image interpretation.
- We introduce a visual substitution methodology revealing striking differences in visual dependency across models, with accuracy drops ranging from 2.4pp to 27.9pp.
- We provide empirical evidence that most current VLMs maintain surprisingly high accuracy with incorrect images, suggesting varying reliance on textual cues rather than robust visual understanding.

2 Related Work

Medical Visual Question Answering. Medical VQA benchmarks such as VQA-RAD [4], PMC-VQA [5], and PathVQA [6] have driven progress in multimodal medical AI. However, recent work has questioned whether these benchmarks truly measure medical understanding or merely test-taking ability [1].

Robustness and Shortcut Learning. The ML community has documented extensive shortcut learning in vision-language models [7], where models exploit spurious correlations rather than learning robust features. In medical imaging, this manifests as reliance on metadata, dataset artifacts, or textual priors rather than genuine visual analysis [8].

Stress Testing Large Models. Recent work by Microsoft Research [1] introduced systematic stress tests revealing that frontier models often succeed for the wrong reasons, maintaining high accuracy even when critical inputs are removed or perturbed. Our work extends this methodology to Italian medical cases with comparative analysis across multiple VLMs.

3 Methodology

3.1 Dataset

We utilized the EuropeMedQA dataset [9], specifically the Italian State Exam for Medical Doctors (SSM) subset. From this collection, we manually curated 60 multiple-choice questions that explicitly require visual interpretation for correct diagnosis. Questions span cardiology (27%), orthopedics (12%), dermatology (13%), neurology (10%), gastroenterology and pulmonology (8% each), and other specialties including preventive medicine/epidemiology (5%), oncology (3%), and hematology, ophthalmology, and trauma surgery (2% each).

Each question includes a clinical vignette in Italian, a medical image (X-ray, CT scan, dermatological photo, ECG, etc.), five answer options (A-E), and the ground truth correct answer.

3.2 Experimental Design

We conducted a visual substitution experiment across four frontier VLMs: Claude Sonnet 4.5, GPT-4o, GPT-5-mini, and Gemini 2.0. For each model:

Original Condition. The model answered questions with correct medical images attached, generating both an answer selection and detailed reasoning.

Substitution Condition. We replaced each medical image with an identical blank placeholder while keeping question text and answer options unchanged. Models truly dependent on visual information should show decreased accuracy when diagnostically relevant images are replaced.

We prompted all models to provide both answer selection and detailed step-by-step reasoning using chain-of-thought prompting, following [1]. This allowed analysis of whether explanations reflected actual image content or hallucinated features.

3.3 Evaluation Metrics

We measured: (1) **Accuracy** in original vs. substitution conditions, (2) **Accuracy drop** as the primary indicator of visual dependency, and (3) **Reasoning quality** through manual analysis of generated explanations for hallucinations and misaligned visual descriptions.

4 Results

4.1 Quantitative Analysis

Table 1 summarizes our comparative findings across 10 repetitions per model. The models show striking variability in visual dependency:

GPT-4o demonstrates the strongest visual grounding with 83.2% accuracy (95% CI: [74.6%, 91.7%]) on real images dropping to 55.3% (95% CI: [44.1%, 66.6%]) with fake images (27.9pp decrease), suggesting substantial reliance on actual visual content for diagnostic reasoning.

GPT-5-mini achieves the highest baseline accuracy (88.0%, 95% CI: [81.3%, 94.7%]) but maintains 79.5% (95% CI: [69.7%, 89.3%]) with substituted images (8.5pp drop), indicating improved textual reasoning but potentially less visual dependency than GPT-4o.

Gemini 2.0 shows 83.7% accuracy (95% CI: [74.3%, 93.0%]) with real images and 81.3% (95% CI: [71.7%, 91.0%]) with fake images (2.4pp drop), demonstrating the smallest performance degradation and suggesting strong reliance on textual cues.

Claude Sonnet 4.5 achieves 82.8% (95% CI: [73.7%, 91.9%]) with real images and 77.2% (95% CI: [66.6%, 87.7%]) with fake images (5.6pp drop), showing moderate visual dependency between GPT-4o and the other models.

Table 1: Comparative performance of four frontier VLMs on Italian medical VQA with correct vs. substituted images (N=60 questions, 10 repetitions per model).

Model	Real Images	Fake Images	Drop
GPT-5-mini	88.0% [81.3, 94.7]	79.5% [69.7, 89.3]	8.5pp
Gemini 2.0	83.7% [74.3, 93.0]	81.3% [71.7, 91.0]	2.4pp
GPT-4o	83.2% [74.6, 91.7]	55.3% [44.1, 66.6]	**27.9pp**
Claude Sonnet 4.5	82.8% [73.7, 91.9]	77.2% [66.6, 87.7]	5.6pp

For context, human performance on the Italian State Exam in 2024 averaged 74.8%, with 9.6% of test takers scoring above 95.6% [10]. All models exceed average human performance with real images, but GPT-4o drops significantly below the human average when images are removed, largely because it refuses to answer the question, while the other models maintain superhuman accuracy even without visual information.

4.2 Qualitative Analysis of Reasoning

We identified three recurring patterns in model-generated explanations across all four VLMs:

Hallucinated Visual Features. Models frequently described specific visual findings absent from images. For example, when shown a blank placeholder for an anterior MI question (correct answer: C describing precordial ST elevation), multiple models confidently described fabricated ECG findings matching various answer options, despite viewing diagnostically empty images.

Answer-Driven Reasoning. Models appeared to select answers first (possibly from textual cues), then construct visual justifications post-hoc. This was evident when identical questions with different images received the same answers but with contradictory visual descriptions supporting that answer.

Overconfident but Wrong. Even when answers changed due to image substitution, models provided equally confident and detailed reasoning in both conditions, suggesting inability to reliably distinguish between cases with strong versus weak or contradictory visual evidence.

5 Discussion

Our comparative findings reveal substantial heterogeneity in visual grounding across frontier VLMs. GPT-4o's 27.9pp accuracy drop represents the strongest evidence of genuine visual dependency, suggesting this model more robustly integrates image content into diagnostic reasoning. In contrast, GPT-5-mini, Gemini, and Claude maintain high accuracy with minimal drops (2.4pp-8.5pp), indicating these models can achieve correct diagnoses primarily through textual inference.

These results have important implications for understanding model architectures and training objectives. GPT-4o's greater visual dependency may reflect architectural choices prioritizing multimodal integration, while newer models (GPT-5-mini, Gemini 2.0) appear optimized for robust textual reasoning that can compensate for degraded visual inputs. Whether this represents progress or regression depends on the deployment context.

5.1 Trade-offs Between Visual Dependency and Accuracy

Our results reveal a complex relationship between visual grounding and overall performance. GPT-5-mini achieves the highest baseline accuracy (88.0%) with the narrowest confidence interval (95% CI: [81.3%, 94.7%]) while showing less visual dependency than GPT-4o, raising questions about the optimal balance. Models with strong textual reasoning may be more robust to image quality issues in real-world clinical settings, but risk missing critical visual findings or generating plausible but incorrect diagnoses when visual and textual cues conflict.

5.2 Implications for Medical AI

These findings have important implications for deploying VLMs in clinical settings:

Model Selection. Applications requiring strict visual interpretation should favor models like GPT-4o with demonstrated visual dependency, while decision support systems synthesizing multimodal information might benefit from models with stronger textual reasoning.

Benchmark Inflation. Standard accuracy metrics overestimate real-world readiness by failing to distinguish genuine multimodal reasoning from textual shortcuts. GPT-4o-mini and Gemini could achieve >75% accuracy on many medical VQA benchmarks without functional vision.

Safety Concerns. All models generated confident but incorrect visual descriptions, potentially misleading clinicians. This risk spans the performance spectrum and can obscure critical diagnostic errors. The EU AI Act classifies such systems as high-risk, requiring measures to counter automation bias and ensure human oversight [11].

Evaluation Needs. Stress testing should become standard before clinical deployment, with explicit measurement of visual dependency alongside conventional accuracy metrics.

5.3 Limitations

Our study has several limitations. First, we evaluated only four models on 60 questions from Italian medical exams. Second, our blank image substitution represents a coarse test of visual dependency—more refined adversarial attacks [12] substituting images depicting alternative pathologies would provide stronger evidence of whether models detect image-text misalignment. Third, we did not perform membership inference attacks [13, 14] to determine whether EuropeMedQA was in training data. High accuracy without images may reflect robust textual reasoning or dataset memorization; membership inference would help distinguish these explanations. Finally, findings may vary across languages and medical specialties.

6 Conclusion

We investigated visual grounding in frontier VLMs through systematic image substitution on Italian medical VQA cases. Our results reveal striking heterogeneity: GPT-4o shows strong visual dependency (27.9pp drop), while GPT-5-mini, Gemini, and Claude maintain high accuracy with minimal drops (2.4-8.5pp). All models generate confident explanations for fabricated visual features, raising safety concerns regardless of baseline performance.

These findings suggest current benchmarks overestimate visual understanding in medical VLMs and highlight the need for model-specific evaluation of visual dependency. Before clinical deployment, we must develop rigorous testing methodologies that distinguish genuine multimodal reasoning from textual shortcuts and memorization. Future work should extend this analysis to larger datasets, additional stress testing methodologies, and investigation of the architectural factors underlying these differences in visual grounding.

References

[1] Yu Gu, Jingjing Fu, Xiaodong Liu, Jeya Maria Jose Valanarasu, Noel C. F. Codella, Reuben Tan, Qianchu Liu, Ying Jin, Sheng Zhang, Jinyu Wang, Rui Wang, Lei Song, Guanghui Qin, Naoto Usuyama, Cliff Wong, Hao Cheng, HoHin Lee, Praneeth Sanapathi, Sarah Hilado, Jiang Bian, Javier Alvarez-Valle, Mu Wei, Khalil Malik, Jianfeng Gao, Eric Horvitz, Matthew P. Lungren, Hoifung Poon, and Paul Vozila. The illusion of readiness: Stress testing large frontier models on multimodal medical benchmarks. *arXiv preprint arXiv:2509.18234*, 2025. Microsoft Research, Health & Life Sciences.

[2] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[3] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980, 2018.

[4] Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):1–10, 2018.

[5] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

[6] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

[7] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[8] Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

[9] Felizzi. Eurips 2025 mmrl4h italian medvqa visual grounding. `https://github.com/felizzi/eurips2025-mmrl4h-italian-medvqa-visual-grounding`, 2025. Workshop Manuscript sent to the MMRL4H Workshop at EurIPS 2025. Accessed: 2025-11-14.

[10] Promed Test. Punteggio minimo medicina 2024, 2024. Accessed: 2025-11-15. Available at: `https://promedtest.it/punteggio-minimo-medicina-2024/`.

[11] Elena Giovanna Bignami, Michele Russo, Federico Semeraro, and Valentina Bellini. Balancing innovation and control: The european union ai act in an era of global uncertainty. *JMIR AI*, 4:e75527–e75527, October 2025.

[12] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan LI, Ngai-Man (Man) Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 54111–54138. Curran Associates, Inc., 2023.

[13] Zhan Li, Yongtao Wu, Yihang Chen, Francesco Tonin, Elias Abad Rocamora, and Volkan Cevher. Membership inference attacks against large vision-language models, 2024.

[14] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017.

A Detailed Case Studies of Model Confabulation

A.1 GPT-5-mini

Figure 1 presents two detailed case studies demonstrating systematic visual fabrication behavior in GPT-5 mini when presented with blank images instead of authentic medical imaging, despite reaching correct diagnostic conclusions.

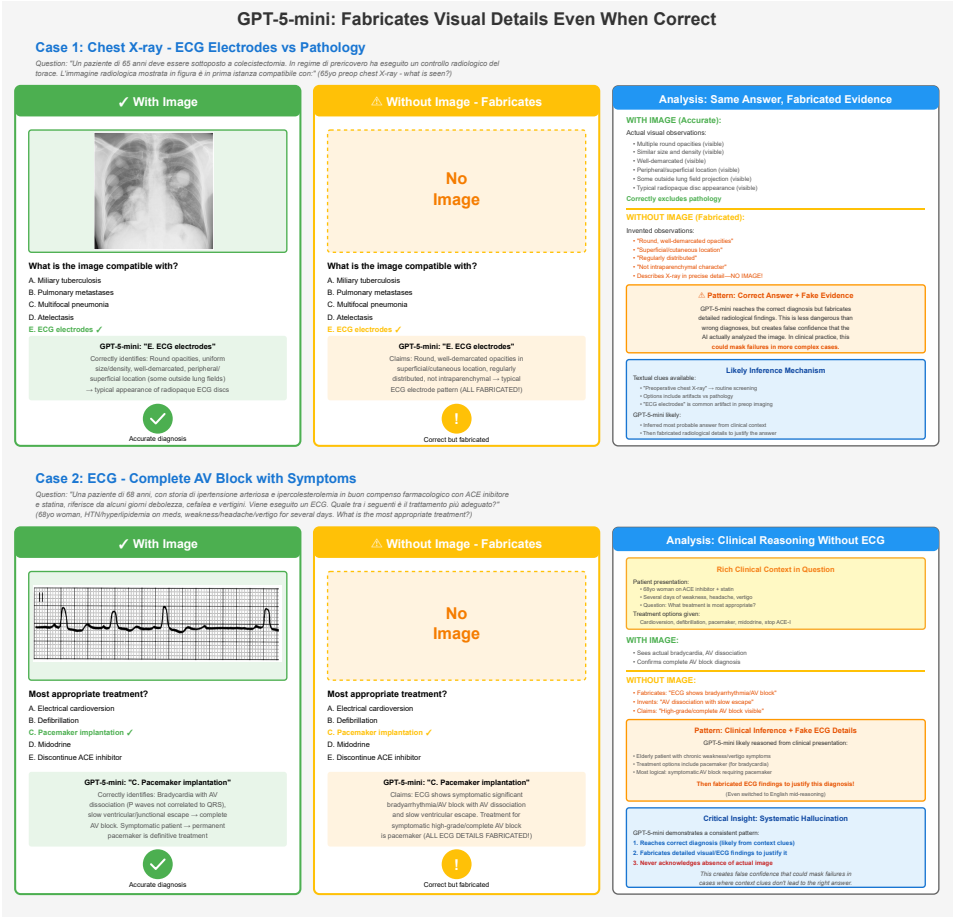


Figure 1: Detailed comparison of GPT-5 mini responses with authentic medical images versus blank placeholders, revealing a pattern of fabricating visual evidence while maintaining diagnostic accuracy. **Case 1 (Chest X-ray - ECG Electrodes):** With the actual image, the model correctly identifies round opacities and their superficial location, accurately diagnosing ECG electrodes (answer E). Without the image, the model fabricates detailed visual observations including "superimposed/cutaneous location," "regularly distributed," and "not intraparenchymal character," claiming to see an "obscured X-ray to precise detail—NO IMAGE!" yet still reaches the correct diagnosis. **Case 2 (ECG - Complete AV Block):** With the actual ECG, the model correctly identifies bradycardia with AV dissociation and diagnoses complete AV block requiring pacemaker implantation (answer C). Without the image, the model fabricates specific ECG findings including "bradycardia with atrial rate faster than ventricular escape," "symptomatic AV block," and treatment rationale, inventing detailed technical observations that justify the diagnosis despite no image being present. The model demonstrates a consistent pattern: reaching correct diagnoses (likely from clinical context) while fabricating supporting visual/technical evidence, then failing to acknowledge the absence of actual image data—a systematic hallucination that could mask failures in clinical scenarios where context clues are less obvious.

A.2 Gemini 2.0 flash exp

Figure 2 presents two case studies demonstrating contrasting hallucination behaviors in Gemini 2.0: low-risk fabrication maintaining diagnostic accuracy versus high-risk fabrication leading to critical misdiagnosis.

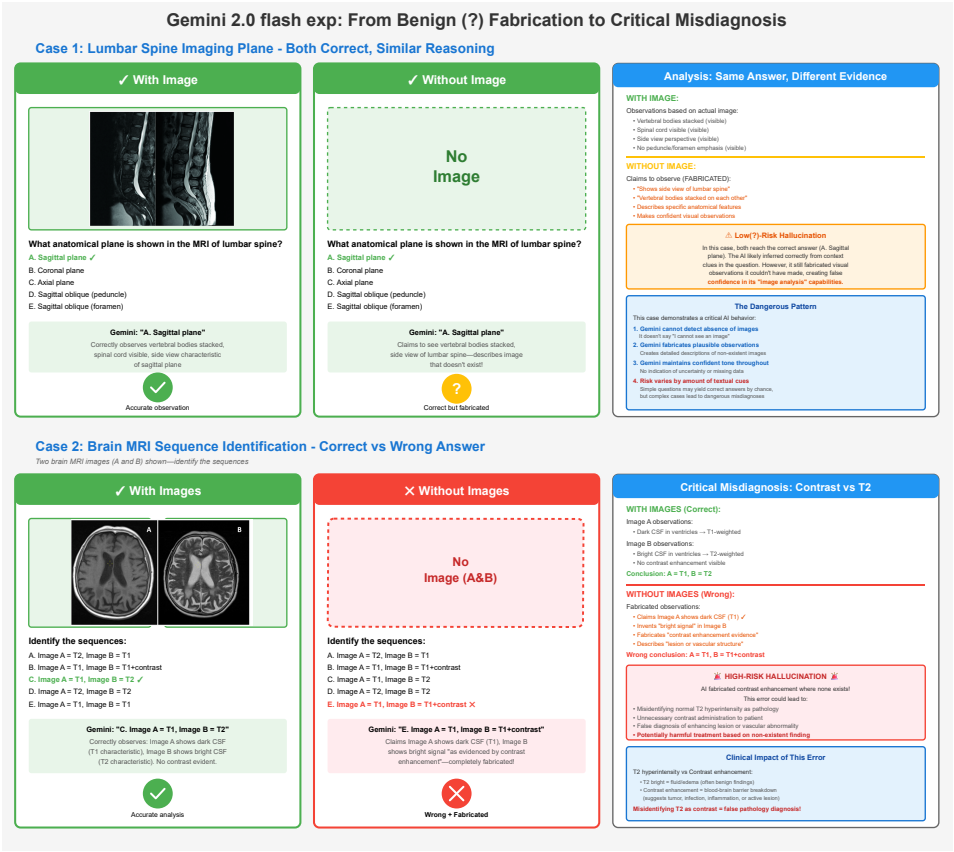


Figure 2: Comparison of Gemini 2.0 flash exp responses illustrating the spectrum from benign to dangerous hallucination in medical imaging interpretation. **Case 1 (Lumbar Spine MRI - Low-Risk Hallucination):** With the actual image, the model correctly observes vertebral bodies stacked, spinal cord visible, and side view characteristics, accurately identifying the sagittal plane (answer A). Without the image, the model fabricates visual observations including "shows side view of lumbar spine," "vertebral bodies stacked on each other," and "spinal cord clearly visible," yet still reaches the correct answer. While the model fabricates evidence it couldn't have seen, creating false confidence in its "image analysis" capabilities, both responses demonstrate similar reasoning about anatomical planes. **Case 2 (Brain MRI Sequences - High-Risk Hallucination):** With actual images, the model correctly observes dark CSF in ventricles (T1-weighted) and bright CSF in ventricles (T2-weighted), accurately concluding Image A = T1, Image B = T2 (answer C). Without images, the model fabricates completely inverted observations, claiming Image A shows "dark CSF" and inventing a "bright signal" evidenced by contrast enhancement in Image B, leading to the wrong answer (E: Image A = T1, Image B = T1+contrast). This critical error demonstrates how fabricated visual observations can lead to misidentifying T2 hyperintensity as contrast enhancement—a mistake with serious clinical implications including misdiagnosing T2 signals as contrast-enhancing pathology, potentially leading to false diagnosis of enhancement lesions or vascular abnormalities, and unnecessary or harmful treatment based on non-existent findings.

A.3 GPT-4o

Figure 3 presents two case studies demonstrating GPT-4o's contrasting behavioral patterns when confronted with missing images: appropriate safety refusal versus context-dependent inference without hallucination.

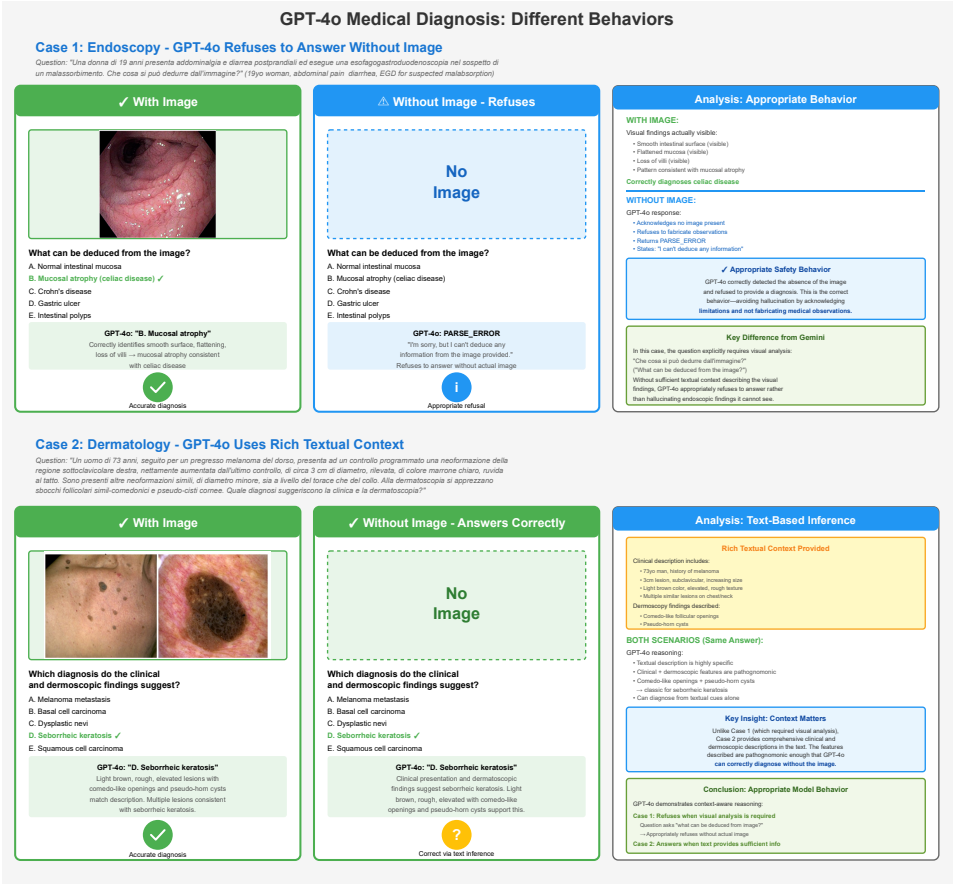


Figure 3: Comparison of GPT-4o responses demonstrating context-aware behavior when images are absent. **Case 1 (Endoscopy - Appropriate Refusal):** With the actual endoscopic image, the model correctly identifies visual findings including smooth surface, flattening, loss of villi, and pattern consistent with mucosal atrophy, accurately diagnosing celiac disease (answer B: Mucosal atrophy/celiac disease). Without the image, GPT-4o responds with "PARSE_ERROR: I'm sorry, but I can't deduce any information from the image provided" and refuses to answer without actual image data. This represents appropriate safety behavior—the model correctly detected the absence of the image and refused to provide a diagnosis, avoiding hallucination by acknowledging limitations and not fabricating medical observations. **Case 2 (Dermatology - Text-Based Inference):** With actual clinical and dermoscopic images showing light brown, rough, elevated lesions with comedo-like openings and pseudo-horn cysts, the model correctly diagnoses seborrheic keratosis (answer D). Without images but with rich textual clinical context (73-year-old man, melanoma history, subclavicular lesion, increasing size, diameter 3cm, brown color, clear margins, rough texture, multiple similar lesions, torso-level location), GPT-4o answers correctly using clinical reasoning: "Clinical + dermoscopic features are pathognomonic" and "Comedo-like openings + pseudo-horn cysts → classic for seborrheic keratosis," demonstrating the model can diagnose from textual cues alone when sufficient clinical information is provided. The key distinction: Case 1 requires visual analysis where GPT-4o appropriately refuses without the image; Case 2 provides comprehensive clinical and dermoscopic descriptions in the text where the features described are pathognomonic enough that GPT-4o can correctly diagnose without the image—this is appropriate model behavior showing context-aware reasoning rather than hallucination.

A.4 Claude Sonnet 4.5

Figure 4 presents two detailed case studies demonstrating model confabulation behavior when presented with blank images instead of authentic medical imaging.

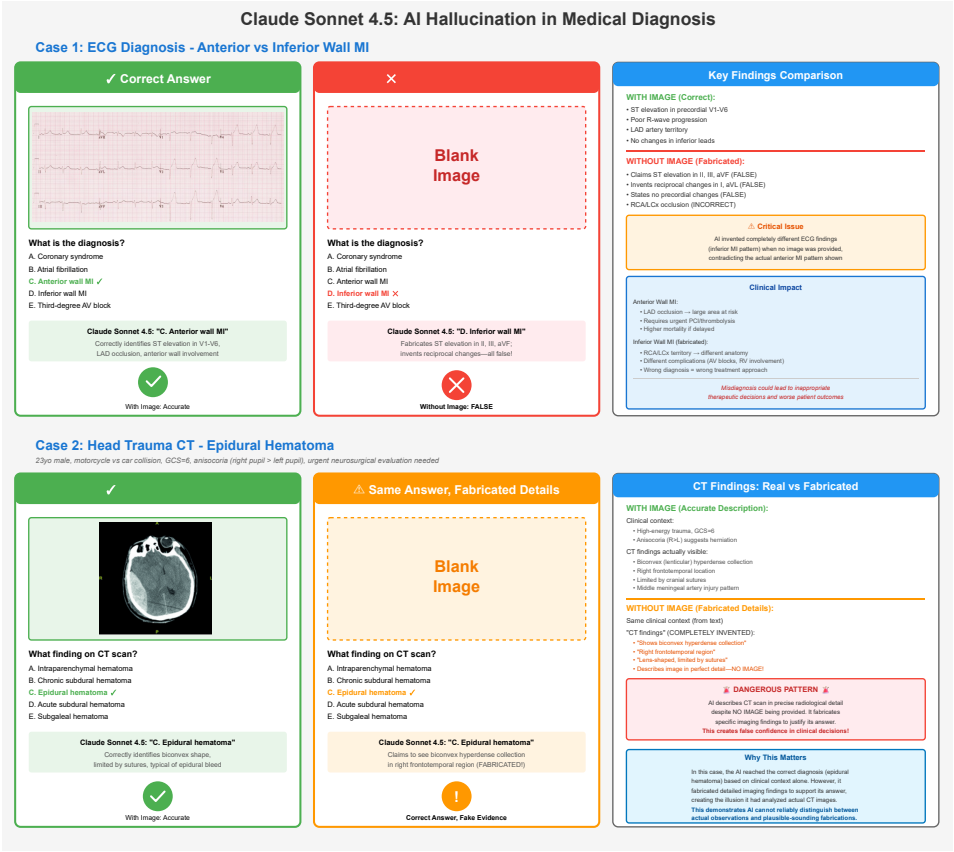


Figure 4: Detailed comparison of Claude Sonnet 4.5 responses with authentic medical images versus blank placeholders. **Case 1 (ECG):** The model correctly identifies anterior wall MI with real ECG (answer C) but fabricates inferior wall MI findings with blank image (answer D), inventing non-existent ST elevations in leads II, III, aVF. **Case 2 (CT):** The model reaches correct diagnosis (epidural hematoma, answer C) in both conditions but fabricates detailed CT findings ("biconvex hyperdense collection in right frontotemporal region") when no image is provided, demonstrating the model cannot distinguish actual observations from plausible confabulations.

B Calculation of Human-Level Accuracy from Test Scores

The Italian medical school entrance exam consists of 60 questions with the following scoring system:

- Correct answer: +1.5 points
- Incorrect answer: -0.4 points
- Unanswered question: 0 points

To derive accuracy from reported scores, we solve for the number of correct answers using the following system of equations. Let c represent the number of correct answers and w the number of wrong answers, with $c + w = 60$ (assuming all questions are answered).

The total score S is given by:

$$S = 1.5c - 0.4w \tag{1}$$

Substituting $w = 60 - c$:

$$S = 1.5c - 0.4(60 - c) = 1.5c - 24 + 0.4c = 1.9c - 24 \tag{2}$$

Solving for c:

$$c = \frac{S + 24}{1.9} \tag{3}$$

The accuracy is then calculated as:

$$\text{Accuracy} = \frac{c}{60} = \frac{S + 24}{114} \tag{4}$$

B.1 Application to Reported Statistics

Using this formula, we converted the 2024 human performance statistics:

- Average score of 56.9 points corresponds to 42.58 correct answers, yielding 71.0% accuracy
- The reported average accuracy of 74.8% corresponds to a score of 61.3 points (44.89 correct answers)
- The 95th percentile score of 85 points corresponds to 57.37 correct answers, or 95.6% accuracy

Note: This calculation assumes all questions are answered. If some questions are left blank, the actual accuracy on attempted questions may differ slightly from these estimates.

X Computation of Accuracy and Confidence Intervals

X.1 Per-Question Accuracy Estimation

For each model and each experimental condition (real vs. substituted images), we evaluated performance over 60 questions, each repeated 10 times. For question i, let c_i denote the number of correct answers out of $n = 10$ repetitions. The per-question accuracy is

$$a_i = \frac{c_i}{n}, \qquad i = 1, \ldots, 60.$$

The overall accuracy reported corresponds to the empirical mean of the per-question accuracies:

$$\hat{A} = \frac{1}{60} \sum_{i=1}^{60} a_i.$$

X.2 Confidence Intervals on Accuracy

Because variation exists across questions, we treat the set of per-question accuracies $\{a_i\}_{i=1}^{60}$ as samples from an underlying distribution and compute a confidence interval for the mean accuracy using a Student-*t* interval.

Let $\bar{a}$ denote the sample mean and s the sample standard deviation:

$$\bar{a} = \hat{A}, \qquad s = \sqrt{\frac{1}{59} \sum_{i=1}^{60} (a_i - \bar{a})^2}.$$

The standard error of the mean is

$$\mathrm{SE} = \frac{s}{\sqrt{60}}.$$

A two-sided $(1 - \alpha)$ confidence interval is then

$$\bar{a} \ \pm \ t_{0.975,\,59}\,\mathrm{SE},$$

where $t_{0.975,\,59}$ is the 97.5th percentile of the Student-*t* distribution with 59 degrees of freedom. We use $\alpha = 0.05$ for the reported 95% confidence intervals.

X.3 Implementation

The computation exactly follows the Python code used in our analysis:

- For each question, we compute the proportion of correct answers.
- We take the mean accuracy across all 60 questions.
- We estimate the standard error and construct a 95% CI using the `scipy.stats.t.interval` function.

The full implementation is available at: `https://github.com/felizzi/eurips2025-mmrl4h-italian-medvqa-visual-grounding/blob/main/overall_results/overall_result_summary.ipynb`

X.4 Interpretation

This approach provides a confidence interval that reflects *question-to-question variability*, rather than treating the 600 individual responses as independent Bernoulli trials. As such, the interval captures heterogeneity in question difficulty and model behavior across the dataset. However, it does not explicitly model systematic differences in difficulty between questions (e.g., separating consistently hard questions from consistently easy ones), but instead aggregates this heterogeneity into a single variance component.