# A Radiomics Perspective on Gastric Lesion Characterization in Endoscopy Images


Nuno D. Teixeira
Master in Data Science
Department of Computer Science
Faculty of Sciences of University of Porto
2025

# A Radiomics Perspective on Gastric Lesion Characterization in Endoscopy Images

Nuno D. Teixeira

Dissertation carried out as part of the Master in Data Science

Department of Computer Science

2025


**Supervisor**
Miguel Coimbra, Full Professor, FCUP

# Acknowledgements

This work was supported by European funds under the European Union's Recovery and Resilience Facility (reference 2024.07584.IACDC/2024), by the Horizon Europe Framework Programme (Grant Agreement No. 101095359), and by UK Research and Innovation (Grant Agreement No. 10058099).

I am deeply grateful to my supervisor Miguel Coimbra for guidance, availability, and scientific rigor throughout this work. I also thank everyone in the CAGE research group for their collaboration, technical discussions and daily support. Finally, I extend my heartfelt thanks to my family and friends for their encouragement, patience, and understanding at every stage of this journey.

# Resumo

Esta tese investiga se a radiómica pode fornecer biomarcadores transparentes e clinicamente relevantes para a caracterização de lesões gástricas em EGD. Foram analisados dois conjuntos de dados de um único centro (IPO-Porto): (A) 1280 imagens NBI anotadas com EGGIM e (B) 199 imagens de WLE com segmentação ao nível do pixel e grau de displasia validado por biópsia. A partir das regiões de interesse das lesões, foram extraídos 104 descritores de forma e textura alinhados com o IBSI, além de características de cor baseadas em histogramas HSV. A associação das características foi medida através de uma combinação do $P_k$ e de um SVM sob validação cruzada estratificada repetida. Adicionalmente, para estudar o compromisso entre interpretabilidade e eficácia, a radiómica foi comparada com radiómica comprimida por PCA e com embeddings de aprendizagem profunda obtidos de uma ResNet-18 pré-treinada com ImageNet.

Os resultados revelam assinaturas específicas da tarefa e da modalidade. Para o EGGIM, as famílias de textura apresentam as maiores associações univariadas, enquanto a cor tem importância limitada. Para a displasia, alguns valores do histograma HSV foram altamente informativos, refletindo alterações cromáticas consistentes com mudanças mucosas e vasculares conhecidas. Os conjuntos de características no top-100 entre tarefas exibiram baixa sobreposição (Jaccard 0,24), salientando que os sinais preditivos dependem da interação entre patologia e modalidade. Embora as características radiómicas fornecessem descritores individuais fortes, os embeddings da ResNet-18 mostraram mais características moderadamente informativas e proporcionaram o melhor desempenho multivariado. As representações via PCA mantiveram desempenho competitivo, mas geralmente inferior.

Este trabalho disponibiliza uma abordagem reprodutível de radiómica para EGD e explora adicionalmente como os descritores manuais se comparam com representações de aprendizagem profunda mais opacas. Apesar do reduzido tamanho amostral e do desequilíbrio de classes, os resultados sustentam a radiómica como uma base compacta e explicável, complementar às arquiteturas de aprendizagem profunda já estabelecidas no domínio endoscópico.

Palavras-chave: Visão Computacional, Radiómica, Esofagogastroduodenoscopia, Metaplasia Gastrointestinal, EGGIM, Displasia

# Abstract

This thesis investigates if radiomics can provide transparent, clinically meaningful biomarkers for gastric lesion characterization in EGD. Two single-center datasets from IPO-Porto were analyzed: (A) 1280 NBI frames annotated with EGGIM, and (B) 199 WLE images with pixel-level segmentation and dysplasia grade validated with biopsy. From the lesion ROIs, 104 IBSI aligned shape and texture descriptors in addition to histogram based HSV color features were extracted. The association of the features was measured using a combination of $P_k$ and an off-the-self SVM under repeated, stratified 10-fold cross-validation. Furthermore, to study the interpretability-effectiveness trade-off, radiomics was benchmarked against PCA compressed radiomics and DL embeddings from a frozen ImageNet-pretrained ResNet-18.

Results show task and modality specific signatures. For EGGIM, texture families have the top univariate associations, while color carried limited significance. For dysplasia, some HSV bins were highly informative, reflecting chromatic shifts consistent with known mucosal and vascular changes. The top-100 ranked feature sets across tasks exhibited low overlap (Jaccard 0.24), highlighting that predictive signals depend on the relation between pathology and modality. Although radiomics provided strong individual descriptors, ResNet-18 embeddings showed more moderately informative features and yielded the best multivariate performance. The PCA representations retained competitive but generally inferior performance.

This work delivers a reproducible radiomics pipeline for EGD, while further exploring how the handcrafted descriptors compare against more opaque DL representations. Despite the small, class-unbalanced dataset, findings support radiomics as a compact, explainable baseline that can be used as a complementary approach to the DL architectures already established in the endoscopic domain.

Keywords: Computer Vision, Radiomics, EGD, GIM, EGGIM, Dysplasia

# Table of Contents

# List of Tables

# List of Figures

# List of Abbreviations

$P_k$ Smith's Prediction Probability. ii, iii, vii, 20, 27, 29– 31, 33– 36, 38– 44, 46

**AI** Artificial Intelligence. 1, 2, 12, 13, 23– 25, 35

**BLI** Blue-light imaging. 8

**CNN** Convolutional Neural Network. vii, 12– 17, 23

**CT** Computed Tomography. 1, 6, 23, 24

**DL** Deep Learning. iii, 12, 13, 17, 18, 22– 24, 36, 38– 40, 42, 44, 46, 47

**EGD** Endoscopic Gastroduodenoscopy. ii, iii, 1– 3, 5– 8, 12, 13, 17, 19, 23, 25, 35, 42, 44, 46, 47

**EGGIM** Endoscopic Grading of Gastric Intestinal Metaplasia. ii– vii, 3, 10, 11, 13, 20, 25, 26, 29, 33, 36, 38– 41, 46, 47

**EUS** Endoscopic Ultrasound. 6, 23

**GC** Gastric Cancer. iv, 4, 5, 7, 10, 13, 23– 25, 35

**GI** Gastrointestinal. 3, 7, 12, 13, 23, 24, 35, 46, 47

**GIM** Gastric Intestinal Metaplasia. iii, 1, 2, 9– 11, 25, 33– 35, 46

**GLCM** Gray-Level Co-occurrence Matrix. 19, 28, 30, 31, 33

**GLDM** Gray-Level Dependence Matrix. 19, 28, 30, 31, 33, 34

**GLRLM** Gray-Level Run-Length Matrix. 19, 28, 30, 31, 33

**GLSZM** Gray-Level Size Zone Matrix. 19, 28, 31

**HSV** Hue–Saturation–Value. ii, iii, 28, 30, 33, 39, 46, 47

**IBSI** Image Biomarker Standardization Initiative. iii, 18, 19, 46

**IEE** Image-enhanced Endoscopy. 8, 9

**LCI** Linked Color Imaging. 8

**MRI** Magnetic Resonance Imaging. 1

**NBI** Narrow-band Imaging. ii, iii, vii, 8– 10, 23, 25, 33, 36, 42, 46, 47

**NGTDM** Neighborhood Gray-Tone Difference Matrix. 19, 28, 31, 34

**OvR** One-vs-Rest. 23, 29, 39

**P-R-E** Proportional-Reduction-in-Error. 20, 21

**PCA** Principal Component Analysis. ii, iii, vii, 35– 38, 41, 43, 44, 46

**ROI** Region of Interest. 17, 18, 26, 27, 36, 38

**SVM** Support Vector Machine. ii, iii, 20, 22, 27, 29, 33, 35, 36, 39, 41, 42, 44, 46

**WLE** White-light Endoscopy. ii, iii, vii, 8– 10, 23, 26, 33, 36, 42, 46, 47

# 1. Introduction

## 1.1. Motivation

According to the most recent statistics, gastric cancer is the fifth most frequently diagnosed cancer and the fifth leading cause of cancer-related mortality worldwide [1]. Despite advances in diagnosis, the prognosis remains poor, mainly due to late-stage detection. Therefore, the development of early diagnosis and risk stratification is of the utmost importance.

Endoscopic Gastroduodenoscopy (EGD) plays a major role in the management of this disease by allowing direct visualization of the gastric mucosa enabling the detection and localization of the presence of Gastric Intestinal Metaplasia (GIM), a precancerous condition that is an initial step in the cascade of events leading to dysplasia and subsequent intestinal-type gastric adenocarcinoma. Dysplasia itself is an important marker for risk stratification, as its severity directly influences surveillance strategies. Current practice relies heavily on visual evaluation by expert endoscopists followed by histopathological confirmation through biopsy, which, while effective, is invasive, subject to sampling error, and dependent on inter-observer expertise.

Recent years have witnessed a surge in the application of Artificial Intelligence (AI) to medical imaging, especially within gastroenterology. Computer vision techniques have the potential to assist clinicians during endoscopic procedures by highlighting suspicious areas and suggesting possible pathological conditions. Radiomics is a particular interesting field of study that extracts quantitative descriptors from medical images that can act as quantitative image biomarkers, providing surrogates of clinical information extracted from gastric biopsies. Although unexplored for EGD imaging, radiomics has already shown promise in other imaging modalities, such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), where it is applied to tumor characterization, treatment response prediction, and outcome prognosis. Examples include studies where radiomic features from CT scans can differentiate between benign and malignant lung nodules with high accuracy, aiding in early cancer detection and personalized treatment planning [2], or applied to breast cancer imaging to assess tumor heterogeneity and predict the likelihood of metastasis, thereby guiding treatment decisions [3]. Importantly, radiomics

does not require additional contrast agents or new imaging modalities, instead leveraging conventional clinical images, enriching them with quantitative biomarkers that often correlate with biological processes invisible to the naked eye.

Unlike radiological modalities, endoscopic imaging presents unique characteristics, such as high variability in lighting, texture, and color. These factors complicate feature extraction but also provide a wide variety of quantitative descriptors if carefully harnessed. Furthermore, while deep learning approaches have become prominent in recent endoscopic AI research, they are often criticized for being opaque and difficult to interpret. Radiomics, by contrast, offers explainable features that can be biologically and visually linked to the appearance of lesions. This interpretability is of great clinical value, as it fosters trust and facilitates integration into medical decision-making.

The motivation behind this work lies in exploring whether radiomics can provide reliable, interpretable, and biologically meaningful biomarkers from standard EGD images, aiming to contribute to the development of a virtual biopsy paradigm. Such an approach would serve as a non-invasive intermediate step between traditional visual inspection and histological biopsy, potentially reducing patient burden while enhancing diagnostic accuracy.

## 1.2. Objectives

The main goal of this thesis is to research the potential of radiomics as a methodology for the characterization of gastric lesions in endoscopy images. The intent is to develop a testing framework to evaluate whether radiomic features can provide interpretable and biologically meaningful biomarkers to support clinical decision-making in gastroenterology. Beyond this, the thesis also aims to examine how the opacity of features influences their predictive performance by comparing purely interpretable radiomic features against more abstract representations.

## 1.3. Contributions

This thesis represents one of the first systematic attempts to apply radiomics to EGD imaging, a modality where quantitative image analysis remains largely unexplored. Focusing on GIM and dysplasia, this work introduces a novel perspective on how endoscopic data can be transformed into measurable and clinically interpretable biomarkers. Beyond this, the work provides a transparency-oriented evaluation framework for EGD, comparing feature effectiveness across diminishing levels of interpretability. This clarifies how radiomic descriptors perform relative to more complex, image-level representations.

## 1.4. Thesis Structure

This thesis structure encapsulates the work done in two papers (one already approved and presented in ENBENG 2025 and the other waiting for submission) with the added context necessary to fully understand the findings. The document is organized as follows: **Introduction** (chapter 1), which outlines the motivation and contextual background of gastric cancer and its clinical management; **Background** (chapter 2) provides the necessary theoretical foundation, describing the biological basis and current clinical management of gastric cancer, the role of endoscopic imaging and the Endoscopic Grading of Gastric Intestinal Metaplasia (EGGIM) protocol; **Computer Vision in Endoscopy** (chapter 3) describes the principles of artificial intelligence approaches in medical imaging, with a particular focus on deep learning and radiomics; **A Radiomics-Based Approach to Gastric Imaging Biomarker Extraction in Upper GI Endoscopy** (chapter 4) presents an analysis framework to validate the usefulness of radiomics in upper Gastrointestinal (GI); **Exploratory Analysis on Feature Effectiveness Across Transparency Levels in Upper GI Endoscopy** (chapter 5) explores how the interpretability of descriptors extracted from EGD images influences their effectiveness; **Conclusion** (chapter 6) summarizes the contributions of this work, reflects on its limitations, and discusses future directions for research in the application of radiomics and artificial intelligence to gastrointestinal imaging.

# 2. Background

## 2.1. Gastric Cancer

Gastric Cancer (GC) is a malignant neoplasm of the stomach and remains a major global health issue, ranking among the leading causes of cancer death worldwide. The latest GLOBOCAN 2022 highlights its recent incidence and mortality, particularly in East Asia and Europe [1]. Clinically, it can be distinguished between cardia and non-cardia GC, depending on the anatomical site. Cardia GC develops in the proximal stomach or near the esophagogastric junction and is often linked to obesity and gastroesophageal reflux disease, while non-cardia GC arises in the remaining area of the stomach and is strongly associated with *Helicobacter pylori* infection and the Correa cascade [4].

The Correa cascade (Figure 2.1) describes the stepwise histopathologic pathway by which most non-cardia, intestinal-type gastric adenocarcinomas develop. It typically begins with chronic active gastritis, which progresses to multifocal atrophic gastritis. As glandular loss advances, the mucosa undergoes intestinal metaplasia, followed by intraepithelial neoplasia (dysplasia) and finally invasive carcinoma [4].

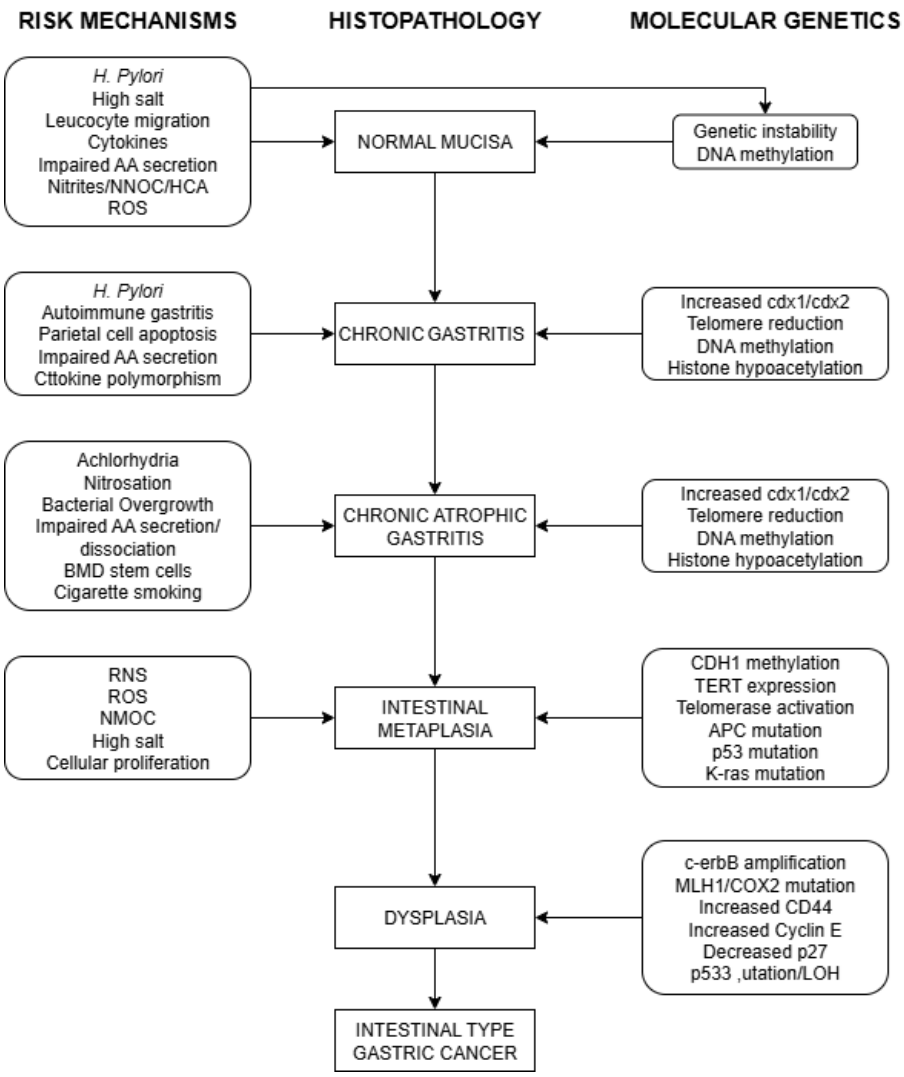


Figure 2.1: Modern Correa pathway of intestinal-type gastric carcinogenesis, with risk factors and host mechanisms in green, molecular genetics in blue, and preventive factors in yellow (adapted from [4]).

This evolution usually unfolds over years to decades and is modulated by host genetics, bacterial virulence factors, and environmental exposures (e.g. high-salt diets, smoking). Eradication of *H. pylori* can halt or partially reverse the earliest steps and lowers cancer risk, but once intestinal metaplasia or dysplasia is established, residual risk persists, which is why endoscopic surveillance and systematic biopsy mapping are recommended [5].

Modern care is a multidisciplinary process that follows a pathway of prevention, detection, staging and treatment (Figure 2.2). Primary prevention emphasizes the identification and eradication of *Helicobacter pylori*, ideally before advanced atrophy/metaplasia develops, reducing the risk of cancer [6]. Furthermore, for high incident countries such as Japan and South Korea, population screening endoscopy is recommended, with organized programs that use periodic upper EGD reducing GC mortality [7].

Similarly, diagnosis begins with high definition EGD to document macroscopic features (e.g. location, size, ulceration) and to obtain tissue for histopathology and predictive biomarkers. EGD is a flexible-scope procedure that directly inspects the esophagus, stomach and duodenum, enabling visual diagnosis, targeted biopsies and therapeutic intervention. A biopsy is a small mucosal tissue sample, taken through the endoscope and then processed for histology to confirm cancer and assess biomarkers. When the background mucosa shows chronic atrophic gastritis or intestinal metaplasia, systematic biopsy mapping may be performed to stage risk, particularly when surveillance decisions are relevant. During work-up of lesions that plausibly meet criteria for curative endoscopic resection, unnecessary repeated biopsies at the planned resection site are avoided to reduce fibrosis and preserve technical curability [8] [9].

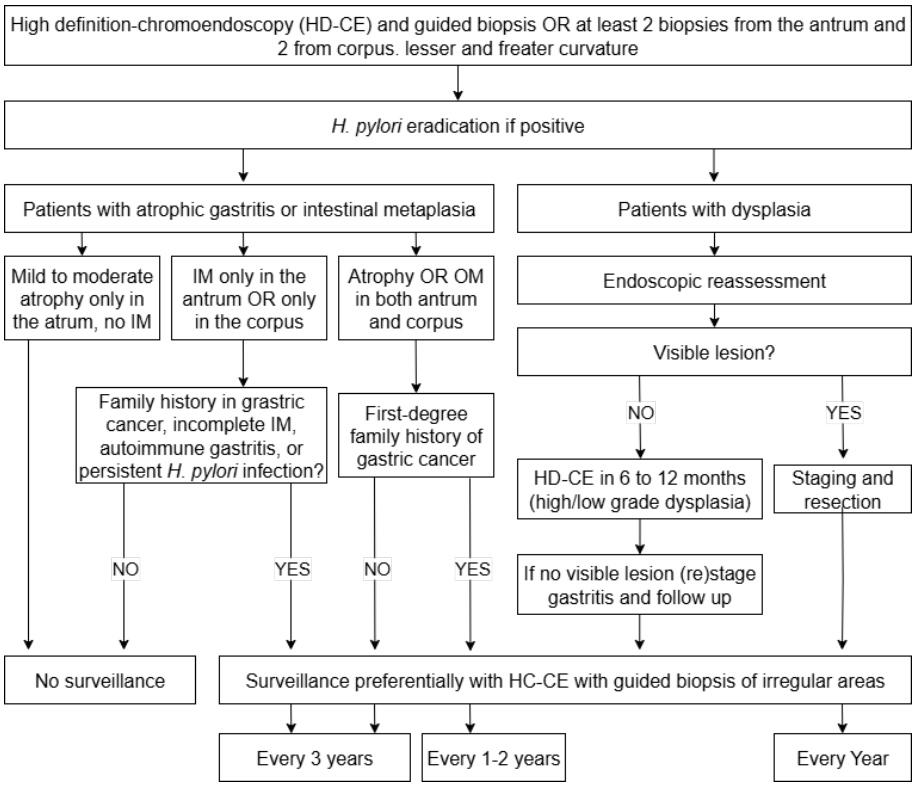


Figure 2.2: Proposed management for patients with atrophic gastritis, gastric intestinal metaplasia, or gastric epithelial dysplasia (adapted from [9]).

Contrast-enhanced CT of the chest, abdomen and pelvis is the backbone of clinical staging. Modern protocols survey the whole torso in one exam, allowing the detection of the primary lesion, estimating local extension, assessing lymph nodes and identifying peritoneal spread. Since CT can't reliably resolve the stomach's wall layers, Endoscopic Ultrasound (EUS) is often used to distinguish between superficial and deep disease [10].

Treatment diverges according to stage and biology. For strictly defined early GC, curative endoscopic resection, preferably endoscopic submucosal dissection, is indicated when the predicted risk of lymph-node metastasis is negligible [11] [12]. For resectable, locally advanced disease, Western guidelines favor perioperative chemotherapy followed by gastrectomy with a D2 lymphadenectomy in experienced centers [13]. In unresectable or metastatic disease, systemic therapy is increasingly biomarker-guided, with multiple chemotherapy based practices being applied depending on the case [8].

### 2.1.1 Endoscopy Imaging

As stated above, EGD is a flexible endoscope examination of the esophagus, stomach, and duodenum performed under topical anesthesia or sedation. The exam requires the patients to be fasted with some services providing a pre-endoscopy drink that improves mucosa visibility. With the patient in the lateral position the endoscope is inserted and the endoscopist proceeds slowly and systematically through the upper GI tract to avoid blind spots. The number of images taken during the exam varies, but the consensus minimum landmark photo-documentation is 8 to 10 images (Figure 2.3) [14]. When applicable, a systematic stomach screening protocol can be performed, which captures around 22 images from defined quadrants to reduce miss rates [15]. International quality frameworks such as RSGE [16] emphasize the need for systematic inspection, mucosa cleanliness, adequate distension, and structured photo-documentation to maximize the effectiveness of the exam, improving detection and facilitating the following procedures.

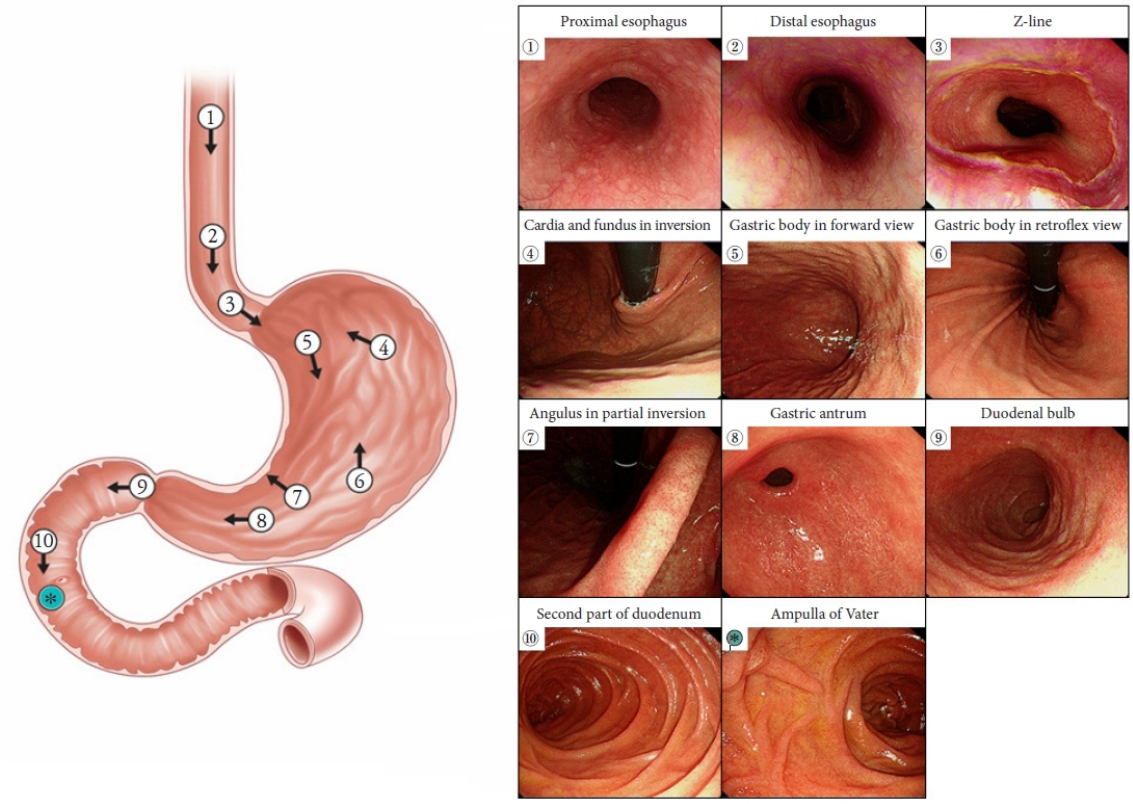


Figure 2.3: Photodocumentation example in a typical endoscopic examination [17].

Modern EGD uses high-definition White-light Endoscopy (WLE) as the baseline [16]. Nevertheless, lesion detection and characterization can be strengthened by Image-enhanced Endoscopy (IEE). Optical laser-based systems such as Narrow-band Imaging (NBI) and Blue-light imaging (BLI) accentuate hemoglobin content to highlight microvascular and microsurface patterns. In addition, digital post-processing systems i-scan/FICE and Linked Color Imaging (LCI) often improves visibility and detection of metaplasia by, respectively, enhancing the sharpness/contrast and color differences of the images [18]. Texture and color enhancement imaging, red dichromatic imaging and other newer modalities further boost subtle surface contrast and deeper vessel visualization [19]. Typically, WLE provides the initial survey and documentation of landmarks. LCI is then commonly used for detection by amplifying the subtle color contrast and helping small lesions stand out against surrounding tissue, while NBI and BLI assist in margin assessment and characterization. When borders remain ambiguous endoscopists can resort to dye-based chromo-endoscopy, where indigo carmine and acetic acid can accentuate the topology and sharpen lesion edges [20].

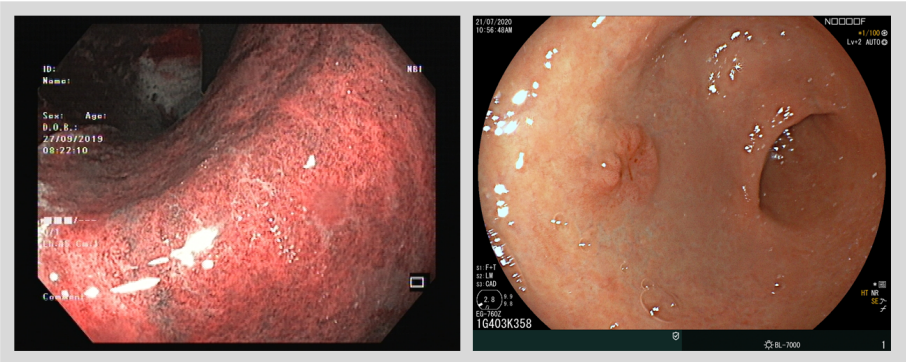


Figure 2.4: Example of image modalities. (A) NBI image collected using an OLYMPUS CV-190 endoscope from the Instituto Português de Oncologia do Porto (IPO-Porto), Portugal; (B) WLE image collected using a FUJIFILM VP-7000 en-doscope also from IPO-Porto

Without rigid standardization of preparation, mapping, and IEE settings, documentation over time becomes a difficult task, and apparent changes may reflect technique rather than biology, making lesion evaluation ambiguous. With these aspects safeguarded, modern imaging resources are increasingly capable of recognizing how the mucosa phenotype evolves along the Correa cascade. The Kyoto classification of gastritis identifies endoscopic features that correlate with the development of *H. pylori* and help classify risk of disease [21]. Aspects such as diffuse redness, enlarged folds, and nodularity are typical signs of infection, while atrophy appears as a pale mucosa with protruding subepithelial vessels.

In WLE, GIM often appears as small, gray-white, slightly elevated plaques on a background of mixed patchy pink and pale mucosa, creating an irregular surface. As glands elongate, the surface can show finger-like patterns, which contrast with the adjacent mucosa [22]. These texture differences can be further magnified with IEE and increase the confidence for targeted biopsies. On NBI, GIM can be further recognized by a characteristic microsurface tubulovillous pattern [23].

Moving forward in the Correa pathway, dysplasia is typically characterized by a color mismatch, usually erythematous or paler than the surrounding mucosa, and loss of surface luster. Moreover, subtle surface irregularities and differences in topography, such as slightly depressed or elevated plaques, small erosions and converging cut gastric folds, can indicate a local lesion (Figure 2.5) [22].

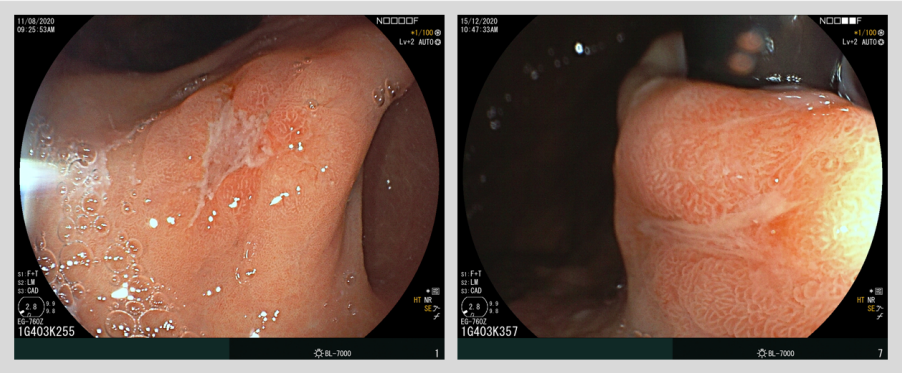


Figure 2.5: Example of WLE images collected using a FUJIFILM VP-7000 endoscope from IPO-Porto containing different levels of dysplasia. (A) low-grade dysplasia; (B) high-grade dysplasia.

### 2.1.2 Endoscopic Grading of Gastric Intestinal Metaplasia (EGGIM) Protocol

The EGGIM Protocol estimates the extent of GIM across five predefined gastric areas: the lesser/greater curvature of the antrum and the corpus, and the incisura. In each area, an expert assigns a value of 0 (no GIM), 1 ($\leq$30% of the area with GIM) or 2 (>30% of the area with GIM), resulting in a total score of 0-10 (Table 2.1) [24].

Table 2.1: Endoscopic grading of EGGIM score: total score varies from 0 (normal endoscopy with no areas suggestive of IM) to 10 (extensive metaplasia in all gastric areas)

| | **Antrum** | | **Incisura** | **Corpus** | |
|---|---|---|---|---|---|
| | *Lesser curvature* | *Greater curvature* | | *Lesser curvature* | *Greater curvature* |
| No GIM | 0 | 0 | 0 | 0 | 0 |
| $\leq$30% GIM | 1 | 1 | 1 | 1 | 1 |
| $>$30% GIM | 2 | 2 | 2 | 2 | 2 |
| **GIM score** | 0–4 | | 0–2 | 0–4 | |

In a prospective study with 250 patients from Italy and Portugal, EGGIM showed promising diagnostic performance for detecting extensive GIM. Using a cut-off of EGGIM > 4 the methodology achieved a sensitivity of 89%, specificity of 95%, and a positive likelihood ratio of 16.5, supporting its use to identify patients who require surveillance for GC risk. Notably, almost all GIM cases captured with targeted biopsies were already suspected endoscopically, with only 3 patients, around 1%, requiring random biopsies for a benign alteration in the gastric mucosa. The main limitation of this approach is that the presence of foveolar hyperplasia can mimic the tubulovillous pattern of GIM in NBI, so

areas without histological GIM can be overestimated. This is increasingly relevant since operator experience can likely influence performance of this scoring system. Nevertheless, the results suggest that the application of EGGIM can potentially reduce the need for random biopsies, reserving histology for targeted lesions, when GIM is seen confidently endoscopically [25].

# 3. Computer Vision in Endoscopy

Computer vision has been part of EGD for decades. Early systems relied on engineered descriptors paired with conventional machine learning classifiers to flag suspicious regions. These hand-crafted pipelines established that quantitative patterns in endoscopy images carry histological meaning that, automated, could standardize what experts see. This ”features-first” idea is the root of radiomics, developed and validated earlier in radiology with the intention to systematically extract image descriptors that serve as candidate biomarkers [2].

A major shift came with the proliferation of Deep Learning (DL). In 2018, Hirasawa et al. trained a single-shot detector on 13000 endoscopy images and reported a sensitivity of 90% for gastric cancer across 2000 test images [26]. This early proof that Convolutional Neural Network (CNN) can generalize to heterogeneous EGD data made it possible to extrapolate the approach to real-time video [27]. Since then, thanks to constant advancements in AI, multiple applications have documented rapid progress in a variety of tasks, such as classification and differentiation of gastric cancer, delineation of margins, and detection of high-risk regions [28].

As AI improves and evidence of its usefulness increases, regulatory authorities are beginning to implement its routine use. Although not in the upper GI, the USA was the first to do so, with the FDA approving multiple applications for colonoscopy, starting in 2021 with Medtronic/Cosmo’s GI Genius [29] and most recently MAGENTIQ-COLO in January 2025 [30]. Europe followed suit in 2024, with the cloud-based AI models CADDIE [31], CADU, and SMARTIBD being clinically deployed along with Odin Vision vendor platforms. For EGD, the most advanced regulatory activity is in Japan where the MHLW/PMDA approved gastroAI-model G for gastric lesion differentiation in December 2024.

Despite their remarkable predictive prowess DL methods are limited by their lack interpretability and heavy training requirements. The ”black-box” nature of neural networks means decision-making is often opaque, with little to no visibility on how the image features influence model prediction. This poses a barrier for both clinical adoption and scientific understanding, since clinicians are less likely to trust results that can’t be meaningfully interpreted and validated against pathological knowledge.

This sets up an additional pathway that can be integrated in parallel with Deep Learning: Explainable AI. DL clearly dominates endoscopic AI, already showing clinical utility in pilot deployments due to its ability to perform tasks efficiently and with high levels of accuracy. Explainable AI, central to this thesis, brings a complementary and interpretable approach by turning images into explicit and readable features that clinicians can analyze and link to histology and visual phenotypes, such as the patterns used in EGGIM. Importantly, instead of being viewed as competing paradigms, both DL and explainable AI can be used in tandem to provide more transparent models that still reach high levels of performance.

## 3.1. Deep Learning

DL is now the most dominant approach for automated image and video analysis, due to its ability to learn directly from pixel intensities rather than from engineered descriptors. This property is especially valuable in upper GI endoscopy, where the images can vary according to device, operator, optics, and mucosal condition, but clinically relevant patterns remain. In EGD, DL is typically used for image-level diagnosis (classification), lesion localization (object detection), margin delineation (segmentation), temporal modeling over video, and even auxiliary tasks such as depth estimation. In classification, the model assigns one or more labels to still frames or short video segments, with typical targets including *H. pylori* status, metaplasia/dysplasia, early GC, or risk stratification. Detection models extend classification by localizing suspicious regions with bounding boxes, enabling real-time prompts during live procedures. Segmentation systems produce pixel-accurate masks of lesions, supporting resection planning and objective quality control. Video-aware architectures aggregate information over-time to stabilize predictions and exploit motion and zoom cues relevant to subsurface assessment.

### 3.1.1 Convolutional Neural Networks

As the most common approach in modern computer vision, CNNs are the backbone of a majority of DL architectures applied to upper GI endoscopy. Ranging multiple image modalities and a variety of tasks such as classification, detection, and segmentation, CNNs have delivered unparalleled results both in efficiency and robustness. CNNs are feed-forward neural networks based on scanning the same small filter across an image, extracting low-level cues (e.g edges, color contrast, textures) that are progressively composed into mid-level motifs and, ultimately, task-specific abstract. This makes CNNs

demand a minimum level of pre-processing compared to other image classification algorithms. This independence to prior knowledge and human effort are the main advantages of its application [32].

The building blocks of CNNs are the convolutional layers, that take an image and pass a small, learnable filter called kernel across it to produce new feature maps. At each spatial location the layer computes a weighted sum of pixel values under the kernel plus a bias. Since the same weights are reused everywhere, the layer is locally connected but shares parameters across the images. This makes CNNs need far fewer parameters than fully connected layers on the same input, and makes it so if a patterns shifts in the image, the response shifts with it. A 2D convolutional layer using a two-dimensional image $I$ as input and applying two-dimensional kernel $K$, producing a feature map $S$ can be written as:

$$S(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i-m,\, j-n)\, K(m,n) \tag{3.1}$$

The commutative property of convolution arises by flipping the kernel relative to the input, so that as *m* increases, the index into the input also increases, while the index into the kernel decreases. While the commutative property is useful for writing proofs, it is not a relevant property of neural networks, so most implementations use a related function called cross-correlation, which is the same as convolution but without flipping the kernel: [33].

$$S(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i+m,\, j+n)\, K(m,n) \tag{3.2}$$

Convolution takes advantage of three important ideas: sparse interactions, parameter sharing, and equivariance. Sparse interaction, allowed by the kernel's receptive field, each output depends on a small neighborhood of the input rather than on every pixel. This significantly reduces the number of learnable weights and computation, while directing the model to locally meaningful structures. With parameter sharing, the same kernel is applied at every spatial location of the input, meaning once set of weights detects the same pattern whenever it appears. This reduces parameters further and allows the model to generalize a feature learned in one region to all regions. Finally, equivariance means that shifting the input also shifts the feature map by the same amount, so in relation to endoscopy, a lesion edge or texture patch elicits the same response regardless of position.

A typical layer of a CNN consists of three stages (Figure 3.1). The first stage performs multiple convolutions in parallel, yielding a bank of linear feature maps. In the second, each map is run through a nonlinear activation function, such as the rectified linear activation function. This stage is sometimes called a detector stage and allows the layer to model non-linear effects. Lastly, the third stage, called pooling, reshapes the previous activations by summarizing values over small neighborhoods to produce a more compact representation. There are multiple pooling methods, such as max, average and L2 pooling, but they all serve the purpose of dampening the sensitivity of small translations by aggregating near by responses [33].

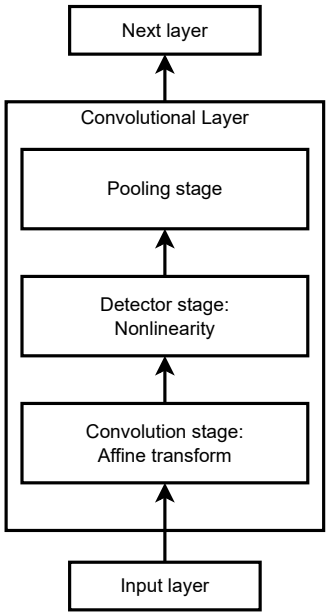


Figure 3.1: Components of a typical CNN layer. In this convention, the convolutional network is viewed as a small number of relatively complex layers, with each having multiple stages.

CNN are trained by empirical risk minimization. Assuming an unknown distribution $p(x, y)$ over images $x$ and their labels $y$, the CNN is a parametric function $f_\theta(x)$ with learnable parameters $\theta$ intended to match the true labels ($y$) from the input ($x$). With that aim, a loss function $l$ is selected to penalize errors, with the population risk being:

$$\mathcal{L}(\theta) = \mathbb{E}_{(x,y)\sim p}\Big[\ell\big(f_\theta(x), y\big)\Big]. \tag{3.3}$$

In practice, the expression cannot be computed, so the error is minimized to the available data, calculating the empirical risk as follows:

$$\hat{\mathcal{L}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\big(f_\theta(x_i),\, y_i\big) \tag{3.4}$$

In classification approaches, the loss function is usually cross-entropy (Equation 3.5), which works well with soft labels, and encourages calibrated probabilities, given heavy penalties for confident mistakes [33].

$$\ell_{\mathsf{CE}}(\mathbf{y}, \hat{\mathbf{p}}) \;=\; -\sum_{k=1}^{K} y_k \,\log \hat{p}_k \;=\; -\log \hat{p}_y, \quad \hat{\mathbf{p}} = \operatorname{softmax}(\mathbf{z}) \tag{3.5}$$

Since labeled medical images are often scarce and heterogeneous, CNNs are commonly initialized from ImageNet-pretrained backbones and fine-tuned to the target task. This allows early convolutional layers to learn generic primitives that transfer well across domains, such as edges, color contrast, and simple textures, while later layers are specialized to the main task. When datasets are closely matched to the target, full-fine tuning is usually better, but when labels are very limited, linear probing or gradual unfreezing tends to be more stable. Nevertheless, even if the final result is similar, pre-training usually accelerates the training process [34].

Regarding evaluation, CNNs are typically with a held-out test set that was never used for training, with patient-level splitting to prevent leakage. After hyperparameter tunning and thresholds defined in the validation set, the frozen model is run on the test set. For classification the main metrics used are:

- **Accuracy:** proportion of correctly predicted instances among the total instances:

$$\text{Accuracy} \;=\; \frac{\mathsf{TP} + \mathsf{TN}}{\mathsf{TP} + \mathsf{FP} + \mathsf{TN} + \mathsf{FN}} \tag{3.6}$$

- **Recall (Sensitivity):** ratio of the true positive predictions to the total actual positives:

$$\text{Recall} \;=\; \frac{\mathsf{TP}}{\mathsf{TP} + \mathsf{FN}} \tag{3.7}$$

- **Specificity:** ratio of the true negative predictions to the total actual negatives:

$$\text{Specificity} \;=\; \frac{\mathsf{TN}}{\mathsf{TN} + \mathsf{FP}} \tag{3.8}$$

- **Precision:** ratio of the true positive predictions to the total predicted positives:

$$\text{Precision} \;=\; \frac{\mathsf{TP}}{\mathsf{TP} + \mathsf{FP}} \tag{3.9}$$

- **F1:** harmonic mean of precision and recall:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.10}$$

where:

- **True Positives (TP):** the actual value is positive and the prediction is also positive.
- **False Positives (FP):** the actual value is negative but the prediction is positive.
- **True Negatives (TN):** the actual value is negative and the prediction is also negative.
- **False Negatives (FN):** the actual value is positive but the prediction is negative.

## 3.2. Radiomics

Modern DL models excel at pattern recognition and are the undisputed state-of-the-art methodology for classification tasks in EGD. However, their internal decision making process is often perceived as opaque making it hard to interpret which visual cues drive prediction. This lack of transparency poses challenges for clinicians where explainability is essential for fostering trust, ensure accountability, and meeting regulatory standards. Radiomics provides a complementary and more explicit approach by converting the images into a set of mathematically defined descriptors, interpretable by humans. These features describe properties such as distributions of pixel intensity, compactness, and irregularities in the shape of lesions, and coarseness of texture patterns in mucosal tissue. Unlike the representations from CNNs, radiomic descriptors are named, standardized, and directly linked to visual characteristics recognized in the images. Additionally, these features can be systematically ranked for importance, tested for reproducibility, and statistically associated with outcomes. The approach was conceptualized by Lambin and colleagues [35] and gained prominence with the landmark study by Aerts et al. [36], which demonstrated that hand-crafted imaging biomarkers from routine CT scans contained prognostic information and reflected underlying histology.

Radiomics is a structured workflow designed to transform raw medical images into high-dimensional, quantitative data (Figure 3.2). The process requires the selection of medical images and the respective delineation of their Region of Interest (ROI), usually provided by manual, semi-automatic, or fully automatic segmentation. This is an essential step,

since the quality of the extracted descriptors depends directly on the precision of the segmentation. Once the ROI is defined, the image can undergo pre-processing to guarantee comparability between datasets and reduce variability introduced by acquisition parameters and devices. Moreover, optical filtering techniques, such as Laplacian-of-Gaussian filter and wavelet decomposition, can be introduced to the images in order to emphasize particular structures and edges, or to highlight micro-texture patterns, respectively. This can be a powerful way to improve the features, but the added complexity requires added care because it makes the descriptors harder to interpret. These features can then be used as input into a variety of classical machine learning methods or be integrated into multimodal approaches that combine engineered radiomic descriptors with latent DL features. In this way, radiomics can provide a more transparent alternative to DL, often with lower computational demands, and can be combined with it to provide hybrid models that balance predictive performance with clinical interpretability.

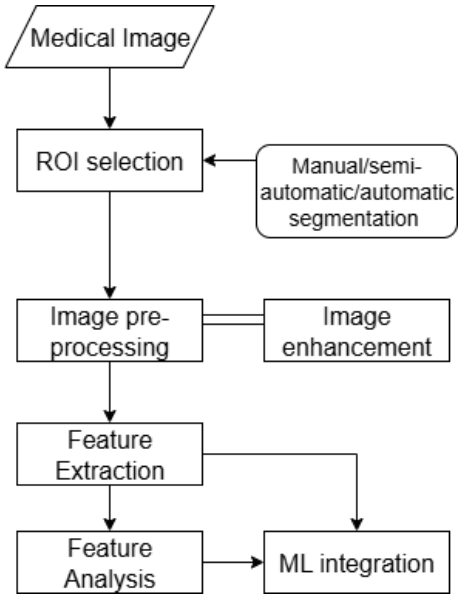


Figure 3.2: Typical radiomic workflow.

### 3.2.1 Feature Extraction

PyRadiomics [37] has become widely used for the extraction of radiomic features in academic and clinical research. Originally introduced in 2017 by van Griethusyen et al. [38], it is an open-source Python implementation of the Image Biomarker Standardization Initiative (IBSI) guidelines, ensuring that the extracted descriptors are mathematically consistent and reproducible between applications. PyRadiomics provides an extensive library of feature classes, grouped according to the type of information they extract from the image ROI:

- **First-order features (intensity histogram):** summarize the distribution of pixel and voxel intensity without considering spatial relationships. Clinically, these features reflect aspects such as brightness or heterogeneity in the mucosal tissue.
- **Shape features (2D/3D):** describes the geometry of the segmented area. Supports 2D and 3D images with the features varying according to the dimension. Quantifies aspects that clinicians already visually assess intuitively, such as irregularities in the form of the lesion.
- **Texture features (spatial gray-level organization):** captures the arrangement of intensities within the image, reflecting heterogeneity and granularity. PyRadiomics implements five major families:
  - **Gray-Level Co-occurrence Matrix (GLCM):** measures how often pairs of gray levels occur in a specific spatial relationship.
  - **Gray-Level Run-Length Matrix (GLRLM):** quantifies consecutive runs of a pixel with the same gray level.
  - **Gray-Level Size Zone Matrix (GLSZM):** captures the size of the distribution of connected regions with the same gray level.
  - **Gray-Level Dependence Matrix (GLDM):** describes the number of connected voxels that depend on a central voxels gray level.
  - **Neighborhood Gray-Tone Difference Matrix (NGTDM):** characterizes the difference between a voxels gray level and the average gray level of its neighbors.

Although quite useful, PyRadiomics is limited in the context of endoscopy by not providing native support for color features, since the library was originally developed for radiology modalities where images are inherently scalar and represented in grayscale intensity values. Consequently, its feature definitions and IBSI conformant implementation assume a single-channel input, focusing on gray-level distributions, shape, and textures rather than color descriptors, which are clinically relevant in endoscopy to access mucosal patterns, erythema, and vascularization. Incorporating color sensitive descriptors will likely be a requirement for radiomics to fully capture the information contained in EGD, not only to increase discriminatory power of radiomic-based models but also to align them with the way endoscopists interpret the images in practice.

### 3.2.2 Feature Evaluation Metrics

Radiomics pipelines can generate hundreds to thousands of candidate features, many of which carry little to no information with respect to the clinical output of interest. Without a principled evaluation stage, redundant features without an association with the disease can hinder downstream predictive models and cause confusion to experts in their analysis of the results. Feature evaluation can be broadly divided into univariate and multivariate approaches. Univariate methods measures the association of each feature individually with the outcome and provide a simple, transparent, and computationally efficient way to compare feature independently. However, by ignoring interactions between the features, univariate approaches may neglect descriptors with modest association that can still become highly informative when combined. Examples include Pearson correlation and Smith's Prediction Probability ($P_k$) [39]. In contrast, multivariate methods consider the joint contribution of multiple features within the context of a predictive model. The trade-off is that multivariate approaches are more complex and require regularization to avoid overfitting, as well as outputting feature importance scores that are less interpretable than univariate rankings. Methods such as LASSO regression, Support Vector Machine (SVM), or tree-based ensembles are notable examples.

For this project $P_k$ was selected as the main measurement of feature effectiveness. $P_k$, unlike Pearson correlation, provides a probability-based solution for discrete variables, particularly tailored for ordinal or graded outcomes, such as EGGIM and dysplasia. Unlike model-based feature importance methods, $P_k$ is model-agnostic and univariate, which means that the evaluation of features is not biased by the classifier or the correlation between features. Finally, $P_k$ is well-established in biomedical signal evaluation and allows for direct clinical interpretation, maintaining a valuable level of transparency on the analysis.

$P_k$ is a rescaling of Kim's $d_{yx}$ measure [40], that based his approach on the Proportional-Reduction-in-Error (P-R-E) criterion. The criterion states that the reduction in the error of the prediction, using variable $X$ to predict $Y$ is proportional to the degree of association between $X$ and $Y$. The P-R-E index is given by:

$$\text{P-R-E index} = \frac{E_{1a} - E_{1b}}{E_{1a}} \tag{3.11}$$

where $E_1a$ and $E_1b$ are the error probabilities under the premise that the value $Y$ is predicted from $X$ assuming that $Y$ is statistically independent of $X$ and $Y$ is a function of $X$, respectively.

Since the data is discrete and ordinal, three relations between labels ($Y$) are possible: $Y_i < Y_j$, $Y_i = Y_j$ and $Y_i > Y_j$ for two samples $i$ and $j$ out of the total number of samples $N$. By also using this logic for $X$ the bivariate distribution can be partitioned into five categories:

$$\begin{array}{lll} P_{xy} & \text{“concordant”} & Y_i > Y_j \quad \text{and} \quad X_i > X_j \\ Q_{xy} & \text{“discordant”} & Y_i < Y_j \quad \text{and} \quad X_i > X_j \\ U_y & \text{“Y-unilateral”} & Y_i \neq Y_j \quad \text{and} \quad X_i = X_j \\ U_x & \text{“X-unilateral”} & Y_i = Y_j \quad \text{and} \quad X_i \neq X_j \\ T_{xy} & \text{“tied on both”} & Y_i = Y_j \quad \text{and} \quad X_i = X_j \end{array}$$

To calculate the error probabilities, it is also necessary to define the ordinal variation of $Y$. This can be derived by:

$$V_y = \sum A_{ij}^2 \tag{3.12}$$

where $A_{ij}$ is 0 for all combinations where $Y_i = Y_j$ and $+/-1$ for $Y_i < Y_j$ and $Y_i > Y_j$, respectively.

Using this information, the error probability $E_{1a}$ is half of the variation in $Y$ ($V_y$) divided by all possible combinations of unique pairs formed from $N$ samples (Equation 3.13). $E_{1b}$, assuming a weakly monotonic relationship, is the number of discordant pairs together with half of the Y-unilateral pairs divided by all possibilities (Equation 3.14).

$$E_{1a} = \frac{\frac{1}{2}V_y}{\frac{1}{2}N(N-1)} \tag{3.13}$$

$$E_{1b} = \frac{Q_{xy} + \frac{1}{2}U_y}{\frac{1}{2}N(N-1)} \tag{3.14}$$

By substituting equations 3.13 and 3.14 into equation 3.11, and applying the necessary transformations, the P-R-E index for ordinal data is obtained:

$$d_{yx} = \frac{P_{xy} - Q_{xy}}{V_y} \tag{3.15}$$

Kim's $d_{yx}$ range is $[-1, 1]$ and indicates no association for $d_{yx} = 0$. In order to get an easier interpretation of the meaning of $d_{yx}$ Smith et al. rescaled the measure and introduced the prediction probability $P_k$ as:

$$P_k = \tfrac{1}{2}\left(d_{yx} + 1\right) \tag{3.16}$$

This association measure ranges from 0 to 1. In analogy to $d_{yx}$, a value of $P_k = 1$ indicates perfect association, whereas $P_k = 0.5$ reflects no association. In the context of this thesis, the *direction* of association is not of primary interest. For example, a value of $-0.7$ represents an inverse association but should be considered comparable in strength to $0.7$. To simplify interpretation while retaining 0.5 as the no-association baseline, the $P_k$ scores were symmetrized by reflecting values below 0.5 across 0.5. This yields a direction-agnostic measure in which stronger associations (whether above or below $0.5$) are mapped to the upper half of the scale (e.g., $P_k = 0$ and $P_k = 1$ both give $1$), while $P_k = 0.5$ remains 0.5:

$$P_k^* = 0.5 + |P_k - 0.5| \tag{3.17}$$

To complement $P_k$ as the primary feature evaluation metric, the classification ability of the features was assessed with a SVM baseline, trained under cross-validation. Historically, before advancements in DL, SVMs trained on hand-crafted features was the state-of-the-art approach for most medical image analysis, being the most popular and often the best performing classifier [41].

SVMs operate by finding a decision boundary that maximizes the margin between the classes. For linear decision margins, the model learns a hyperplane $wx + b = 0$ that separates the data while maximizing the margin. Soft-margin classification introduces slack variables $\xi_i$ to allow some level of missclassification depending on a regularization parameter $C$. In multiclass settings, SVMs use decomposition strategies:

- **One-vs-One (OvO):** creates a binary classifier for all possible pairs of classes. Each classifier is responsible for identifying the margins between two specific classes.

Particularly effective with a small number of classes.

- **One-vs-Rest (OvR):** trains one binary classifier for each class. Each class is treated as the positive class, while the other are grouped as the negative class.

## 3.3. Related Work in GC Characterization

Computer vision applications on upper GI endoscopy have shown consistent evolution in the past few years. Moving from traditional machine learning models trained with engineered features to data-driven DL methodologies that directly learn from images. The current landscape of AI in EGD focus on frame or clip level classification tasks such as *H. pylori* status, diagnosis of lesions, histology differentiation, and invasion-depth estimation.

Pioneering work by Shichijo and colleagues trained and fine-tuned a CNN on routine endoscopic images for *H. pylori* gastrisis. In a 397 patient test with 11,481 WLE images, the architecture achieved 88.9% sensitivity and 87.4% specificity on a per-patient basis, outperforming endoscopists in both accuracy and reading time [42]. Building on this, Li et al. used magnifying NBI to also train a CNN that discriminated GC, demonstrating high accuracy on external test data [43]. Beyond single examples, meta-analytic syntheses now indicate that image-based AI for EGD routinely performs well across centers and devices with a 2022 review by De Luo reporting a pooled AUC of 0.94 with sensitivity and specification around 0.87-0.88 [44].

Despite or maybe because of DLs dominance in EGD, radiomics has seen very limited use in upper GI applications. Endoscopy-radiomics reports for upper GI are typically for EUS, where the images are in grayscale and consequently more aligned with the usual radiomic framework. Cai et al. developed an EUS-based radiomic model to discriminate small gastric submucosal tumors. The optimized pipeline demonstrated an area under the curve of 0.76 for the diagnosis of GIST and 0.73 for risk stratification of the same cancer, outperforming experienced endoscopists by more than 20% [45].

By contrast, radiomics is already widely used for GC in CT modalities, with multiple systematic reviews validating the clinical utility of the methodology. A 2023 meta-analysis by HajiEsmailPoor et al. aggregated 15 studies with more than 7000 GC patients and reported a pooled area under the curve of 0.85 for the prediction of lymph node metastasis with radiomic-based models [46]. A complementary 2024 systematic review centered on radiomic-based machine learning models for neoadjuvant chemotherapy response and survival prediction in GC patients pooled 14 studies involving 3000 patients and showed

a sensitivity of 0.67 and specificity of 0.77, increased to 0.78 and 0.73, respectively, when clinical data was integrated [47].

Recent works shows that AI for GC characterization has been almost entirely driven by DL applications, which provides strong image-level performance and robustness across centers and modalities. Radiomics has shown great promise in CT, being capable of capturing prognostic and predictive qualities in images. However, radiomic approaches are almost completely unexplored for upper GI applications. This gap in the research presents an opportunity to translate the maturity of radiomics in CT to the endoscopic domain and analyze how it can be used in modalities where color is an essential visual descriptor for the decision making of clinicians.

# 4. A Radiomics-Based Approach to Gastric Imaging Biomarker Extraction in Upper GI Endoscopy

This chapter reproduces, almost verbatim, the content of my peer-reviewed conference paper accepted for presentation at the 8th ENBENG 2025, ”A Radiomics-Based Approach to Gastric Imaging Biomarker Extraction in Upper GI Endoscopy”. To preserve the integrity of the reviewed work, the text, figures, and results are carried over with light editorial adjustments to fit the thesis.

## 4.1. Motivation and Objectives

The motivation of the paper follows the thesis-level motivation (see 1.1). Briefly GC is one of the most common and fatal cancers at a worldwide scale. EGD plays a major role in the management of this disease enabling on one hand the detection and localization of the presence of GIM, a precancerous condition that is an initial step in the cascade of events leading to dysplasia subsequent intestinal-type gastric adenocarcinoma, making it an important indicator for GC risk stratification. AI, and more specifically computer vision, has the potential to create an intermediate virtual biopsy step between today's endoscopy and biopsy procedures, by exploring how radiomics features can act as quantitative image biomarkers, providing surrogates of clinical information extracted from gastric biopsies.

The main objective of this work is to explore how a conventional radiomics approach behaves when applied to a gastroenterology scenario. More specifically, texture, shape and color features are extracted from two EGD scenarios ( GIM characterization, dysplasia characterization), assessed for their association with the clinical characterization obtained by physicians, and analyzed using simple statistical classifiers. The study explicitly explores how standard EGD imaging datasets, traditionally evaluated qualitatively by expert clinicians, can benefit from radiomic feature extraction, thus providing quantitative and interpretable biomarkers.

## 4.2. Materials

### 4.2.1 Dataset A - IPO EGGIM

Two datasets collected using an OLYMPUS CV-190 endoscope from the Instituto Português de Oncologia do Porto (IPO-Porto), Portugal, were merged and used in this study. One is a retrospective dataset with 414 NBI images, collected between December 2019

and September 2020. The second is a prospective dataset collected in 2024 with 866 NBI images from 80 patients. A total of 1280 images annotated by experts using the EGGIM scoring system were used. Dataset A is markedly imbalanced across diagnostic classes, with an EGGIM score of 1 comprising only 19% of the images (Figure 4.1). Since the analysis is performed on an image basis and not on a patient level, this needs to be addressed in the classification portion of the work to ensure the results are correctly interpreted.

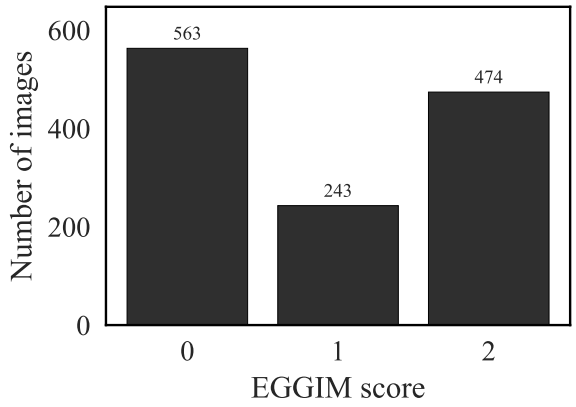


Figure 4.1: Distribution of EGGIM score across Dataset A images.

Each image of the dataset was manually framed into a 224x224 visual patch, according to the following criteria: (i) prioritize the center of the image, (ii) avoid medical devices (like the endoscope), and (iii) ensure representativeness of the anatomical location. Feature extraction is therefore performed within this ROI, so shape and size-based descriptors will not carry any meaningful information.

### 4.2.2 Dataset B - IPO Dysplasia

A retrospective dataset was also collected at IPO-Porto on regular clinical practice and includes 199 WLE images. These images were collected using a FUJIFILM VP-7000 endoscope. Each image was classified as either low-grade or high-grade dysplasia, based on the results of its corresponding biopsy. Similar to Dataset A, Dataset B has a substantial imbalance in the number of images for each label, with high-grade dysplasia images being 4 times more common than low-grade ones (Figure 4.2). Reiterating, this requires extra consideration during the evaluation of the classification prowess of the radiomic features produced by this dataset.

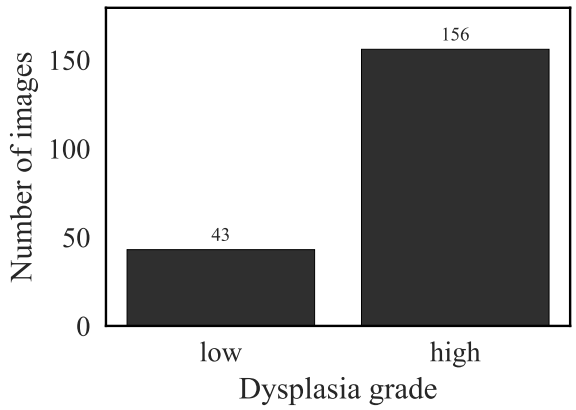


Figure 4.2: Distribution of dysplasia grade across Dataset B images.

For this dataset lesions were fully segmented at a pixel level by experts, yielding binary masks that delineate the extent of the disease. All radiomic computations are restricted to this ROI, which enables shape features and reduces background contamination.

## 4.3. Methods

### 4.3.1 Analysis Framework

The framework (Figure 4.3) is composed of four main modules: Materials, containing the datasets previously explained; Feature Extraction, radiomic features are extracted from the ROI of the images; Features Selection, the descriptors are evaluated using $P_k$ and sorted accordingly; Classification, SVM baseline model is used to evaluate the predictive strength of the features.

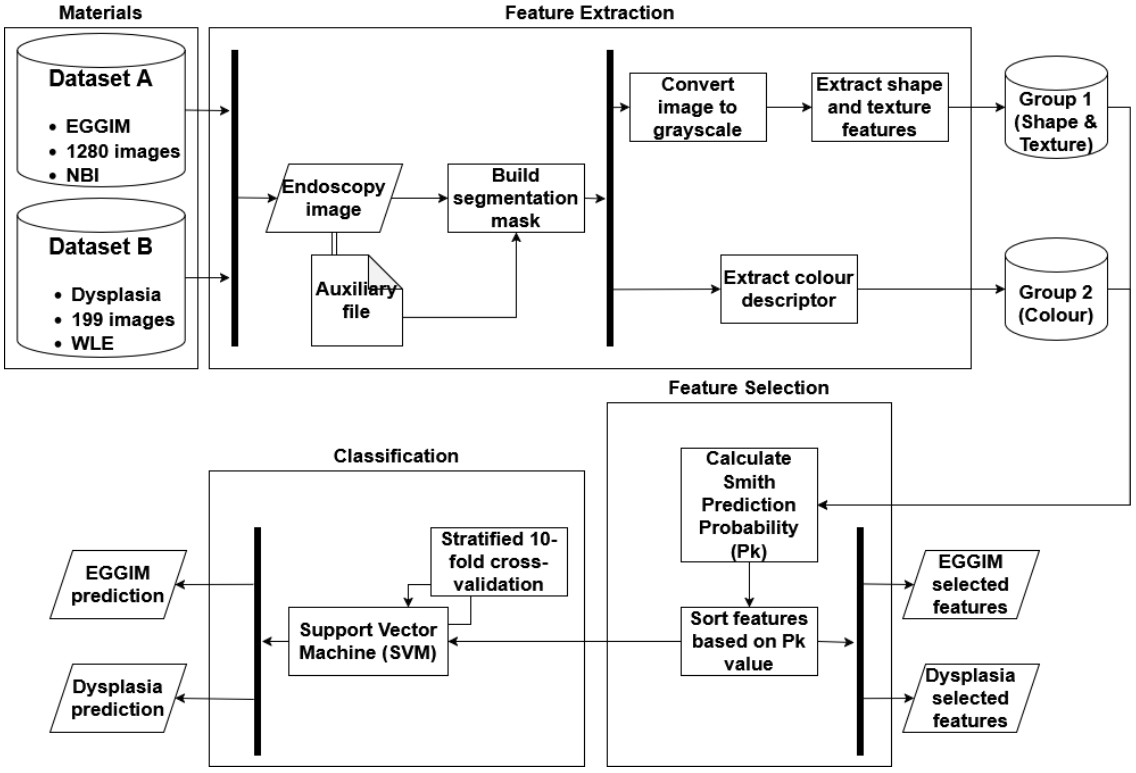


Figure 4.3: Analysis framework used in Chapter 4

### 4.3.2 Feature Extraction

**Group 1 - Shape and Texture Features**

This group comprises a diverse set of handcrafted radiomic features capturing both shape-related and textural characteristics of the lesions, extracted using the open-source PyRadiomics library. Specifically, features from seven core families were included: First Order Statistics (19 features); Shape-based (2D) features (10 features); and five texture feature classes. These include the GLCM (24 features); GLRLM (16 features); GLSZM (16 features); NGTDM (5 features); GLDM (14 features). Together, these 104 features offer a comprehensive characterization of shape and texture patterns within endoscopic images.

**Group 2 - Color Features**

This group captures the chromatic characteristics of the endoscopic images through a histogram-based representation in the Hue–Saturation–Value (HSV) color space, which is well-suited for separating color information from intensity. For each image, a color histogram was computed using a uniform quantization of Hue (16 bins), Saturation (4 bins), and Value (4 bins), resulting in a total of 256 unique combinations used as color

features. In this specific scenario, the majority of these bins are actually zero, due to the absence of a large variety of colors.

### 4.3.3 Feature Analysis

To preliminarily assess the discriminative power of each extracted radiomic feature with respect to clinically relevant outcomes, the $P_k$ measure was employed (explained in 3.2.2). In this study, $P_k$ was calculated individually for all extracted features in relation to two distinct classification tasks: (i) the EGGIM score level (ordinal values: 0, 1, 2) in Dataset A, and (ii) the severity of dysplasia (binary classes: low-grade, high-grade) in Dataset B. Features yielding higher $P_k$ values were considered more informative for downstream analysis and model development, supporting an initial, task-specific filtering of the high-dimensional feature space.

### 4.3.4 Classification

To verify the practical advantage of the retained features and validate the effectiveness of the feature selection process, a SVM classifier was selected. The classifier was applied separately to the two classification tasks defined in this study: (i) prediction of EGGIM score levels using Dataset A, and (ii) discrimination between low-grade and high-grade dysplasia using Dataset B. For each task, the SVM model using a subsets of features with the highest $P_k$ values, as determined in the feature selection step. Model performance was evaluated using repeated stratified 10-fold cross-validation to ensure robustness and generalizability of the results across different data splits. In all experiments, the SVM used a radial basis function (RBF) kernel and a class weighting inversely proportional to the class frequency, adjusting to the imbalance seen in both datasets. For Dataset A, the multi-class nature of the prediction was addressed using an OvR scheme.

## 4.4. Results

### 4.4.1 Feature Analysis

Figure 4.4 provides an overview of the distribution of the $P_k$ values across the different feature groups for each dataset. As previously described, $P_k$ values deviating further from 0.5 (indicated by the red dashed line) reflect stronger associations with the corresponding clinical outcome.

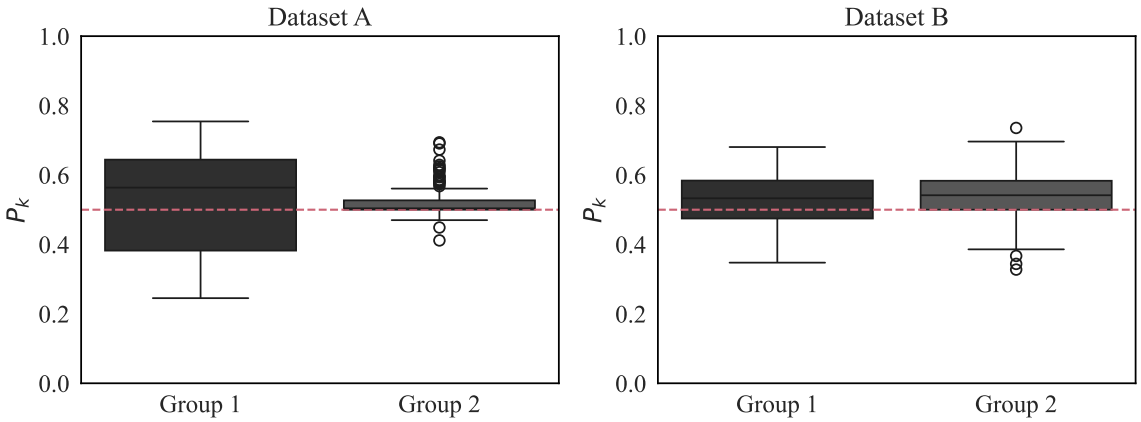


Figure 4.4: Distribution of $P_k$ for the grouped radiomic features.

It is immediately evident that the majority of features from Group 2 (color features) exhibit minimal association with the outcome classes, as reflected by their $P_k$ values very close to the red dashed line. This observation aligns with the underlying distribution of the HSV histograms, where color information is heavily concentrated within a limited number of bins. Nonetheless, a subset of non-zero color descriptors demonstrated notable discriminative power. This is particularly evident in Dataset B, where the feature with the highest $P_k$ score was, in fact, a color feature. On the other hand, in Dataset A, the highest $P_k$ values are observed among features from Group 1 (shape and texture). To further illustrate the contribution of individual feature families within Group 1, Figure 4.5 disaggregates the group into its subgroups, providing a more detailed view on which specific features have a greater influence on class differentiation.

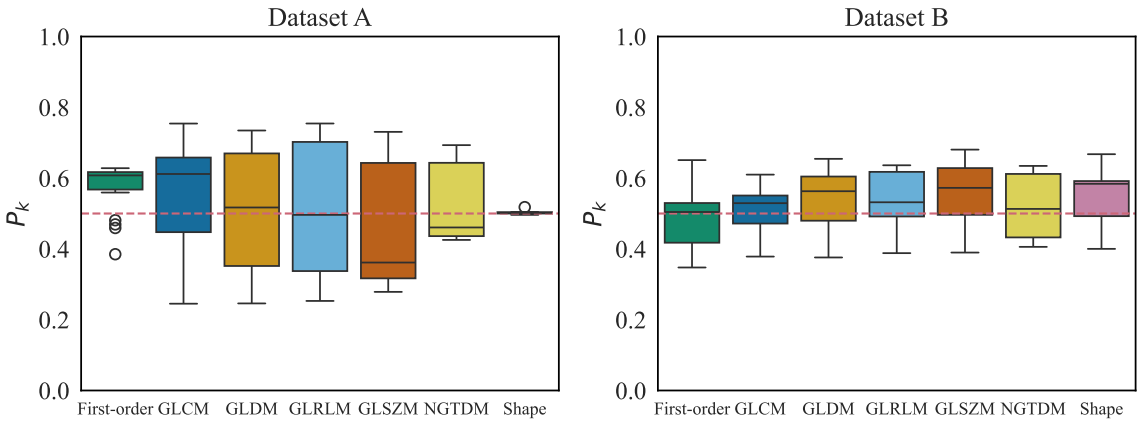


Figure 4.5: Distribution of $P_k$ for Group 1 (Shape and Texture) radiomics features.

Texture features appear to be particularly informative for Dataset A, with GLCM, GLDM, and GLRLM exhibiting $P_k$ values deviating more than 0.25 from the baseline of 0.5 in both directions, indicating strong discriminatory capacity. In contrast, as also reflected

in Figure 4.4, Group 1 features show less predictive relevance in Dataset B. The most notable deviations from 0.5 in this dataset are observed in GLSZM and shape features, though none exceed a 0.2 difference. This behavior can be better seen in Figure 4.6 that presents the $P_k$ scores for all features, irrespective of directionality ($P_k^*$), sorted in descending order to highlight the most informative descriptors across both datasets.

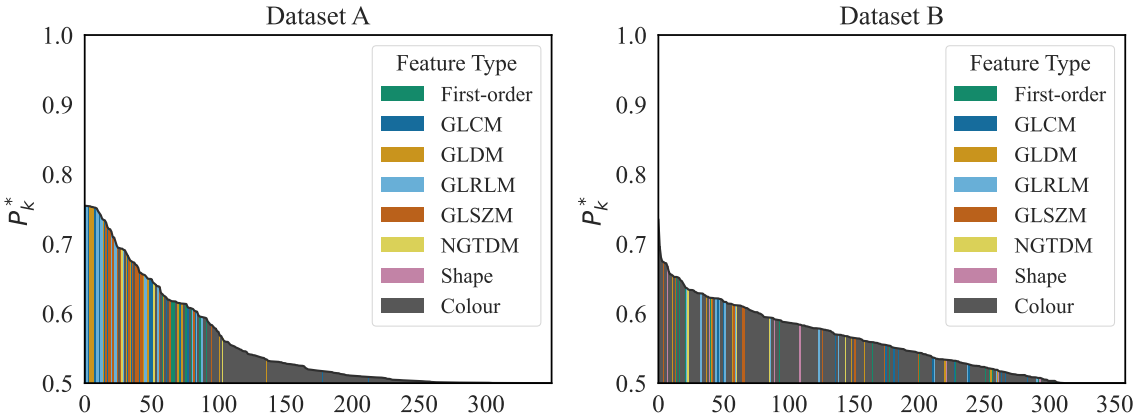


Figure 4.6: $P_k$, irrespective of direction ($P_k^*$ - see 3.17), for each individual feature in descending order.

There is limited overlap in the most predictive features between the two datasets. When considering the top 100 features ranked by $P_k^*$ (explained in Equation 3.17), the Jaccard similarity between the feature sets is 0.24, indicating a low degree of intersection. The distribution of shared features across subgroups is as follows: 5% from NGTDM, 10% each from GLCM and GLDM, 15% from first-order statistics, 17% from GLRLM, and 20% from both GLSZM and color features.

### 4.4.2 Classification

Following the feature selection process, the predictive performance of the selected features was evaluated using a SVM classifier. To assess how the most informative features (ranked by their $P_k$ values) contribute to the classification of clinical outcomes, a fixed SVM base model was employed while varying the number of top-ranked features included in the input. This iterative approach allowed us to explore the relationship between feature dimensionality and classification performance across both datasets. It is important to reiterate that both datasets exhibit considerable class imbalance. Accordingly, 4.7 illustrates the progression of the F1 score for each dataset as the number of top-ranked features included in the model increases.

For both datasets, the best average results are achieved when combining the features from both groups. However, when considering the groups individually, images from Dataset A seem to be better classified using texture features, while Dataset B seems to perform better with color features. In the case of Dataset A, the performance of the combined model mirrors that of Group 1 alone for the top 25 ranked features, as color descriptors are only introduced beyond this point. A noticeable increase in performance follows, attributable to the higher-ranked color features, which prove to be informative, as reflected by a sharp rise in F1 score when the top 15 color descriptors are included. Nevertheless, a subsequent drop in performance occurs as lower-ranked, less informative color features are introduced, highlighting their contribution to noise. The model's performance stabilizes after approximately 60 features, with only minor variations in the F1 score beyond this point. In Dataset B, color features outperform shape and texture features, reaching a local maximum at around 40 descriptors, after which performance slightly declines. For Group 1 in this dataset, a notable increase in F1 score is observed at the 40-feature mark, stabilizing again beyond 60 features. This improvement also positively influences the performance of the combined feature group. A similar trend can be observed in Dataset A, where an increase in performance is seen between 25 and 40 features in Group 1. Interestingly, reevaluating the feature similarity between these intervals in both datasets reveals no significant overlap, suggesting that different features are driving the improvements in each case.

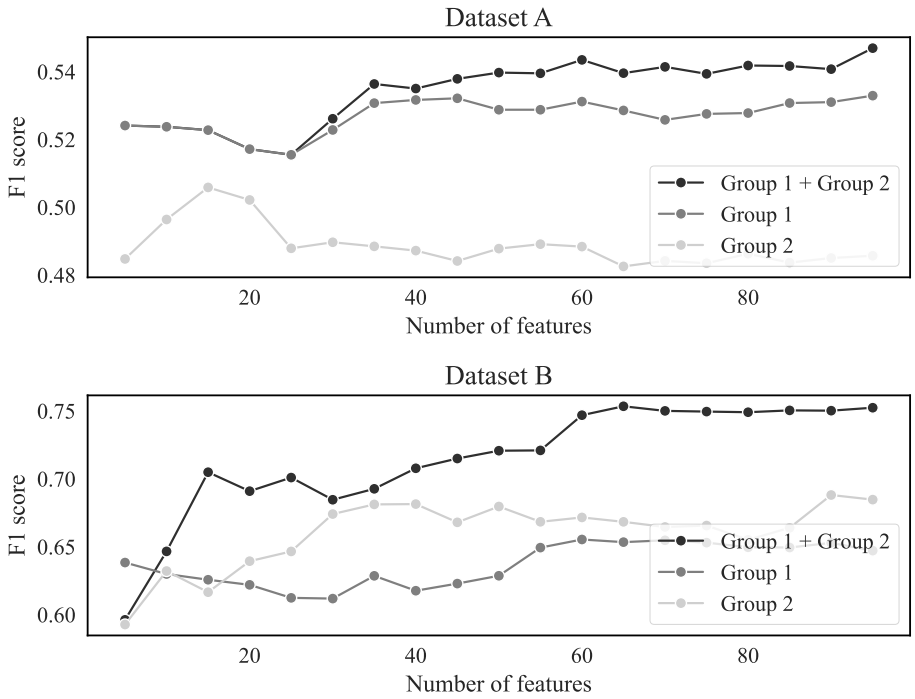


Figure 4.7: Average F1 score of the SVM classifier evaluated using 10-fold cross-validation, repeated 10 times. The SVM was trained and tested using Group 1 (shape and texture features), Group 2 (color features), and a combination of both groups.

## 4.5. Discussion

This study investigated the potential of radiomic and color features to support clinical classification tasks in endoscopic image analysis, focusing on two datasets with distinct characteristics. The combined use of the $P_k$ metric for feature selection and SVM classifiers for performance evaluation provided a robust framework to explore feature relevance and predictive utility. The feature selection process revealed notable differences in the predictive relevance of radiomic and color descriptors between the two datasets. In Dataset A ( EGGIM scoring), texture-related features, particularly those from GLCM, GLDM, and GLRLM, emerged as the most informative. These texture related radiomic features have high probability of being impacted GIM related changes in the stomach mucosa, as shifts in glandular and mucin types, accumulation of goblet cells [48] and changes in the capillary patterns [49]. Another important factor is the poorer color information present in NBI when compared to the WLE used in Dataset B.

In contrast, Dataset B (dysplasia classification) exhibited a stronger contribution from color features. This is especially intriguing considering that, overall, the color space in both datasets was sparsely populated due to the nature of the HSV histogram representation. Despite this,

certain bins encapsulated clinically meaningful chromatic variations, as evidenced by the presence of a color feature with the highest $P_k$ score in Dataset B. Dysplastic gastric mucosa vascular (angiogenesis and inflammation) [50] and structural (altered mucosa topography, shifted glandule type, cellular crowding and mucin depletion) [50][51] changes have high potential to interfere with light propagation and impact color radiomic features.

From a biological perspective, these results make sense. Observation of these images does indeed hint at the richer texture variability of GIM, and to the distinct color features (an observed 'redness') of dysplasia lesions.

These findings highlight the dataset-specific relevance of feature types. The low Jaccard similarity (0.24) between the top 100 features of each dataset further reinforces the notion that distinct radiomic signatures are associated with different pathological conditions and classification tasks. This low overlap was consistent across subgroups, with texture features such as NGTDM and GLDM showing minimal intersection. The classification results mirrored the insights gained from feature selection. In both datasets, combining features from all groups yielded the highest average F1 scores. However, the contribution of each group varied considerably depending on the dataset. For Dataset A, Group 1 (shape and texture) alone provided strong classification performance, particularly in the top 25 features, beyond which color features began to enhance the model. Notably, a spike in F1 score occurred with the addition of the top-ranked color descriptors, though performance declined when less informative color bins were introduced, underscoring the importance of selective feature inclusion to avoid overfitting or introducing noise. In Dataset B, the advantage of color features was more pronounced, with peak performance achieved at around 40 descriptors. This suggests that chromatic information plays a more central role in the visual differentiation of dysplasia severity. Interestingly, texture features also contributed meaningfully, with a substantial performance gain noted after the 40-feature threshold. This pattern was consistent across both datasets, where performance stabilized after the inclusion of 60 features, indicating a plateau in incremental value.

As final conclusions of this study, confirmed that radiomics is a valid and interesting approach to gastric imaging biomarker extraction, in which visually explainable descriptors and be potentially mapped to biological meaning, offering a viable and complementary alternative to the opaque nature of current deep neural network approaches.

# 5. Exploratory Analysis on Feature Effectiveness Across Transparency Levels in Upper GI Endoscopy

This chapter conducts an exploratory comparison of feature effectiveness across transparency levels in upper GI endoscopy. Three representation families are evaluated under a common pipeline: (i) transparent handcrafted radiomics, (ii) semi-transparent radiomics compressed by PCA (on the full set and after supervised preselection), and (iii) opaque deep embeddings from a frozen ImageNet-pretrained ResNet-18. Building on the materials and general procedures established previously, features are extracted from lesion ROIs, ranked by $P_k$ for two clinical tasks (EGGIM levels and dysplasia grade), and assessed with a fixed RBF SVM under repeated stratified cross-validation while sweeping the number of top-ranked features. This design isolates the effect of representation and transparency, enabling a controlled, like-for-like appraisal of the interpretability–effectiveness trade-off. The chapter opens with a concise statement of the motivation and experimental objectives (Section 5.1), followed by a summary of the materials (Section 5.2) and the methodology (Section 5.3). Results are then presented in detail (Section 5.4) and interpreted in the context of the task and prior chapters (Section 5.5).

## 5.1. Motivation and Objectives

The motivation of this work follows the thesis-level motivation (see 1.1). Succinctly, GC is the fifth most common and deadly cancer worldwide. EGD is central to the management of the disease by enabling detection and precise mapping of GIM. This work explores whether AI driven computer vision can provide an alternative to biopsy, utilizing radiomics as quantitative image biomarkers that approximate histological information.

The main objective of the work is to quantify how the effectiveness of features extracted from EGD images changes according to the level of transparency. Concretely, three families of features will be explored: fully interpretable radiomic features (Transparent), dimensionality-reduced radiomic features via Principal Component Analysis (PCA) (Semi-Transparent), and deep learning features from a ResNet-18 backbone pretrained with ImageNet (Opaque). The testing framework is designed to map the trade-off between interpretability and effectiveness and provide an additional, empirically grounded validation of radiomics usefulness in upper GI applications.

## 5.2. Materials

The materials used are identical to those detailed in 4, refer to 4.2 for full specifications.

### 5.2.1 Dataset A - IPO EGGIM

Comprises 1280 NBI images from IPO-Porto (OLYMPUS CV-190 endoscope), merged from a retrospective set (414 images, 2019–2020) and a prospective set (866 images, 2024). Images are annotated with EGGIM (0/1/2) and the distribution is highly imbalanced, with EGGIM=1 at 19% (Figure 4.1).

### 5.2.2 Dataset B - IPO Dysplasia

Contains 199 WLE images from IPO-Porto (FUJIFILM VP-7000 endoscope), each labeled as low-grade or high-grade dysplasia per biopsy. The dataset is substantially imbalanced, with high-grade images approximately four times more frequent than low-grade (Figure 4.2).

## 5.3. Methods

### 5.3.1 Analysis Framework

The framework (Figure 5.1) is composed of four main modules: Materials, containing the datasets previously explained; Feature Extraction, radiomic and DL features are extracted from the ROI of the images. A PCA is applied to the radiomic features; Features Selection, the descriptors are evaluated using $P_k$ and sorted accordingly; Classification, SVM baseline model is used to evaluate the predictive strength of the features.

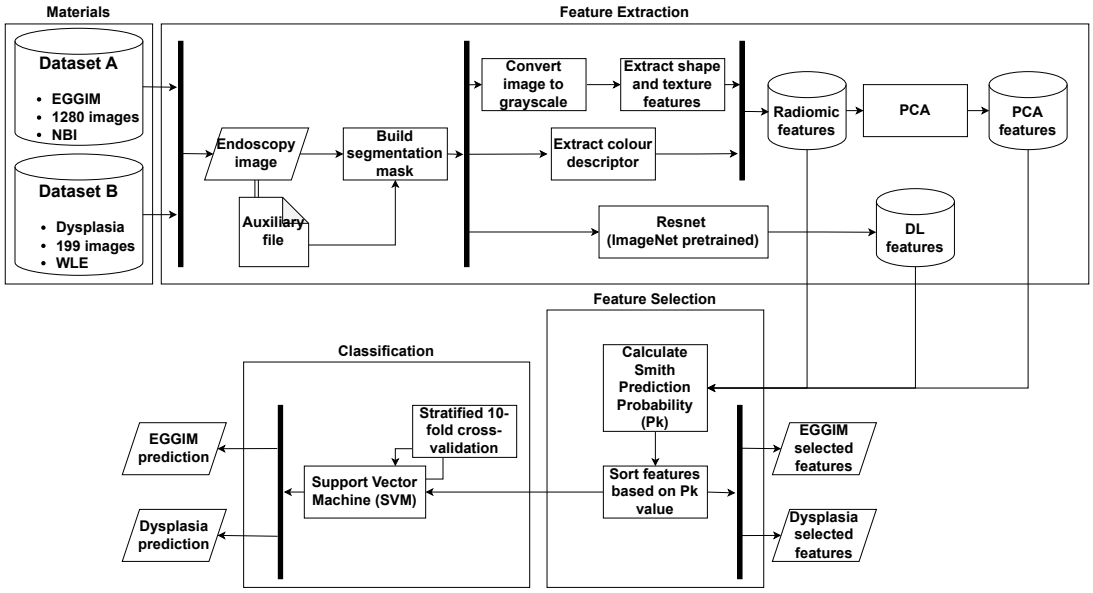


Figure 5.1: Analysis framework used in Chapter 5

### 5.3.2 Feature Extraction

To span different levels of transparency three feature representations were selected: (i) hadcrafted radiomics (transparent and clinically interpretable descriptors), (ii) PCA-compressed radiomics (semi-transparent linear mixtures that retain the variance while reducing collinearity), (iii) Resnet embeddings (opaque deep features that prioritize transfer performance sacrificing interpretability). This three families provide a compact yet representative spectrum for probing the interpretability-effectiveness trade-off.

**Group 1 - Radiomics (transparent features)**

The radiomic feature set and extraction parameters are identical to those defined in the previous chapter (see 4.3.2). Specifically, the 104 descriptors from PyRadiomics (First-Order Statistics, Shape-based (2D) and Texture), and a 256-bin HSV color histogram.

**Group 2 - Principal Component Analysis (semi-transparent features)**

A PCA is an unsupervised linear transform that takes high-dimensional, often correlated feature sets and provides a way to decorrelate and compress the information into a lower-dimensional representation that retains most of the variance of the original data. Concretely, after z-scoring the features, PCA calculates the eigenvectors of the covariance matrix (orthogonal directions where the data varies most) and orders them by decreasing explained variance (eigenvalues). Projecting the original data onto the eigenvectors gives the principal components, which are uncorrelated thanks to the orthogonal axes. Keeping

only the components with highest eigenvalues, that yield most of the variance from the original data, tends to make classifiers more stable.

In this study, all principal components are retained and dimensional adjustments are deferred to downstream feature selection with $P_k$, enabling quantitative identification of clinically informative components without imposing an *a priori* cutoff. This is especially important because, as an unsupervised method, PCA may relegate low-variance but label informative descriptors to trailing components. To further address this issue two PCA pipelines will be considered: (i) PCA applied to the full radiomic feature set and (ii) PCA applied to a subset comprising the most informative features preselected by $P_k$.

**Group 3 - Deep Learning (opaque features)**

A ResNet-18 backbone pretrained on ImageNet is used strictly as a fixed feature extractor. This is possible by removing the classification head, and the 512-dimensional embedding produced after global average pooling serves as the DL feature vector for each image. No fine-tuning is performed and the network remains frozen, so the model weights remain those learned from ImageNet, and observed performance differences arise from input variance and not task-specific adaptation. For Dataset A the images are cropped to the ROI (already 224x224), while for Dataset B the segmentation mask is firstly applied to zero out non-lesion pixels and then resized to 224x224. Images are subsequently normalized with ImageNet channel statistics and fed in batches through the frozen Resnet-18 backbone, and the post global average polling are flattened to a 512-D vector. Resnet-18 architecture was used because it produces a similar amount of features to the radiomic dataset.

### 5.3.3 Feature Analysis

As in 4.3.3, the same discriminative screening is used, now applied uniformly across all feature families described in Section 5.3.2: Radiomics, PCA on the full radiomic set, PCA after $P_k$-based preselection, and pretrained Resnet-18 embeddings. For each family, $P_k$ is calculated per feature with respect to two tasks: (i) EGGIM score (ordinal values: 0, 1, 2) in Dataset A, and (ii) dysplasia grade (binary classes: low-grade, high-grade) in Dataset B.

### 5.3.4 Classification

The classification procedure mirrors the protocol already detailed in 4.3.4. An RBF-kernel SVM with inverse frequency class weight ( OvR for Dataset A), and repeated stratified 10-fold cross-validation. Here, the same pipeline is applied to all feature families, using top-ranked $P_k$ features selected to validate how the classifier performs across different opacity levels.

It is important to note that Group 1 and Group 2 have fewer features than Group 3. Given the sparse population of the HSV histogram most of the bins are 0 and the respective color descriptors are meaningless, therefore being removed for the classification portion. However, the Resnet-18 architecture outputs 512 features, creating a discrepancy between the groups. To circumvent this, the classifier will only be trained up to the top-100 $P_k$ rated features, filtering out other descriptors with little to no association with the clinical output. This will penalize the DL approach that has more moderately associated features, so the full results will be addressed in Section 5.5.

## 5.4. Results

### 5.4.1 Feature Analysis

Figure 5.2, shows the distribution of $P_k$ for the four feature families. Values near the red dotted line ( $P_k = 0.5$) indicate no association with the output label ( EGGIM score and dysplasia grade for Dataset A and B, respectively), whereas larger deviations from the line express stronger discriminative signal. Group 1 (radiomic features) has already been examined in 4.4, with the only notable difference from this representation is that all descriptor families are pooled into a single set. In Dataset A, the chromatic homogeneity makes most HSV histogram bins carry little signal and cluster around $P_k = 0.5$. As a result, in the box-plot the features with higher $P_k$ lie far from the interquartile range and are flagged as ”outliers”. This is purely a visualization effect and those points are not anomalies, but the most informative descriptors in Group 1.

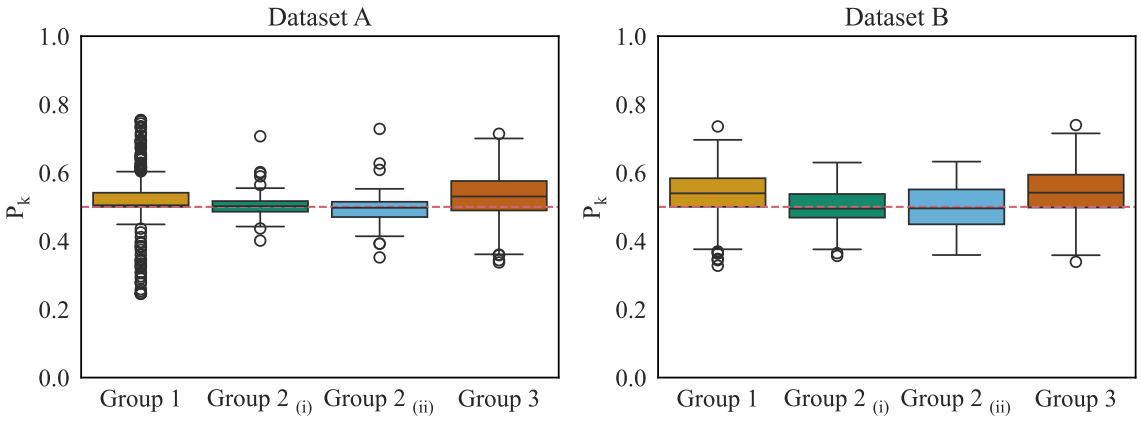


Figure 5.2: Distribution of $P_k$ for the feature families.

Regarding the PCA families, there are not substantial differences between Group $2_{(i)}$ (PCA on all radiomic features) and Group $2_{(ii)}$ (PCA on preselected radiomic features). For Dataset A, most Group $2_{(i)}$ descriptors show little or no association with the EGGIM score, with some notable outliers that, although representative, do not match the best features from Group 1. Group $2_{(ii)}$ shows a similar pattern, yet its components exhibit overall higher $P_k$ than Group $2_{(i)}$, indicating better alignment with the clinical outcome. Conversely, for Dataset B, there are no components with significant association with the labels. The best component for both Group $2_{(i)}$ and Group $2_{(ii)}$ is at a $P_k^*$ value of 0.64, lacking the discriminative prowess of the other feature families.

Group 3 (ResNet-18 embeddings) shows, on average, stronger association with the clinical labels in both datasets. In Dataset A, however, the top-tier features from Group 1 still exceed the best from Group 3, whereas in Dataset B the DL features surpass the base radiomic descriptors across the board.

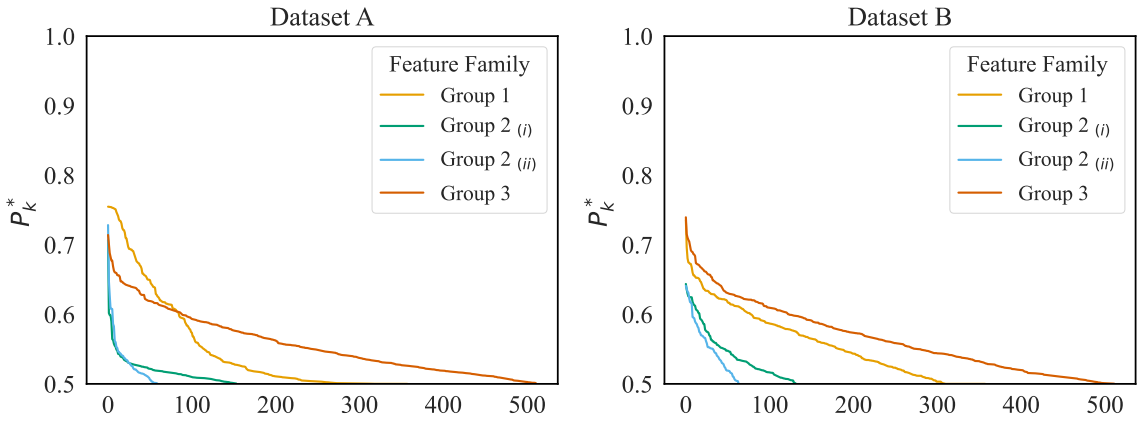


Figure 5.3: $P_k$, irrespective of direction ($P_k^*$ - see 3.17), for each feature across the four feature families.

These patterns are further illustrated in Figure 5.3, where $P_k^*$ (explained in Equation 3.17) values are ranked in descending order for each feature family. In Dataset A, Group 1 provides the strongest top-ranked features. However, after the 87th rank, Group 3 descriptors surpass the radiomic features and exhibit a smoother, slower decay, indicating many moderately informative descriptors. Both PCA families show a brief early spike followed by a rapid drop towards 0.5, with Group $2_{(ii)}$ achieving a higher peak and sustaining more components with nontrivial association to the EGGIM score. In Dataset B, Group 3 dominates across most ranks, retaining more label-associated features. Group 1 maintains mid-level signal with a smooth, consistent decline, whereas both sub-groups from Group 2 are outperformed by all other families.

### 5.4.2 Classification

To enable a fair, like-for-like comparison across feature families, predictive performance was assessed with a fixed RBF- SVM. For each family and dataset, the input comprised the top-$k$ features ranked by $P_k$, with $k$ swept to trace dimensionality–performance curves. Given the class imbalance, Figure 5.4 shows the progression of the F1 score as $k$ increases.

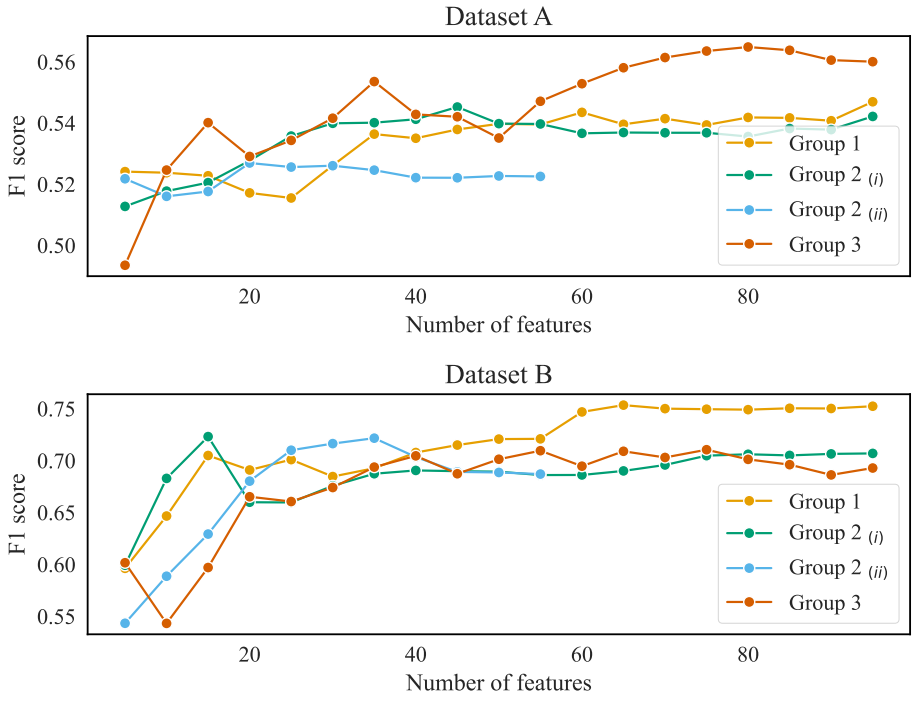


Figure 5.4: Average F1 score of the SVM classifier evaluated using 10-fold cross-validation, repeated 10 times. The SVM was trained and tested using Group 1 (radiomic features), Group 2 (PCA features), and Group 3 (Resnet-18 embeddings)

In Dataset A, Group 3 improves rapidly and after $k = 55$ surpasses the other groups by a considerable margin, plateauing at an F1 of 0.56. Group 1 and Group $2_{(i)}$ level off at a similar feature count, with the PCA on all radiomics slightly outperforming the base radiomic when the $k$ is lower (between 15 and 45). Group $2_{(ii)}$ lags throughout, peaking near F1 = 0.52, just after 20 features. For Dataset B, Group 1 performs better than Group 3 with a lower number of features, but as the $k$ increases the Resnet-18 embeddings improve steadily to almost match it. Both PCA variants follow a similar trend, peaking early, dip modestly, and then stabilize around a F1 score of 0.7, with Group $2_{(i)}$ reaching the maximum at 15 features and Group $2_{(ii)}$ only at 35.

## 5.5. Discussion

This chapter set out to examine whether the effectiveness of features for gastric lesion characterization changes with their degree of interpretability, and to benchmark four representation families under a common evaluation pipeline: handcrafted radiomics (Group 1), PCA on all radiomics (Group $2_{(i)}$), PCA after supervised preselection (Group $2_{(ii)}$) and ImageNet-pretrained ResNet-18 embeddings (Group 3). The analysis was conducted on two tasks, EGGIM level prediction (Dataset A) and dysplasia grade discrimination (Dataset B) using $P_k$ and a fixed RBF SVM has measurements of feature usefulness.

Feature selection indicates that pretrained DL embeddings capture a broad range of generic visual primitives cues that transfer well to EGD. While the very best radiomic descriptors are sometimes as, or even more, associated with the clinical outcome, the ResNet-18 representation supplies a longer tail of moderately informative dimensions, which explains its superior performance as the number of selected features increases. A key contextual factor is the imaging modality, NBI and WLE for Dataset A and Dataset B, respectively. Because the backbone was pretrained on ImageNet (white light images), the domain shift is larger for NBI, whose narrow spectral bands accentuate mucosal microvasculature and alter the color/contrast statistics of the images. It is therefore expected that Group 3 would face a harder transfer in Dataset A than in Dataset B, yielding smaller performance margins in A and stronger gains in B. This modality gap suggests that modest domain adaptation (e.g. modality-specific normalization or light fine-tuning) could further enhance deep-feature effectiveness. Interestingly, although the univariate association of radiomic descriptors exceeds that of deep embeddings in Dataset A, the multivariate predictive performance of the ResNet-18 embedding is superior. This apparent discrepancy stems from $P_k$ not measuring redundancy and inter-feature interactions. Since many

high $P_k$ radiomic features from Dataset A belong to correlated texture families, their information overlaps, meaning adding more of them yields diminishing returns. In contrast, Group 3 distributes the signal across multiple dimensions with moderate $P_k$ individually but complementary when combined. This behavior can also be seen in Dataset B, where the selection of the top 100 $P_k$ ranked features greatly limits the performance of Group 3. Figure 5.5 shows that when using the entire 512 descriptors, the classifier using the Resnet-18 embeddings matches the F1 score from Group 1.

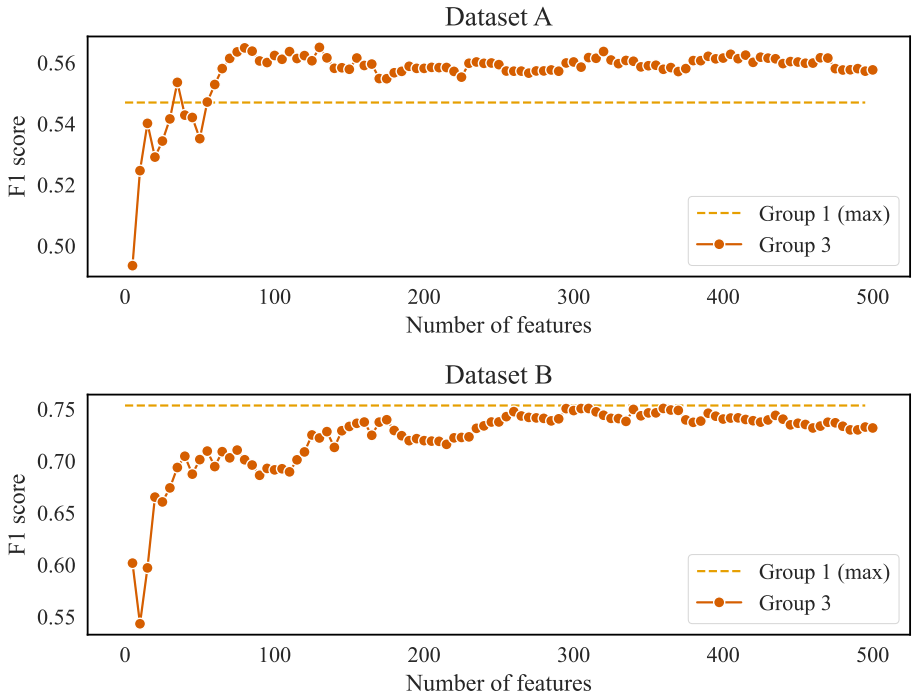


Figure 5.5: Average F1 score of the SVM classifier evaluated using 10-fold cross-validation, repeated 10 times. Comparison between the performance of Group 3 (Resnet-18 embeddings) using all the 512 descriptors with the maximum F1 score reached for Group 1 (radiomic features)

Across both tasks, PCA-based representations did not surpass the raw radiomics or the Resnet-18 features. In Dataset A, the highest $P_k$ radiomic descriptors cluster within texture families that contain strong correlations the PCA can compress into a few early components with strong association. By contrast, in Dataset B, the most informative descriptors are more diffusely distributed across feature families, so shared variance is weaker and the PCA cannot concentrate label-relevant signal in the leading axes. Figure 5.6 is consistent with this hypothesis, showing there is a mild negative association between component index and $P_k$ in Dataset A. Higher $P_k$ values cluster among the first 10 principal components the directions that, by construction, explain more variance and are enriched by texture derived variance, whereas later components sit close to chance. In Dataset B,

no clear monotonic trend is evident and the $P_k$ values are scattered across component indices for both PCA families, indicating that label association is not concentrated in the leading variance directions.

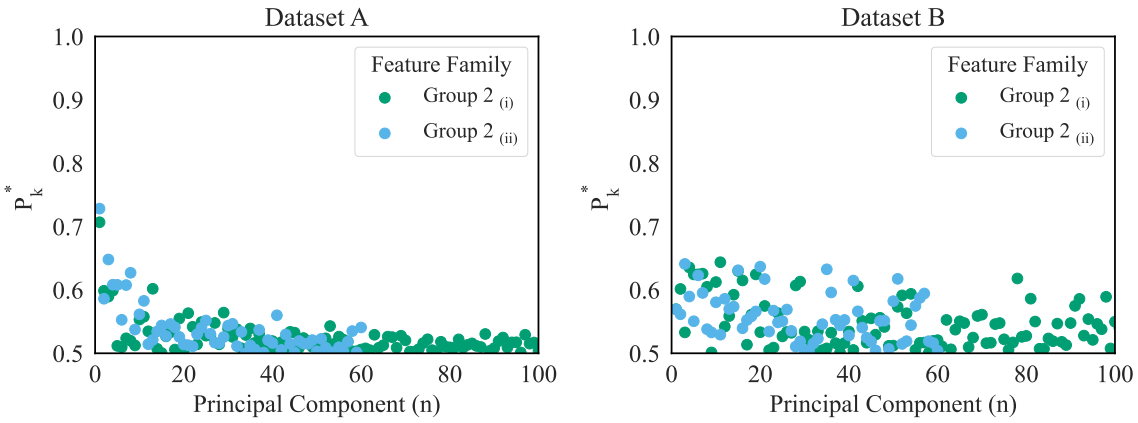


Figure 5.6: $P_k^*$ (see 3.17) across principal components for both PCA feature families: (i) with all radiomic features; (ii) with the preselected top-60 $P_k$-ranked features

Although individual principal components show only modest univariate association with the labels, the PCA representation still provides competitive SVM performance because the classifier depends on the joint geometry of the features. PCA orthogonalizes and concentrates the signal into a set of complementary directions, forming a configuration that the RBG kernel can exploit thanks to the reduced multicollinearity. Preselecting only the top $P_k$ features before applying the PCA can discard moderately informative but complementary descriptors, so some cross-feature relation the PCA and subsequently the SVM could exploit are lost. In Dataset A the highest $P_k$ features are overwhelming texture-based, resulting in a highly redundant preselected matrix. Consequently, PCA in Group $2_{(ii)}$ concentrates most variance into the early few components, hence the similar F1 in the first 20 features, but afterwards little label relevant variance remains and performance stagnates. On the contrary, Group $2_{(i)}$ retains a more diverse structure allowing it to increase performance further as more components are added. In Dataset B, the top $P_k$ features are more heterogeneous across families, therefore preselection preserves a richer mix of signals and Group $2_{(i)}$ and Group $2_{(ii)}$ perform more similarly.

In conclusion, the study validated the effectiveness of radiomic for EGD as an alternative approach to less transparent methodologies. As expected the barebones DL features delivered the best results, indicating strong transfer without any fine-tunning. PCA did not clearly surpass radiomics but still retained meaningful predictive value despite

lower univariate association. Radiomics nonetheless, provided a robust, compact baseline, with many informative descriptors that reached competitive performance, with clear interpretability.

# 6. Conclusions

## 6.1. Contributions

This thesis presents a first systematic evaluation of radiomics in upper GI endoscopy, framed around two clinically relevant problems, EGGIM and dysplasia. Additionally, the work benchmarked radiomics against semi-transparent (PCA) and opaque (ImageNet-pretrained ResNet-18) representations under a common pipeline.

Concretely, the work contributes with a reproducible, IBSI aligned radiomics pipeline for EGD, where using PyRadiomics, shape and texture descriptors were extracted from the images and then assessed for their univariate effectiveness using $P_k$ and multivariate performance with a fixed SVM under repeated, stratified cross-validation. Moreover, given the clinical relevance attached to color in the endoscopic domain, a HSV histogram was added to the standard radiomics features. For EGGIM on NBI images, texture features dominated the top $P_k$ ranks, while color carried limited signal. On the contrary, for dysplasia under WLE, sparse but informative HSV bins surfaced among the strongest individual descriptors, including the single highest $P_k$ feature. The low Jaccard similarity of 0.24 between the top-100 features across tasks highlights that predictive signatures differ between pathology and imaging modality rather than reflecting a single universal best subset. The strongest radiomic cues aligned with the known mucosal changes, which indicate texture heterogeneity for GIM and chromatic differences for dysplasia, supporting radiomics as a visually and biologically explainable bridge towards a ”visual biopsy”. The second part of the thesis provided a calibrated comparison across transparency levels. Bare-bones deep embeddings provided a long tail of moderately informative features, translated into the strongest SVM performance. Radiomic descriptors remained competitive with fewer, high $P_k$ features and delivered the most direct interpretability. PCA did not surpass the raw radiomics in either task but retained meaningful predictive value.

Taken together, these findings position radiomics as a viable complementary alternative to DL in EGD. The deep embeddings without any fine-tuning already outperformed the handcrafted descriptors, indicating the unlikeliness that radiomics can fully replace deep neural networks. Regardless, the competitive performance in addition to the higher degree of transparency motivates the development of hybrid models that reconcile performance with interpretability.

## 6.2. Implications for GI Imaging

There is undoubtedly a clinical significant signal within routine EGD frames. Even with small unbalanced datasets, heterogeneous devices, and different modalities, standard descriptors captured pathological-relevant phenomena. This suggests that quantitative image biomarkers can be extracted from routine upper GI images, potentially augmenting the visual assessment of endoscopists with consistent auditable evidence. The testing framework also highlighted the importance of task specificity. The results were not uniform across datasets, emphasizing that data acquisition protocol, image modality and clinical outcome are important aspects to consider to fully understand the meaning behind the features. In contrast to opaque DL approaches, radiomics provides named, standardized, and explainable features that map to visible clinical occurrences, improving trust and facilitating dialogue with pathology. In a field that is moving from proof-of-concept to real-world deployment, transparent predictors can assist already established and capable DL architectures to pass clinical scrutiny and integrate into guideline-driven pathways.

## 6.3. Future Research Direction

Current evidence is limited by small, class-imbalanced datasets from a single center and by reliance on two imaging modalities ( NBI for EGGIM and WLE for dysplasia). Larger multicenter cohorts with patient-level splits are needed to improve generalization and fairness. A cross-over design, which evaluated EGGIM on WLE and dysplasia on NBI would also help understand the effects of modality, task biology, and their interaction. Data preparation and feature extraction also warrant a stronger, domain-aware pipeline. Priorities include systematic image quality control and color normalization. Beyond a simple HSV histogram, the development of clinical aware color radiomics that better encapsulate aspects such as hemoglobin absorption in the vessels, would greatly improve the interpretability of the features and likely their effectiveness across tasks. Expanding on this, a deeper, domain-informed analysis of the descriptors is essential to explicitly map them to visible endoscopic phenomena. An analysis co-developed with endoscopists and pathologists to pursue clinical correlation at lesion and patient levels can elevate the descriptors into clinically actionable biomarkers that inform surveillance decisions.

# References

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229–263, Feb. 2024. [Cited on pages 1 and 4.]

[2] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, Feb. 2016. [Cited on pages 1 and 12.]

[3] H. Li, Y. Zhu, E. S. Burnside, K. Drukker, K. A. Hoadley, C. Fan, and M. L. Giger, "Quantitative mri radiomics in the prediction of molecular classifications of breast cancer subtypes in the tcga/tcia dataset," *NPJ Breast Cancer*, vol. 2, p. 16012, May 2016. [Cited on page 1.]

[4] J. W. T. Toh and R. B. Wilson, "Pathways of gastric carcinogenesis, helicobacter pylori virulence and interactions with antioxidant systems, vitamin c and phytochemicals," *International Journal of Molecular Sciences*, vol. 21, no. 17, p. 6451, Sep. 2020. [Cited on pages vii, 4, and 5.]

[5] L. Yang, X. Ying, S. Liu, G. Lyu, Z. Xu, X. Zhang, H. Li, Q. Li, N. Wang, and J. Ji, "Gastric cancer: Epidemiology, risk factors and prevention strategies," *Chinese Journal of Cancer Research*, vol. 32, no. 6, pp. 695–704, Dec. 2020. [Cited on page 5.]

[6] P. Malfertheiner, F. Megraud, T. Rokkas, J. P. Gisbert, J.-M. Liou, C. Schulz, A. Gasbarrini, R. H. Hunt, M. Leja, C. O'Morain, M. Rugge, S. Suerbaum, H. Tilg, K. Sugano, E. M. El-Omar, E. Helicobacter, and M. S. Group, "Management of *Helicobacter pylori* infection: the maastricht vi/florence consensus report," *Gut*, vol. 71, no. 9, pp. 1724–1762, Sep. 2022. [Cited on page 5.]

[7] K. Yashima, M. Shabana, H. Kurumi, K. Kawaguchi, and H. Isomoto, "Gastric cancer screening in japan: A narrative review," *Journal of Clinical Medicine*, vol. 11, no. 15, p. 4337, Jul. 2022. [Cited on page 5.]

[8] J. A. Ajani, T. A. D'Amico, D. J. Bentrem, C. U. Corvera, P. Das, P. C. Enzinger, T. Enzler, H. Gerdes, M. K. Gibson, P. Grierson, G. Gupta, W. L. Hofstetter, D. H. Ilson, S. Jalal, S. Kim, L. R. Kleinberg, S. Klempner, J. Lacy, B. Lee,

F. Licciardi, S. Lloyd, Q. P. Ly, K. Matsukuma, M. McNamara, R. P. Merkow, A. M. Miller, S. Mukherjee, M. F. Mulcahy, K. A. Perry, J. M. Pimiento, D. M. Reddi, S. Reznik, R. E. Roses, V. E. Strong, S. Su, N. Uboha, Z. A. Wainberg, C. G. Willett, Y. Woo, H. H. Yoon, N. R. McMillian, and M. Stein, "Gastric cancer, version 2.2025, nccn clinical practice guidelines in oncology," *J Natl Compr Canc Netw*, vol. 23, no. 5, pp. 169–191, May 2025. [Online]. Available: https://jnccn.org/view/journals/jnccn/23/5/article-p169.xml [Cited on pages 6 and 7.]

[9] P. Pimentel-Nunes, D. Libânio, R. Marcos-Pinto, M. Areia, M. Leja, G. Esposito, M. Garrido, I. Kikuste, F. Megraud, T. Matysiak-Budnik, B. Annibale, J.-M. Dumonceau, R. Barros, J.-F. Fléjou, F. Carneiro, J. E. van Hooft, E. J. Kuipers, and M. Dinis-Ribeiro, "Management of epithelial precancerous conditions and lesions in the stomach (maps ii): European society of gastrointestinal endoscopy (esge), european helicobacter and microbiota study group (ehmsg), european society of pathology (esp), and sociedade portuguesa de endoscopia digestiva (sped) guideline update 2019," *Endoscopy*, vol. 51, no. 4, pp. 365–388, Mar. 2019. [Online]. Available: https://www.esge.com/assets/downloads/pdfs/guidelines/2019_a_0859_1883.pdf [Cited on pages vii and 6.]

[10] T. Giandola, C. Maino, G. Marrapodi, M. Ratti, M. Ragusi, V. Bigiogera, C. Talei Franzesi, R. Corso, and D. Ippolito, "Imaging in gastric cancer: Current practice and future perspectives," *Diagnostics*, vol. 13, no. 7, p. 1276, Mar. 2023. [Cited on page 6.]

[11] Japanese Gastric Cancer Association, "Japanese gastric cancer treatment guidelines 2021 (6th edition)," *Gastric Cancer*, vol. 26, no. 1, pp. 1–25, Jan. 2023. [Cited on page 7.]

[12] ASGE Standards of Practice Committee, N. Forbes, S. E. Elhanafi, M. A. Al-Haddad, N. C. Thosani, P. V. Draganov, M. O. Othman, E. P. Ceppa, V. Kaul, M. M. Feely, I. Sahin, J. L. Buxbaum, A. H. Calderwood, J. M. Chalhoub, N. Coelho-Prabhu, M. Desai, L. L. Fujii-Lau, D. R. Kohli, R. S. Kwon, J. D. Machicado, N. B. Marya, S. Pawa, W. Ruan, S. G. Sheth, A. C. Storm, N. R. Thiruvengadam, and B. J. Qumseya, "American society for gastrointestinal endoscopy guideline on endoscopic submucosal dissection for the management of early esophageal and gastric cancers: summary and recommendations," *Gastrointestinal Endoscopy*, vol. 98, no. 3, pp. 271–284, Sep. 2023. [Cited on page 7.]

[13] L. Sisic, N. Crnovrsanin, H. Nienhueser, J.-O. Jung, S. Schiefer, G. M. Haag, T. Bruckner, M. Schneider, B. P. Müller-Stich, M. W. Büchler, and T. Schmidt, "Perioperative chemotherapy with 5-fu, leucovorin, oxaliplatin, and docetaxel (flot) for esophagogastric adenocarcinoma: ten years real-life experience from a surgical perspective," *Langenbeck's Archives of Surgery*, vol. 408, no. 1, p. 81, Feb. 2023. [Cited on page 7.]

[14] S. Marques, M. Bispo, P. Pimentel-Nunes, C. Chagas, and M. Dinis-Ribeiro, "Image documentation in gastrointestinal endoscopy: Review of recommendations," *GE Portuguese Journal of Gastroenterology*, vol. 24, no. 6, pp. 269–274, Nov. 2017. [Cited on page 7.]

[15] K. Yao, N. Uedo, M. Muto, H. Ishikawa *et al.*, "Development of an e-learning system for teaching endoscopists how to diagnose early gastric cancer: basic principles for improving early detection," *Gastric Cancer*, vol. 20, no. Suppl 1, pp. 28–38, Mar. 2017. [Cited on page 7.]

[16] R. Bisschops, M. Areia, E. Coron, D. Dobru, B. Kaskas, R. Kuvaev, O. Pech, K. Ragunath, B. Weusten, P. Familiari, D. Domagk, R. Valori, M. F. Kaminski, C. Spada, M. Bretthauer, C. Bennett, C. Senore, M. Dinis-Ribeiro, and M. D. Rutter, "Performance measures for upper gastrointestinal endoscopy: a european society of gastrointestinal endoscopy (esge) quality improvement initiative," *Endoscopy*, vol. 48, no. 9, pp. 843–864, Sep. 2016. [Cited on pages 7 and 8.]

[17] S. Y. Kim and J. M. Park, "Quality indicators in esophagogastroduodenoscopy," *Clinical Endoscopy*, vol. 55, no. 3, pp. 319–331, May 2022, epub 2022 May 16. [Online]. Available: https://www.e-ce.org/journal/view.php?doi=10.5946/ce.2022.094 [Cited on pages vii and 8.]

[18] X. Shu, G. Wu, Y. Zhang, Y. Wang, Y. Zheng, Q. Guo, R. Ji, and Y. Zhou, "Diagnostic value of linked color imaging based on endoscopy for gastric intestinal metaplasia: a systematic review and meta-analysis," *Annals of Translational Medicine*, vol. 9, no. 6, p. 506, Mar. 2021. [Cited on page 8.]

[19] J. Y. Park, "Image-enhanced endoscopy in upper gastrointestinal disease: focusing on texture and color enhancement imaging and red dichromatic imaging," *Clinical Endoscopy*, vol. 58, no. 2, pp. 163–180, Mar. 2025, epub 2024 Nov 6. [Cited on page 8.]

[20] Y. Kono, Y. Yamasaki, R. Takenaka, K. Takemoto, S. Fujiki, Y. Kawahara, H. Okada, K. Hori, S. Kawano, T. Miyake, and K. Yamamoto, "Chromoendoscopy of gastric adenoma using an acetic acid indigocarmine mixture," *World Journal of Gastroenterology*, vol. 20, no. 17, pp. 5092–5097, May 2014. [Cited on page 8.]

[21] O. Toyoshima and T. Nishizawa, "Kyoto classification of gastritis: Advances and future perspectives in endoscopic diagnosis of gastritis," *World Journal of Gastroenterology*, vol. 28, no. 43, pp. 6078–6089, Nov. 2022. [Online]. Available: https://www.wjgnet.com/1007-9327/full/v28/i43/6078.htm [Cited on page 9.]

[22] M. Banks, D. Graham, M. Jansen, T. Gotoda, S. Coda, M. di Pietro, N. Uedo, P. Bhandari, D. M. Pritchard, E. J. Kuipers, M. Rodriguez-Justo, M. R. Novelli, K. Ragunath, N. Shepherd, and M. Dinis-Ribeiro, "British society of gastroenterology guidelines on the diagnosis and management of patients at risk of gastric adenocarcinoma," *Gut*, vol. 68, no. 9, pp. 1545–1575, Sep. 2019. [Online]. Available: https://gut.bmj.com/content/68/9/1545 [Cited on page 9.]

[23] T. Rokkas and K. Ekmektzoglou, "Current role of narrow band imaging in diagnosing gastric intestinal metaplasia: a systematic review and meta-analysis of its diagnostic accuracy," *Annals of Gastroenterology*, vol. 36, no. 2, pp. 149–156, Mar. 2023. [Online]. Available: https://www.annalsgastro.gr/index.php/annalsgastro/article/view/6931 [Cited on page 9.]

[24] P. Pimentel-Nunes, M. Dinis-Ribeiro, J. B. Soares, R. Marcos-Pinto, C. Santos, C. Rolanda, R. P. Bastos, M. Areia, L. Afonso, J. Bergman, P. Sharma, T. Gotoda, R. Henrique, and L. Moreira-Dias, "A multicenter validation of an endoscopic classification with narrow band imaging for gastric precancerous and cancerous lesions," *Endoscopy*, vol. 44, no. 3, pp. 236–246, Mar. 2012, epub 2012 Jan 31. [Cited on page 10.]

[25] G. Esposito, P. Pimentel-Nunes, S. Angeletti, R. Castro, D. Libânio, G. Galli, E. Lahner, E. Di Giulio, B. Annibale, and M. Dinis-Ribeiro, "Endoscopic grading of gastric intestinal metaplasia (eggim): a multicenter validation study," *Endoscopy*, vol. 51, no. 6, pp. 515–521, Jun. 2019, epub 2018 Dec 21. [Cited on page 11.]

[26] T. Hirasawa, K. Aoyama, T. Tanimoto, S. Ishihara, S. Shichijo, T. Ozawa, T. Ohnishi, M. Fujishiro, K. Matsuo, J. Fujisaki, and T. Tada, "Application of artificial intelligence

using a convolutional neural network for detecting gastric cancer in endoscopic images," *Gastric Cancer*, vol. 21, no. 4, pp. 653–660, Jul. 2018, epub 2018 Jan 15. [Cited on page 12.]

[27] M. Ishioka, T. Hirasawa, and T. Tada, "Detecting gastric cancer from video images using convolutional neural networks," *Digestive Endoscopy*, vol. 31, no. 2, pp. e34–e35, Mar. 2019, epub 2018 Dec 21. [Cited on page 12.]

[28] E. Klang, A. Soroush, G. N. Nadkarni, K. Sharif, and A. Lahat, "Deep learning and gastric cancer: Systematic review of ai-assisted endoscopy," *Diagnostics*, vol. 13, no. 24, p. 3613, Dec. 2023. [Cited on page 12.]

[29] A. Seager, L. Sharp, L. J. Neilson, A. Brand, J. S. Hampton, T. J. W. Lee, R. Evans, L. Vale, J. Whelpton, N. Bestwick, C. J. Rees, and C.-D. trial team, "Polyp detection with colonoscopy assisted by the gi genius artificial intelligence endoscopy module compared with standard colonoscopy in routine colonoscopy practice (colo-detect): a multicentre, open-label, parallel-arm, pragmatic randomised controlled trial," *The Lancet Gastroenterology & Hepatology*, vol. 9, no. 10, pp. 911–923, Oct. 2024. [Cited on page 12.]

[30] M. H. J. Maas, H. Neumann, H. Shirin, L. H. Katz, A. A. Benson, A. Kahloon, E. Soons, R. Hazzan, M. J. Landsman, B. Lebwohl, S. K. Lewis, V. Sivanathan, S. Ngamruengphong, H. Jacob, and P. D. Siersema, "A computer-aided polyp detection system in screening and surveillance colonoscopy: an international, multicentre, randomised, tandem trial," *The Lancet Digital Health*, vol. 6, no. 3, pp. e157–e165, Mar. 2024. [Cited on page 12.]

[31] R. Kader, C. Hassan, Á. Lanas, and et al., "P146: Cloud-based artificial intelligence for detection of colorectal neoplasia — a randomized controlled trial (eagle trial)," *Gut*, vol. 74, no. Suppl 1, p. A176.1, Jun. 2025, bSG LIVE'25 conference abstract. [Cited on page 12.]

[32] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015. [Online]. Available: https://www.nature.com/articles/nature14539 [Cited on page 14.]

[33] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016, chapter 9: Convolutional Networks. [Online]. Available: https://www.deeplearningbook.org/ [Cited on pages 14, 15, and 16.]

[34] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," *BMC Medical Imaging*, vol. 22, no. 1, p. 69, Apr. 2022. [Online]. Available: https://bmcmedimaging.biomedcentral.com/articles/10.1186/s12880-022-00793-7 [Cited on page 16.]

[35] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. P. M. van Stiphout, P. Granton, C. M. L. Zegers, R. Gillies, R. Boellard, A. Dekker, and H. J. W. L. Aerts, "Radiomics: extracting more information from medical images using advanced feature analysis," *European Journal of Cancer*, vol. 48, no. 4, pp. 441–446, 2012. [Online]. Available: https://doi.org/10.1016/j.ejca.2011.11.036 [Cited on page 17.]

[36] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Communications*, vol. 5, p. 4006, 2014. [Online]. Available: https://doi.org/10.1038/ncomms5006 [Cited on page 17.]

[37] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. W. L. Aerts, "Pyradiomics: Open-source python package for the extraction of radiomics features from medical imaging," https://github.com/Radiomics/pyradiomics, 2017, version 3.1.0. Available at https://pyradiomics.readthedocs.io/. [Cited on page 18.]

[38] ——, "Computational radiomics system to decode the radiographic phenotype," *Cancer Research*, vol. 77, no. 21, pp. e104–e107, 2017. [Online]. Available: https://doi.org/10.1158/0008-5472.CAN-17-0339 [Cited on page 18.]

[39] W. D. Smith, R. C. Dutton, and N. T. Smith, "Measuring the performance of anesthetic depth indicators," *Anesthesiology*, vol. 84, no. 1, pp. 38–51, 1996. [Online]. Available: https://doi.org/10.1097/00000542-199601000-00005 [Cited on page 20.]

[40] J.-O. Kim, "Predictive measures of ordinal association," *American Journal of Sociology*, vol. 76, no. 5, pp. 891–907, 1971. [Online]. Available: https://doi.org/10.1086/224993 [Cited on page 20.]

[41] J.-G. Lee, S. Jun, Y.-W. Cho, H. Lee, G. B. Kim, J. B. Seo, and N. Kim, "Deep learning in medical imaging: General overview," *Korean Journal of Radiology*, vol. 18, no. 4, pp. 570–584, 2017. [Online]. Available: https://doi.org/10.3348/kjr.2017.18.4.570 [Cited on page 22.]

[42] S. Shichijo, S. Nomura, K. Aoyama, Y. Nishikawa, M. Miura, T. Shinagawa, H. Takiyama, T. Tanimoto, S. Ishihara, K. Matsuo, and T. Tada, "Application of Convolutional Neural Networks in the diagnosis of Helicobacter pylori infection based on endoscopic images," *EBioMedicine*, vol. 25, pp. 106–111, Nov. 2017. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29056541/ [Cited on page 23.]

[43] L. Li, Y. Chen, Z. Shen, X. Zhang, J. Sang, Y. Ding, X. Yang, J. Li, M. Chen, C. Jin, C. Chen, and C. Yu, "Convolutional neural network for the diagnosis of early gastric cancer based on magnifying narrow band imaging," *Gastric Cancer*, vol. 23, no. 1, pp. 126–132, Jan. 2020. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31332619/ [Cited on page 23.]

[44] D. Luo, F. Kuang, J. Du, M. Zhou, X. Liu, X. Luo, Y. Tang, B. Li, and S. Su, "Artificial intelligence–assisted endoscopic diagnosis of early upper gastrointestinal cancer: A systematic review and meta-analysis," *Frontiers in Oncology*, vol. 12, p. 855175, Jun. 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fonc.2022.855175 [Cited on page 23.]

[45] M. Cai, B. Song, Y. Deng, P. Gao, S. Cai, A. Yalikong, E. Xu, Y. Zhong, J. Yu, and P. Zhou, "Automatically optimized radiomics modeling system for small gastric submucosal tumor (<2 cm) discrimination based on EUS images," *Gastrointestinal Endoscopy*, vol. 99, no. 4, pp. 537–547.e4, Apr. 2024, epub 2023 Nov 11. [Cited on page 23.]

[46] Z. HajiEsmailPoor, P. Tabnak, B. Baradaran, F. Pashazadeh, and L. Aghebati-Maleki, "Diagnostic performance of CT scan-based radiomics for prediction of lymph node metastasis in gastric cancer: a systematic review and meta-analysis," *Frontiers in Oncology*, vol. 13, p. 1185663, oct 2023. [Cited on page 23.]

[47] D. Adili, A. Mohetaer, and W. Zhang, "Diagnostic accuracy of radiomics-based machine learning for neoadjuvant chemotherapy response and survival prediction in gastric cancer patients: A systematic review and meta-analysis," *European Journal of Radiology*, vol. 173, p. 111249, Apr. 2024, epub 2023 Dec 5. [Cited on page 24.]

[48] S. C. Shah, A. J. Gawron, R. A. Mustafa, and M. B. Piazuelo, “Histologic subtyping of gastric intestinal metaplasia: Overview and considerations for clinical practice,” *Gastroenterology*, vol. 158, no. 3, pp. 745–750, Feb. 2020. [Cited on page 33.]

[49] S. Paghadhar, M. Jain, B. Mahadevan, and J. Venkataraman, “Interpretation of benign gastric mucosal lesions using narrow-band imaging,” *Journal of Digestive Endoscopy*, vol. 11, no. 2, pp. 106–111, Jun. 2020. [Cited on page 33.]

[50] Y. Cao, D. Wang, G. Mo, Y. Peng, and Z. Li, “Gastric precancerous lesions: Occurrence, development factors, and treatment,” *Frontiers in Oncology*, vol. 13, p. 1226652, Aug. 2023. [Cited on page 34.]

[51] M. Fornasarig, A. Capuano, S. Maiero, E. Pivetta, G. Guarnieri, V. Canzonieri, A. Zucchetto, M. Mongiat, R. Cannizzaro, and P. Spessotto, “pcle highlights distinctive vascular patterns in early gastric cancer and in gastric diseases with high risk of malignant complications,” *Scientific Reports*, vol. 11, no. 1, p. 21053, Oct. 2021. [Cited on page 34.]

# A. Appendix

## A.1. PyRadiomics Features Description

- **First-order features (intensity histogram):** summarize the distribution of pixel and voxel intensity without considering spatial relationships. Clinically, these features reflect aspects such as brightness or heterogeneity in the mucosa tissue.
  - Energy: sum of squared intensities, reflecting magnitude.
  - Entropy: randomness/heterogeneity of intensity distribution.
  - Mean: average intensity value.
  - Median: middle intensity value.
  - Variance: dispersion of intensities.
  - Standard deviation: spread around the mean.
  - Skewness: asymmetry of intensity distribution.
  - Kurtosis: peakedness/flatness relative to normal distribution.
  - Minimum/Maximum: extreme intensity values.
  - Range: difference between max and min.
  - 10th/90th percentile: intensity cutoffs.
  - Interquartile range: middle 50% spread.
  - Mean absolute deviation: average deviation from the mean.
  - Robust mean absolute deviation: deviation within central range (excludes outliers).
  - Root mean squared: quadratic mean of intensities.
  - Uniformity: measure of distribution uniformity (inverse of entropy).
- **Shape features (2D/3D):** describes the geometry of the segmented area. Supports 2D and 3D images with the features varying according to the dimension. Quantifies aspects that clinicians already visually assess intuitively, such as irregularities in the form of the lesion.
  - 2D area / 3D volume: size of the ROI.
  - Perimeter / surface area: boundary length/area.

  - Major/Minor axis length: longest and shortest diameters.
  - Elongation: degree to which shape is stretched.
  - Flatness: degree of compression along one axis.
  - Sphericity: how close the shape is to a sphere.
  - Compactness (1 & 2): compactness relative to sphere.
  - Roundness (2D): how circular the region is.
  - Rectangularity: ratio of shape area to bounding box.
  - Maximum 2D diameter: largest distance across the ROI.

- **Texture features (spatial gray-level organization):** captures the arrangement of intensities within the image, reflecting heterogeneity and granularity. Pyradiomics implements five major families:

  - **Gray-Level Co-occurrence Matrix (GLCM):** measures how often pairs of gray levels occur in a specific spatial relationship.
    * Contrast: intensity difference between neighbors.
    * Correlation: linear dependency between neighbors.
    * Homogeneity: closeness of distribution to diagonal (low contrast).
    * Energy (Angular Second Moment): uniformity of GLCM.
    * Entropy: randomness of co-occurrence patterns.
    * Dissimilarity: average absolute difference between pairs.
    * ASM (angular second moment): squared elements sum, uniformity.
    * Cluster shade/prominence: skewness/peakedness of clusters.
    * Maximum probability: strongest co-occurrence.
    * Inverse difference (IDM): local homogeneity.
  - **Gray-Level Run-Length Matrix (GLRLM):** quantifies consecutive runs of a pixel with the same gray level.
    * Short-run emphasis (SRE), Long-run emphasis (LRE): prevalence of short or long runs.
    * Gray-level non-uniformity (GLN), Run length non-uniformity (RLN): variability of gray levels or run lengths.
    * Low gray-level run emphasis (LGRE), High gray-level run emphasis (HGRE): distribution of low or high intensity runs.

  - SRLGE, SRHGE, LRLGE, LRHGE: combined metrics for run length and intensity.
  - Run percentage (RP): proportion of runs relative to ROI size.
- **Gray-Level Size Zone Matrix (GLSZM):** captures the size of the distribution of connected regions with the same gray level.
  - Small area emphasis (SAE), Large area emphasis (LAE): weight of small or large zones.
  - Gray-level non-uniformity (GLN), Zone size non-uniformity (ZSN): variability of gray levels or zone sizes.
  - Zone percentage (ZP): number of zones relative to ROI size.
  - LGZE, HGZE: emphasis on low or high gray-level zones.
  - SALGLE, SAHGLE, LALGLE, LAHGLE: combined metrics for zone size and intensity.
- **Gray-Level Dependence Matrix (GLDM):** describes the number of connected voxels that depend on a central voxels gray level.
  - Small dependence emphasis (SDE), Large dependence emphasis (LDE): Prevalence of small or large dependencies.
  - Gray-level non-uniformity (GLN), Dependence non-uniformity (DN): Variability across gray levels or dependence sizes.
  - Dependence entropy: Randomness of dependence distributions.
  - Low gray-level emphasis (LGL), High gray-level emphasis (HGL): Weight of low or high gray-level dependencies.
  - SDLGLE, SDHGLE, LDLGLE, LDHGLE: Combined metrics of dependence size and intensity.
- **Neighborhood Gray-Tone Difference Matrix (NGTDM):** characterizes the difference between a voxels gray level and the average gray level of its neighbors.
  - Coarseness: inverse of spatial rate of change (smoothness).
  - Contrast: local intensity variation strength.
  - Busyness: rate of intensity change between neighbors.
  - Complexity: intensity differences weighted by frequency.
  - Strength: overall contrast weighted by neighborhood frequency.