Politecnico di Torino

Politecnico di Torino

Computer Engineering - Artificial Intelligence and Data Analytics

A.a. 2025/2026

Graduation Session March 2026

Domain Adaptation for AI-Based Endoscopic Gastric Risk Stratification

Relatori:

Giuseppe Bruno Averta

Miguel Coimbra

Candidati:

Claudia Maggiulli

Abstract

The classification of gastric mucosal inflammation (IM) is a critical task in gastroenterology, as accurate identification of gastric conditions is essential for diagnosis and treatment. However, medical imaging data often exhibit significant variability due to differences in image quality, equipment used, and clinical practices across datasets. This variability presents a challenge for developing models that can generalize well across different clinical environments, making it difficult to apply models trained in one domain to another without significant performance degradation.
Domain adaptation (DA) provides a solution to this challenge by allowing models to adapt to new domains without requiring labeled data from the target domain. DA is particularly valuable in medical imaging, where annotated datasets can be scarce, and the quality of data may differ between clinical settings. The use of domain adaptation techniques enables the transfer of models from controlled environments to routine clinical practice, addressing the issue of domain shift and improving model generalization.
To tackle this problem, the EGGIM (Endoscopic Gastric Inflammation Grading) score is used as a quantitative measure of gastric inflammation. The EGGIM score ranges from 0 (normal mucosa) to 2 (severe inflammation) and is computed at the patient level by aggregating the scores across five anatomical regions of the stomach. This scoring system is essential for identifying high-risk patients and assessing the severity of inflammation in a standardized manner.
Several domain adaptation (DA) techniques are employed, including self-supervised learning methods such as pseudo-labeling, which maximize the use of available labeled data. Additionally, Domain Adversarial Neural Networks (DANN) are applied, where the network is trained to distinguish between domains while simultaneously learning to classify images. This approach enables the model to learn domain-invariant features, improving its ability to generalize across different datasets. These methods, combined with the EGGIM score, provide an effective solution to the challenge of domain adaptation in gastric inflammation classification.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduzione

1.1 Artificial Intelligence in Medicine

Artificial intelligence (AI), and in particular deep learning techniques, in recent years have become increasingly prominent in the medical field. This growth has been driven by the progressive digitalization of clinical data and by significant advances in computational capabilities. The adoption of deep neural network–based models has enabled the development of automated methods for medical image analysis, clinical signal processing, and decision support systems, contributing to improvements in both the efficiency and accuracy of diagnostic workflows.
Across multiple application domains, artificial intelligence systems have demonstrated strong performance, in some cases comparable to that of human specialists under controlled experimental conditions. These results have fostered considerable interest in the integration of AI-based tools into clinical practice, with the aim of supporting healthcare professionals, reducing inter-operator variability, and promoting more objective and reproducible decision-making processes. [1].
Among the areas in which AI has found the widest application, medical imaging plays a central role. Diagnostic modalities such as radiology, digital histopathology, and endoscopy generate large volumes of visual data, making them particularly well suited to data-driven approaches. In this context, deep learning models have proven effective in the automatic recognition of complex visual patterns associated with pathological conditions, often capturing subtle features that are difficult to quantify through purely manual or subjective assessment.
Despite these promising results, the large-scale clinical adoption of artificial intelligence systems remains constrained by several critical challenges. In particular, deep learning models applied to medical imaging are strongly dependent on the data distribution used during training. Variations in acquisition devices, clinical protocols, patient populations, and annotation procedures can introduce substantial

changes in data characteristics, thereby limiting the ability of models to generalize effectively to new and unseen contexts.
As a consequence, models that achieve high performance on specific datasets may experience significant performance degradation when applied to data originating from different domains, reducing their reliability in real-world clinical settings. This limited generalization capability represents one of the main barriers to the clinical deployment of artificial intelligence systems and highlights the need for methodological approaches that ensure greater robustness and transferability of learned models.

1.2 Application Context: Gastric Endoscopy and the EGGIM Score

Gastric endoscopy is a fundamental diagnostic tool for the clinical assessment of gastric diseases, enabling direct visualization of the gastric mucosa and the identification of inflammatory and pre-neoplastic conditions. Accurate evaluation of gastric inflammation plays a crucial role in clinical practice, as chronic inflammatory processes are closely associated with the development of more severe pathologies and represent a key factor for patient risk stratification and clinical management. In particular, atrophic gastritis and intestinal metaplasia are recognized as important stages in the progression toward gastric cancer. Early detection of these conditions during endoscopic examination is therefore essential for the implementation of appropriate surveillance programs and timely therapeutic interventions. However, endoscopic assessment of these conditions is often challenging, as mucosal alterations may be subtle and subject to considerable inter-observer variability, even among experienced endoscopists.
To mitigate these limitations, standardized endoscopic scoring systems have been introduced to support a more objective evaluation of gastric mucosal changes. Among these, the Endoscopic Grading of Gastric Intestinal Metaplasia (EGGIM) score is a clinically validated tool designed to assess the presence and extent of intestinal metaplasia based on endoscopic imaging. The EGGIM system assigns a score to predefined gastric regions according to characteristic visual patterns, enabling a structured and reproducible assessment. Within automated analysis frameworks, the score can be estimated at the level of localized image patches, allowing for fine-grained modeling of regional mucosal alterations.
Despite the advantages offered by standardized scoring systems, gastric endoscopy represents an inherently complex multi-center setting. Variability in endoscopic equipment, imaging modalities, acquisition protocols, and operator expertise introduces substantial heterogeneity in the acquired data. This multi-center variability poses significant challenges for the development of robust and reliable artificial

intelligence models, as differences across clinical sites can strongly affect model performance and limit generalization to unseen domains. [2]

1.3 Objectives and Contributions

The main objective of this thesis is to analyze the generalization capabilities of deep learning models for the classification of gastric intestinal metaplasia in endoscopic images, with particular attention to the challenges arising from data heterogeneity across different clinical settings. The work aims to investigate the factors that contribute to limited model generalization and to better understand the sources of difficulty in endoscopic image–based classification.
To address these challenges, this thesis explores the application of domain adaptation techniques to improve model robustness when transferring knowledge between different domains. The proposed methods are evaluated on datasets acquired under varying imaging conditions, enabling a systematic analysis of cross-domain performance and its potential implications for clinical use.

1.4 Outline of the Work

This thesis is structured as follows. It begins with a review of the state of the art concerning the application of deep learning to gastric endoscopic image classification and the main domain adaptation approaches proposed for endoscopic imaging. This is followed by an introduction to the medical and clinical background necessary to contextualize the application domain, with particular attention to gastric anatomy, relevant pathologies, and the EGGIM scoring system.
The core part of the work is devoted to the analysis of the generalization problem and domain shift, first addressed from a general perspective and subsequently examined within the specific context of gastric endoscopy, in order to motivate the methodological choices adopted. The datasets employed in this study, together with the data acquisition procedures, annotation strategies, and preprocessing methods, are then described.
Finally, the proposed methodologies and experimental results are presented, including comparisons with baseline approaches and an analysis of cross-domain performance. The thesis concludes with a discussion of the main findings, a summary of the contributions, and an outline of possible directions for future work.

Chapter 2

State of the Art

2.1 Deep Learning for Gastric Endoscopic Classification

Deep learning has emerged as a key enabling technology for the analysis of gastric endoscopic images, with the aim of supporting the automated recognition of inflammatory, atrophic, and pre-neoplastic conditions, including intestinal metaplasia. Among deep learning methods, convolutional neural networks (CNNs) are particularly well suited to this task, as they can learn discriminative visual patterns directly from raw images, even when mucosal alterations are subtle and difficult to interpret consistently.

Early evidence of the clinical potential of CNN-based systems in gastrointestinal endoscopy was provided by Hirasawa et al. [3], who developed an automated detector for gastric cancer. The model was trained on 13,584 endoscopic images and evaluated on 2,296 images from 69 patients, including 77 cancer lesions. The system processed the entire test set in 47 seconds and achieved a sensitivity of 92.2% (71/77 detected lesions). However, the positive predictive value was limited to 30.6%, as 161 non-cancerous regions were incorrectly classified as malignant [3]. This study demonstrated the feasibility of large-scale automated detection while also highlighting the importance of addressing false positives in real-world clinical deployment.

Following these initial results in cancer detection, subsequent research shifted focus toward the identification of precancerous gastric conditions, where visual patterns are often even less evident. Ligato et al. [4] proposed a patch-based CNN approach for recognizing intestinal metaplasia in the gastric corpus, using 200×200 pixel image patches and a ResNet-50 backbone. At the patch level, the model achieved an accuracy of 74%, with a precision of 76% and a recall of 72% [4]. Patch-level predictions were subsequently aggregated using a voting

strategy to produce image-level classifications, yielding test accuracies of up to 78%. In certain configurations, the system reached very high recall values (up to 100%), at the expense of reduced precision (approximately 70%) [4]. This two-stage strategy reflects a common paradigm in endoscopic AI, where local analysis enhances sensitivity and aggregation mechanisms provide clinically meaningful outputs.

A broader perspective on the field is offered by survey and review studies. Xin et al. [5] provide an overview of artificial intelligence applications in gastrointestinal endoscopic oncology, covering tasks such as lesion detection, characterization, and decision support. They emphasize several open challenges, including dataset shift, inter-observer variability in annotations, and the need for large-scale prospective validation. From a methodological standpoint, most existing systems rely on established CNN architectures, such as VGG-like networks and ResNet variants, and frequently employ transfer learning from large-scale natural image datasets (e.g., ImageNet) to mitigate the scarcity of annotated medical data. Patch-based processing is also widely adopted for fine-grained diagnostic tasks, as highlighted in systematic reviews of AI-based endoscopic diagnosis [6].

A closely related research direction involves the automated assessment of standardized endoscopic scoring systems. In particular, the Endoscopic Gastric Inflammation Grading Index (EGGIM), proposed and clinically validated by Esposito et al. [7], provides a structured framework for estimating the extent of gastric inflammation and intestinal metaplasia through the evaluation of predefined anatomical regions. Within this context, Martins et al. [8] investigated whether small, appropriately selected patches can predict the EGGIM grade of an entire still frame, motivated by a "self-similarity" hypothesis for the mucosal structural changes observed in gastric intestinal metaplasia. Using a ResNet-50 model and a leave-one-patient-out cross-validation protocol, they reported that patch-based predictions could correctly stratify the risk of 57 out of 65 patients, achieving perfect sensitivity on an extremely biased dataset [8]. These findings strengthen the rationale for patch-based learning not only for lesion detection, but also for estimating standardized endoscopic scores (and thus surveillance-relevant risk categories), while suggesting that careful sampling/selection of informative regions may be as important as the network architecture itself.

More recent studies have explored gastric cancer risk stratification using multi-modal learning approaches. Ma et al. [9] introduced an attention-based model (Attention-GT) that integrates gastroscopic images with tongue images and basic clinical indicators, including age, gender, and *H. pylori* infection status. Evaluated on a longitudinal cohort of 384 participants, the model achieved an AUC of 0.83 for distinguishing precancerous lesions of gastric cancer and an AUC of 0.84 for identifying disease progression among baseline non-PLGC patients [9]. Despite these promising results, both multi-modal approaches and purely image-based

classifiers often exhibit limited generalization across different clinical centers and imaging settings.
Overall, the literature demonstrates that deep learning represents a promising solution for gastric endoscopic classification and precancerous risk assessment. Nevertheless, most reported results are obtained on controlled and relatively homogeneous datasets. This limitation motivates the development of methods aimed at improving robustness and generalization, particularly in multi-center scenarios, providing the rationale for the domain adaptation techniques.

2.2 Domain Adaptation in Medical and Endoscopic Imaging

One of the main obstacles to the clinical translation of deep learning models in medical imaging is represented by the phenomenon of *domain shift*, namely the statistical discrepancy between the data distribution used during the training phase and that encountered during real-world deployment. Numerous studies have shown that models trained and validated on controlled datasets may experience a significant degradation in performance when applied to data originating from different clinical environments, with substantial reductions in accuracy, sensitivity, and area under the ROC curve [10, 11].
In medical imaging, domain shift exhibits characteristics that distinguish it from traditional computer vision tasks. Major sources of variability include differences in acquisition devices, clinical protocols, hardware and software sensor settings, patient population composition, and annotation procedures. In endoscopic imaging, these issues are further amplified by the strong dependence of image appearance on operator-related factors, such as viewing angle, distance from the mucosa, illumination conditions, and the quality of patient preparation. Differences among endoscopic devices from different manufacturers, across clinical centers adopting non-standardized acquisition protocols, and in post-processing procedures contribute to significant variations in the distribution of visual features, thereby compromising model generalization [2].
The impact of domain shift and limited generalizability in gastrointestinal endoscopy has been repeatedly observed in practice, motivating evaluation protocols that include external test sets and multicentre settings.
To address the domain shift problem, the literature has introduced two main paradigms: *Domain Adaptation* and *Domain Generalization*. In Domain Adaptation, the model has access during training to data from the target domain, typically unlabeled, with the goal of reducing the discrepancy between source and target domain distributions. In contrast, Domain Generalization aims to learn robust representations using only source-domain data, without any prior information about

the target domain. Although Domain Generalization represents a more ambitious objective, Domain Adaptation is particularly relevant in the medical domain, where unlabeled data from new clinical centers are often available [11].
From a methodological perspective, many Domain Adaptation techniques rely on feature-level representation alignment. In this context, the adversarial learning framework introduced by Ganin and Lempitsky [12] represented a major breakthrough, enabling the learning of domain-invariant features through the introduction of a domain discriminator and a gradient reversal mechanism. This approach, known as the Domain-Adversarial Neural Network (DANN), has been widely adopted in medical imaging and has demonstrated consistent improvements in unsupervised adaptation scenarios.
Alongside adversarial approaches, methods based on self-training and pseudo-labeling have been proposed, in which the model assigns pseudo-labels to unlabeled target data and iteratively uses them to refine training. These strategies can be particularly effective when strong structural correlations exist between source and target domains, but they may be sensitive to noise in the pseudo-labels. In gastrointestinal endoscopy, related source-free / self-training-style domain adaptation approaches have been investigated, for example for polyp detection, to adapt models without target labels while improving robustness to domain shift [13]. Image-level adaptation approaches have also been explored, often relying on generative adversarial networks; however, in medical imaging such methods may introduce visual artifacts that are potentially incompatible with clinical interpretation.
The application of Domain Adaptation to gastrointestinal endoscopy is a relatively recent but rapidly growing research area. Kim et al. [2] review the main domain adaptation strategies proposed for gastrointestinal endoscopy and medical imaging, highlighting that adversarial feature learning and image translation approaches can help mitigate domain shift and improve cross-center robustness. Overall, their analysis emphasizes the need for careful validation across devices and hospitals, and for adaptation methods that preserve clinically relevant features while reducing domain-specific biases.
More recently, the research community has shifted attention toward architectures based on Vision Transformers, which have demonstrated a superior ability to model global relationships compared to traditional CNNs. Alijani et al. [14] showed that Vision Transformers exhibit increased robustness to domain shift in both Domain Adaptation and Domain Generalization settings, reporting average improvements in AUC and F1-score over convolutional backbones across different application contexts. Nevertheless, in medical and endoscopic imaging, the high computational cost and the requirement for large amounts of annotated data remain significant barriers to their widespread clinical adoption.
Overall, the literature agrees that domain adaptation constitutes a key component for the effective transfer of deep learning models from experimental settings to

clinical practice. The most promising approaches combine adversarial learning, self-supervised strategies for exploiting unlabeled data, and architectures capable of integrating both local and global information. In the context of gastric endoscopic classification, integrating Domain Adaptation techniques with standardized scoring systems such as EGGIM appears particularly relevant, as it enables improved model robustness while preserving clinical interpretability. These considerations form the methodological foundation of the work presented in this thesis.

Chapter 3

Medical and Clinical Background

3.1 Anatomy and Physiology of the Stomach

The stomach is a hollow organ of the upper gastrointestinal tract with fundamental roles in the mechanical and chemical digestion of food. From an anatomical point of view, it extends between the esophagus and the duodenum and is characterized by a complex structure adapted to perform secretory, motor, and barrier functions. The gastric wall is composed of multiple layers, among which the gastric mucosa represents the most important component. The mucosa is the innermost layer and is directly observable during endoscopic examination. It consists of a columnar secretory epithelium, lamina propria, and muscularis mucosae, and it is responsible for the secretion of mucus, hydrochloric acid, and digestive enzymes. From a clinical and endoscopic perspective, the mucosa constitutes the primary site of interest, as it is the location where inflammatory, atrophic, and metaplastic alterations develop.
From an anatomo-functional perspective, the stomach is divided into different regions, each characterized by specific histological and functional properties. For the purposes of endoscopic evaluation and the application of the EGGIM score, five anatomical regions are considered:

- **Antrum LC**: the distal portion of the stomach along the lesser curvature, located proximal to the pylorus and characterized by a predominantly mucous glandular mucosa. This region is frequently affected by inflammatory processes and represents a common site of *Helicobacter pylori* infection [15].
- **Antrum GC**: the distal gastric region along the greater curvature, presenting morphological characteristics that may differ from those observed along the

lesser curvature.

- **Incisura**: the transition area between the antrum and the gastric body, anatomically identifiable as the angular incisure. This region has high clinical relevance, as it is often an early site of atrophic and metaplastic changes.

- **Body LC**: the portion of the gastric body along the lesser curvature, characterized by oxyntic glands responsible for acid secretion. The mucosa in this region exhibits morphological features that differ from those of the antrum.

- **Body GC**: the portion of the gastric body along the greater curvature, also characterized by oxyntic mucosa with a predominantly secretory function. Inflammatory and atrophic alterations may also occur in this region, particularly in advanced stages of disease.

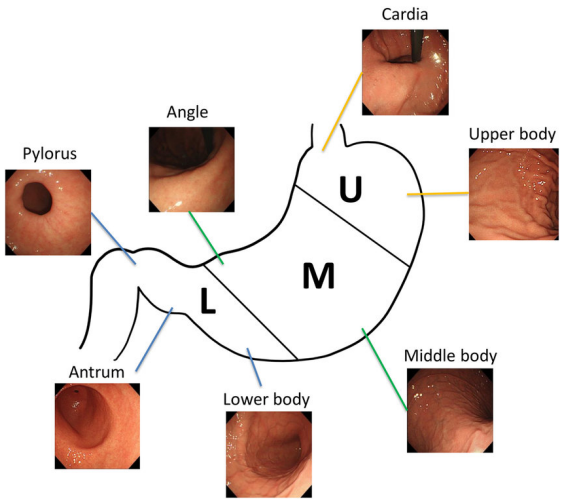


Figure 3.1: Anatomy of the Stomach

The anatomical subdivision of the stomach into these regions enables a systematic and standardized endoscopic assessment of the gastric mucosa. This approach is fundamental for the identification and quantification of pathological alterations and constitutes the anatomical basis upon which endoscopic scoring systems, such as the EGGIM score, are built, as described in the following sections.

3.2 Gastric Diseases and Inflammation

Gastric diseases encompass a wide spectrum of inflammatory and pre-neoplastic conditions affecting the gastric mucosa, which are clinically relevant due to their

potential progression toward more severe pathological states. Among these conditions, gastritis, intestinal metaplasia, and dysplasia represent successive stages of mucosal alteration and are closely associated with an increased risk of gastric cancer.
Gastritis can occur in acute or chronic forms. Acute gastritis is typically characterized by transient inflammation of the gastric mucosa and is often related to irritative factors such as medications or chemical agents. Chronic gastritis, which is more relevant from a clinical standpoint, is a persistent inflammatory condition that may lead to long-term structural changes of the gastric mucosa. Major etiological factors include *Helicobacter pylori* infection, prolonged use of non-steroidal anti-inflammatory drugs, smoking, and environmental influences.
Intestinal metaplasia is defined by the replacement of normal gastric epithelium with intestinal-type epithelium. It is recognized as a pre-neoplastic condition and is associated with an increased risk of gastric carcinoma. From a clinical perspective, intestinal metaplasia can be classified into different subtypes based on histological features, which are associated with different levels of cancer risk. Its detection and assessment are therefore essential for patient stratification and surveillance.
As mucosal alterations progress, dysplasia may develop, characterized by more pronounced cellular and architectural abnormalities. Dysplasia represents an advanced stage in the pathological continuum and is considered a significant risk factor for the development of gastric carcinoma, requiring careful clinical monitoring.
Gastric carcinoma represents the final outcome of this pathological progression and remains one of the leading causes of cancer-related mortality worldwide. Prognosis is strongly dependent on the stage at diagnosis, highlighting the importance of early identification of pre-neoplastic conditions. Relevant prognostic factors include tumor stage, histological differentiation, and the presence of metastatic disease.
A key role in the development of many of these gastric conditions is played by *Helicobacter pylori* infection, which is widely recognized as a major etiological factor in chronic gastric inflammation. Persistent infection is associated with progressive structural and functional alterations of the gastric mucosa, increasing the risk of pre-neoplastic and neoplastic transformations. These progressive changes provide the clinical basis for endoscopic screening and surveillance strategies aimed at identifying patients at increased risk [16].

3.3 Gastric Endoscopy: Techniques and Procedures

As previously discussed, various gastric pathologies, including chronic gastritis, intestinal metaplasia, dysplasia, and gastric cancer, are often identified through endoscopic examination. The ability to detect and monitor these conditions in

a timely and accurate manner is crucial for effective diagnosis and treatment. Endoscopy, specifically gastroscopy, remains the primary method for assessing the gastric mucosa and diagnosing such conditions. It provides direct visualization of the mucosal surface, allowing for the identification of subtle abnormalities that may not be evident through other diagnostic methods.

Conventional gastroscopy (white light endoscopy) remains the most widely used method for visualizing the gastric mucosa. It involves the use of a flexible endoscope with a white light source, which provides detailed images of the gastric lining. The procedure is performed using standard equipment and protocols, ensuring a consistent approach for routine examination and diagnosis.

In addition to conventional endoscopy, advanced imaging techniques have been developed to enhance the diagnostic capabilities of endoscopy. Among these, Narrow Band Imaging (NBI) utilizes specific wavelengths of light to highlight mucosal changes and improve the visualization of vascular patterns, making it particularly useful for detecting early-stage lesions. Chromoendoscopy involves the application of special dyes to the gastric mucosa to enhance contrast and identify subtle abnormalities. Magnification endoscopy allows for a closer examination of the mucosal surface, facilitating the detection of microscopic changes in the epithelium.

Image acquisition protocols are essential to ensure the consistency and quality of endoscopic images. International guidelines are available to standardize procedures, including the positioning of the endoscope, image documentation, and image processing techniques. Systematic photographic documentation is also critical to allow for comparison over time and provide accurate records for clinical decision-making.

However, there is significant variability in endoscopic imaging, which can affect the reliability of diagnostic outcomes. Differences in equipment between manufacturers, such as Olympus, Fujinon, and Pentax, may lead to variations in image quality, resolution, and color rendering. Additionally, inter-operator and inter-center variability can impact the consistency of image interpretation, as factors such as experience and technique influence the examination. Technical factors, including lighting, angle of view, and working distance, further contribute to this variability, requiring careful consideration during procedures to minimize potential diagnostic errors.

3.4 The EGGIM Scoring System

The Endoscopic Grading of Gastric Intestinal Metaplasia (EGGIM) score is a standardized system designed to assess the presence and extent of gastric intestinal metaplasia based on endoscopic observations. The clinical rationale behind the EGGIM score is to provide a structured, reproducible, and non-invasive method for evaluating pre-neoplastic gastric changes, which helps in risk stratification and in planning surveillance strategies.

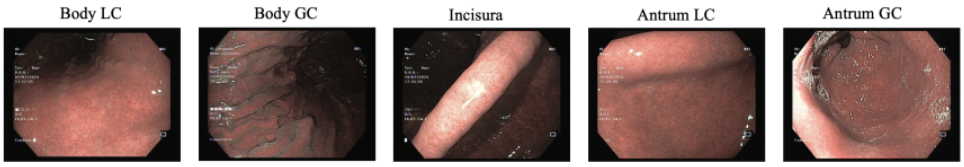


Figure 3.2: Endoscopic Regions for EGGIM Scoring

The stomach mucosa is assessed by dividing it into several regions or patches observed during the endoscopy, each of which is evaluated independently. These regions typically include areas such as the Antrum LC, Antrum GC, Body LC, Body GC, and the Incisura. Each of these patches is assigned a score based on the visual characteristics observed during the endoscopic examination. For each patch, the scoring follows a three-point scale: 0 indicates the absence of intestinal metaplasia, 1 indicates a focal or limited presence of intestinal metaplasia, and 2 indicates the extensive or diffuse presence of intestinal metaplasia. Each of these regions is evaluated separately by the endoscopist, with the score reflecting the extent of mucosal change in that specific area. The use of advanced imaging techniques, such as Narrow Band Imaging (NBI) or Chromoendoscopy, helps to enhance the visualization of subtle mucosal alterations, ensuring more accurate detection of metaplasia.
Once each patch has been evaluated, the scores are aggregated at the patient level, typically by summing or averaging the individual patch scores, depending on the approach used. The aggregated score represents the overall degree of intestinal metaplasia in the patient's stomach. For instance, if five regions (patches) are evaluated for a patient, and the scores for these regions are 0, 1, 2, 0, 1, the sum of the scores would be 4. Alternatively, the average score could be used (in this case, the average would be 1.2). The final aggregated score is then interpreted in clinical terms to assess the patient's risk of developing more severe conditions such as gastric cancer. The higher the score, the greater the extent of metaplasia and the higher the patient's risk of progression toward dysplasia and carcinoma.
The individual and aggregated scores are used for several purposes: risk stratification, follow-up recommendations, and therapeutic decisions. The aggregated score

helps classify patients based on their risk of developing gastric cancer. For instance, patients with higher scores (indicating extensive metaplasia) may be categorized as high-risk and may require more frequent monitoring. A higher EGGIM score generally indicates the need for more intensive surveillance to monitor the progression of metaplasia and the potential development of dysplasia or cancer. The EGGIM score also influences therapeutic strategies, such as the decision to initiate *H. pylori* eradication therapy or to recommend surgical intervention for high-risk patients. The EGGIM system has been clinically validated in multicenter studies, demonstrating its reliability and reproducibility among different observers. The score has been shown to have high inter-observer agreement, making it a useful tool for consistent endoscopic evaluations. In clinical practice, the EGGIM score is widely used for screening and monitoring, especially for patients at risk of gastric cancer, as it provides a quantifiable way to track changes over time. It also aids in determining surveillance intervals for patients, particularly those with higher levels of metaplasia, and supports treatment decisions based on the severity of mucosal changes [7, 15, 16].

Chapter 4

Domain Adaptation and Generalization in Healthcare: Variability and Shifts

4.1 Generalization and Domain Adaptation

In machine learning, the generalization problem refers to the model's ability to perform well on new, unseen data that comes from a distribution different from that of the training data. This challenge becomes even more complex when the model is applied to data from a different domain, where the distribution of the data may not match the one the model was trained on. The concept of domain adaptation (DA) is central to addressing this problem, especially when there are domain shifts between the source domain (training data) and the target domain (testing or deployment data) [10, 11].

The source domain is the dataset on which a model is trained, and the target domain is the dataset where the model is applied. A key issue in domain adaptation is that models trained on one dataset often fail to generalize well when applied to a dataset from a different domain. This discrepancy between the source and target domains is called domain shift. Domain shift can occur in various forms, including differences in feature distributions, label distributions, or both.

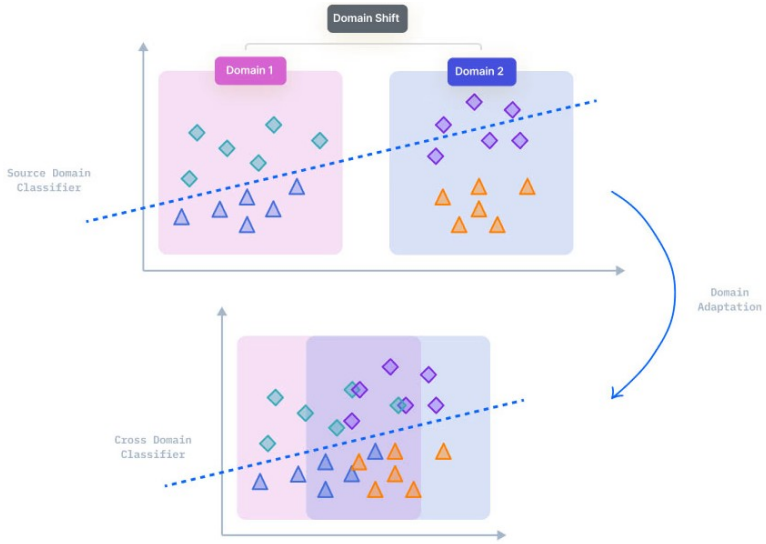


Figure 4.1: Illustration of Domain Shift between Source and Target Domains

One of the most common forms of domain shift is covariate shift, where the distribution of the input features changes between the source and target domains, but the conditional distribution of the labels remains the same. Another type of shift, known as label shift, occurs when the distribution of labels differs between the domains, even though the distribution of the input features remains similar. Lastly, prior distribution shift happens when both the feature and label distributions change.

The challenge of out-of-distribution (OOD) generalization arises when models are tested on data that is different from what they have seen during training. Models are typically designed to learn patterns that are present in the training data, and when the data distribution changes, the model often fails to perform effectively. This failure occurs because the model may overfit to specific characteristics of the training data, making it unable to generalize to new or diverse data distributions. Domain adaptation aims to address this issue by enabling models to transfer knowledge learned from one domain to another. Techniques such as adversarial training, self-supervised learning, and feature alignment are used to bridge the gap between the source and target domains. The goal of these techniques is to make the model more robust to domain shifts and improve its performance when applied to new, unseen data.

The generalization problem and domain adaptation are critical in many applications, including computer vision, natural language processing, and time-series forecasting. In these domains, datasets can vary significantly in terms of acquisition conditions, sensor types, environmental factors, and other sources of variability, all of which contribute to domain shift. Proper domain adaptation techniques can

help mitigate these differences and ensure that models perform consistently across diverse datasets.

4.2 Domain Adaptation Approaches

To address the problem of domain shift, domain adaptation (DA) techniques are specifically designed to bridge the gap between source and target domains. The challenge is not only to transfer knowledge from one domain to another, but to do so in a way that ensures robust and consistent model performance across different domains. Approaches such as feature alignment, adversarial learning, and self-training have proven effective in mitigating the effects of domain shift, allowing models trained on a specific domain to be successfully applied to others.

One of the primary strategies in domain adaptation is feature alignment, which aims to minimize the discrepancy between feature distributions in the source and target domains. Among the most widely used methods for achieving this alignment are Maximum Mean Discrepancy (MMD) and Correlation Alignment (CORAL). MMD measures the distance between the distributions of data in two domains, with the objective of reducing differences in feature representations. CORAL, on the other hand, aligns the covariance matrices of the source and target feature distributions, minimizing discrepancies in feature covariance. Both methods are widely adopted to reduce domain gaps and enhance the generalization capability of machine learning models.

Another widely used approach in domain adaptation is adversarial learning, inspired by the principles of Generative Adversarial Networks (GANs). This strategy involves training two neural networks simultaneously: a classification network and a domain discriminator. The discriminator attempts to distinguish between samples originating from the source and target domains, while the feature extractor and classification network are trained to confuse the discriminator by learning domain-invariant representations. This process is typically implemented through the use of a Gradient Reversal Layer (GRL), which reverses the gradients during backpropagation, encouraging the network to extract features that are useful for the classification task while being independent of the domain. Although adversarial learning can significantly improve cross-domain performance, it may lead to classification confusion if the discriminator becomes overly dominant, potentially causing the model to overlook task-relevant features. Domain-Adversarial Neural Networks (DANN) represent one of the most prominent implementations of adversarial domain adaptation [12]. In this framework, the classification network and the domain discriminator are trained jointly. The classification component focuses on the primary learning task, while the discriminator attempts to identify the domain of origin of each sample. The GRL plays a critical role by reversing the gradients

flowing to the feature extractor, thereby enforcing the learning of domain-invariant features. This approach is particularly effective when dealing with heterogeneous datasets collected from different sources. However, improper balancing between the classification and adversarial objectives may result in overfitting to the source domain or excessive suppression of discriminative features, ultimately degrading performance on the target domain.

Self-training constitutes another effective domain adaptation strategy, especially in scenarios where labeled data from the target domain is limited or unavailable. In this approach, a model is first trained using labeled data from the source domain and subsequently applied to unlabeled target domain data to generate pseudo-labels. These pseudo-labeled samples are then incorporated into the training process to progressively adapt the model to the target domain. While self-training can significantly enhance model adaptability, it is sensitive to errors in pseudo-label generation, as incorrect labels may propagate and negatively affect model performance.

Additional measures commonly employed to quantify and improve feature alignment include Cosine Similarity and Wasserstein Distance. Cosine Similarity evaluates the angular similarity between feature vectors and is particularly suitable for high-dimensional representations, as it focuses on directional similarity rather than magnitude. Wasserstein Distance, also known as Earth Mover's Distance, computes the optimal transport cost required to transform one probability distribution into another. This metric is especially effective for capturing complex differences between continuous and high-dimensional distributions and has been successfully applied in domain adaptation and image-to-image translation tasks.

Despite their effectiveness, domain adaptation techniques present several inherent limitations. One major issue is negative transfer, which occurs when adaptation degrades performance rather than improving it, often due to excessive dissimilarity between the source and target domains. Another challenge is overfitting to the source domain, where models learn domain-specific features that do not generalize well to the target domain. Furthermore, many domain adaptation methods implicitly assume a certain level of similarity between domains; when this assumption is violated, adaptation may fail altogether.

Finally, classification confusion remains a potential drawback, particularly in adversarial and self-training approaches. If domain invariance is enforced too strongly, models may fail to preserve task-discriminative information, leading to incorrect predictions and reduced reliability. Consequently, achieving an appropriate balance between domain invariance and task-specific feature learning is essential for the successful application of domain adaptation technique

4.3 Domain Shift in the Medical Field and Gastric Endoscopy

In the medical field, domain shift presents a particularly critical challenge due to the unique nature of healthcare data. Medical data tends to be highly heterogeneous, with variations across patient populations, healthcare facilities, and diagnostic tools [10, 17]. This makes domain adaptation particularly crucial, as models trained on data from one setting may struggle to generalize when applied to new, unseen data from different domains.

One significant issue is the limited availability of labeled data. Medical annotations often require expert knowledge, and acquiring sufficient annotated data for training is a slow and costly process, especially in specialized medical domains. This challenge is compounded by the fact that medical data is often collected under varying conditions, which results in large variability in the data distributions between different domains. For instance, different hospitals might use different medical imaging devices or have varying protocols for data collection, which can cause discrepancies in the data distribution.

In addition to these challenges, domain shift in medicine carries an inherent clinical risk. Misclassifications, inaccurate diagnoses, or treatment decisions based on poorly adapted models can result in patient harm. The risk is particularly high in fields like gastroenterology, where medical decisions are heavily reliant on accurate image-based analysis and interpretation.

One of the areas where domain shift is particularly evident is gastric endoscopy. In this domain, the problem is exacerbated by factors such as variability in instruments, protocols, and annotations. The instrument variability in gastric endoscopy is significant, as different brands and models of endoscopic devices (e.g., Olympus, Fujinon, Pentax) vary in terms of image quality, lighting, resolution, and even color reproduction. Images taken with these devices can differ greatly, which poses a challenge when training models on one set of endoscopic images and applying them to data from different devices.

Protocol variability further complicates the issue. Different medical centers or even individual practitioners may follow slightly different procedures during the endoscopic examination. These differences can include variations in camera positioning, lighting settings, and the length of time the endoscope is used in specific regions. As a result, endoscopic images captured under different protocols may look significantly different, making it harder for models trained on one protocol to perform accurately on data captured with another protocol.

Finally, annotation variability is a substantial issue in gastric endoscopy. The interpretation of endoscopic images often relies on subjective human judgment, and different clinicians may assign varying labels to the same image. For example, one

clinician might score a region as mildly inflamed, while another might score it as moderately inflamed, introducing inter-observer variability. This inconsistency in annotations can introduce noise into the data and make it difficult for machine learning models to achieve reliable performance when transferred across different centers or practitioners.
The combination of these sources of instrumental, protocol, and annotation variability makes domain adaptation a critical aspect of gastric endoscopy. Without robust adaptation techniques, models trained in one clinical setting or with one type of imaging equipment may not perform reliably in another. Developing domain adaptation strategies tailored to these challenges is essential for ensuring that automated systems in gastric endoscopy can provide accurate and trustworthy results in diverse healthcare environments.

4.4 Domain Adaptation as a Solution in the Context of EGGIM Classification and Metaplasia Classification

Domain Adaptation (DA) is essential for EGGIM classification and intestinal metaplasia classification (normal vs. metaplasia) due to the variability in data collected from multiple centers, each with its own imaging protocols, equipment configurations, and data quality. These variations can lead to models trained on data from one center performing poorly when applied to data from another center. This challenge is particularly critical when attempting to deploy models across different healthcare settings, as discrepancies in data distributions between the source and target domains are common.
In the case of EGGIM classification, this variability means that a model trained on data from one medical center may not generalize well when applied to data from another center with different imaging equipment or protocols. Domain adaptation enables the model to learn domain-invariant features, allowing it to generalize across diverse datasets without the need for extensive retraining or re-annotation. This flexibility is particularly important in clinical settings, where it is essential for automated systems to function reliably across various hospitals and healthcare facilities with differing data distributions.
For the classification of intestinal metaplasia, the challenge is even more pronounced. Metaplasia in gastric tissue can be subtle and requires expert interpretation. The variability in data quality and data distribution between centers further complicates the task. Domain adaptation helps address these challenges by enabling the model to learn from diverse data sources and adapt to differences in data quality. This improves its ability to accurately classify whether tissue is normal or shows signs

of metaplasia, even when the data comes from different centers or protocols.
Compared to traditional methods like retraining from scratch or fine-tuning, domain adaptation offers significant advantages. Retraining from scratch would require creating a new model for each dataset, which is computationally expensive and time-consuming. Fine-tuning, while more efficient, still requires a significant amount of labeled data from the target domain and assumes that the distribution of the target domain is sufficiently similar to that of the source domain. However, given the variability in medical datasets, these assumptions are often not valid. Domain adaptation, on the other hand, allows for knowledge transfer from one domain to another, improving model generalization without the need for extensive retraining or re-annotation, making it both cost-effective and scalable.
By leveraging DA, models can handle the diverse data distributions across multiple centers, ensuring consistent performance and reliable classification in clinical environments. This approach enables models to generalize effectively to new, unseen data without compromising diagnostic accuracy, making it a practical and efficient solution for real-world clinical applications.

Chapter 5

Dataset and Preprocessing

5.1 Description of the Datasets Used

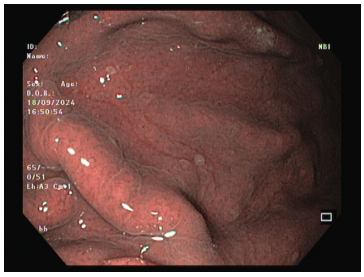
For this study, three main datasets were employed, each contributing unique characteristics that were essential for the research. The first two datasets, IPO and TOGAS, include annotations for image patches and the EGGIM scores. On the other hand, the Erasmus dataset contains images without EGGIM annotations and was used for a binary classification task in detecting metaplasia.

5.1.1 Datasets overview and acquisition protocols

The IPO dataset was collected retrospectively from routine clinical practices at the IPO-Porto between December 2019 and September 2020. It consists of frames taken from various patients, selected based on the quality of the images. The dataset includes only images that met quality requirements, meaning that frames which were blurry, out of focus, or obstructed by food or foam were excluded. The final dataset contains images of normal mucosa, atrophic mucosa, and gastric intestinal metaplasia (GIM) across all anatomical areas necessary for EGGIM scoring. These include the lesser and greater curvatures of the antrum, the incisura angularis, and the lesser and greater curvatures of the corpus. The dataset represents a broad spectrum of conditions observed in the clinical setting, making it valuable for training models on real-world data.
The second dataset, TOGAS, was prospectively collected in 2024 under a high-quality controlled acquisition protocol. It includes images from 80 patients who underwent upper gastrointestinal endoscopy, performed by experienced gastroenterologists. These procedures were conducted using Olympus high-definition endoscopes (185º or 190º series) equipped with Narrow Band Imaging (NBI), ensuring high-quality imaging. The dataset contains a mix of frames, including normal and

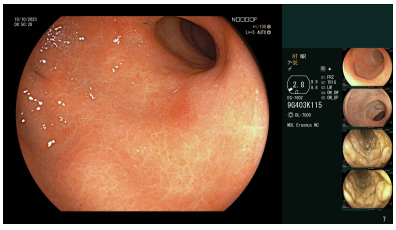
abnormal gastric conditions. The images were captured in JPG format with a resolution of 640×480 pixels and a 24-bit color depth, providing sufficient detail for analysis. This dataset was particularly valuable for its controlled environment and high-quality data, which allowed for more precise model training and validation.
The third dataset, Erasmus, was collected in 2024 using FUJIFILM ELUXEO® 8000 endoscopes at the Erasmus University Medical Clinic (ErasmusMC) in Rotterdam, the Netherlands. This dataset comprises 307 white-light (WL) and 313 blue-light imaging (BLI) images from 80 patients. All images have a minimum resolution of 1232x1048 pixels. While this dataset is substantial in terms of image resolution and the number of patients, it does not include EGGIM annotations, which made it unsuitable for certain types of analyses.
For the purpose of binary classification, however, the IPO dataset was excluded from the analysis. This decision was made due to the significant class imbalance in the dataset, and the poor quality of some images, which made them unsuitable for training a robust model. As a result, the final model training focused on the normal mucosa and intestinal metaplasia classes, both of which were more balanced and suitable for a binary classification task.

5.1.2 Representative examples



(a) Example TOGAS

(b) Example IPO



(c) Example ERASMUS

Figure 5.1: Representative examples from the three datasets used in this study.

5.1.3 General statistics

In the following tables, general statistics of the datasets are presented, including the total number of images, the number of patients, the images per patient, and the class distribution. These statistics help provide an overview of the data before the preprocessing and balancing steps were applied. In the domain adaptation experiments discussed in the following chapters, TOGAS is used as the *source* domain. This choice is motivated by its controlled acquisition protocol and overall higher image quality, together with a larger number of samples per patient and a more stable class distribution compared to the other datasets (see Tables 5.1 and 5.2). In light of the previously discussed considerations,for the three-class EGGIM classification task, TOGAS was used as the source domain and IPO as the target domain (TOGAS $\rightarrow$ IPO).Whereas for the binary classification task (Normal vs Metaplasia), TOGAS was again used as the source domain, while ERASMUS served as the target domain (TOGAS $\rightarrow$ ERASMUS). These configurations enable the evaluation of model robustness under different levels of domain shift, as previously quantified through similarity and distribution metrics. Table 5.2 highlights notable differences in dataset structure. TOGAS provides the highest number of images per patient, offering a richer and more consistent sampling of the gastric mucosa, whereas IPO contains relatively few images per patient, reflecting its retrospective nature and increasing the risk of patient-level sparsity during training. ERASMUS lies in between in terms of images per patient but differs in acquisition modality and resolution, which can introduce additional covariate shift when transferring models across centers.
Finally, Table 5.1 shows that the three-class (EGGIM) setting is strongly imbalanced in IPO (with a clear dominance of class 2), while TOGAS exhibits a more gradual distribution across classes. This mismatch anticipates a challenging label-shift scenario when adapting from TOGAS to IPO.

Dataset	Class	Samples	Percentage
TOGAS	0	509	58.8%
TOGAS	1	188	21.7%
TOGAS	2	169	19.5%
IPO	0	54	13.0%
IPO	1	55	13.3%
IPO	2	305	73.7%

Table 5.1: Class distribution in the datasets used.

Dataset	Total images	Patients	Images/patient	Class 0	Class 1	Imbalance
TOGAS	866	80	10.82	509	357	1.43
IPO	414	247	1.68	54	360	6.67
Erasmus	489	75	6.52	179	310	1.73

Table 5.2: General statistics for the datasets used.

5.1.4 Label shift and class distributions

Figures 5.2 and 5.3 illustrate the class distributions of the considered datasets, highlighting relevant discrepancies that directly impact the domain adaptation scenarios investigated in this work.

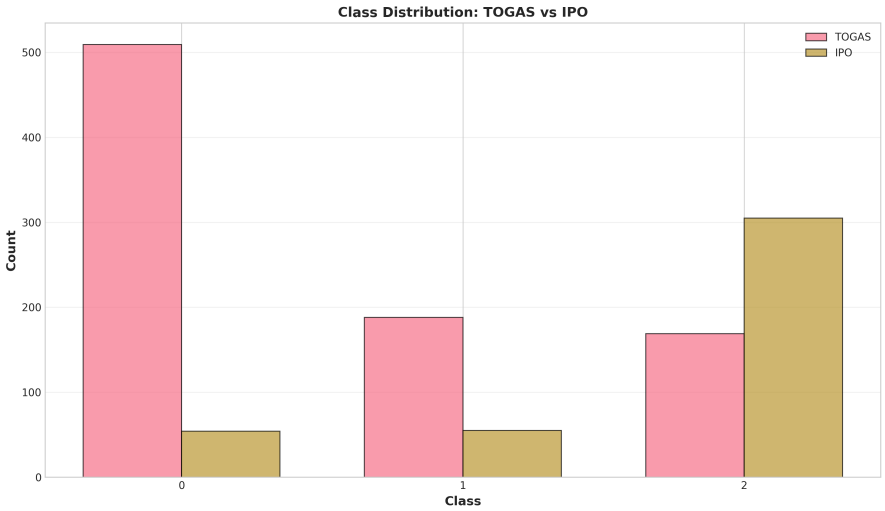


Figure 5.2: EGGIM score distribution in the IPO dataset.

In the three-class classification setting (EGGIM 0, 1, 2), the TOGAS dataset exhibits a relatively more balanced distribution across classes compared to IPO. In particular, TOGAS contains a higher number of samples in class 0 and a more gradual decrease toward higher EGGIM scores, whereas IPO shows a markedly different structure, with a substantial concentration of samples in class 2 and a severe underrepresentation of class 0. This discrepancy induces a distribution shift not only at the feature level but also at the label level, resulting in a combined covariate and label shift scenario.

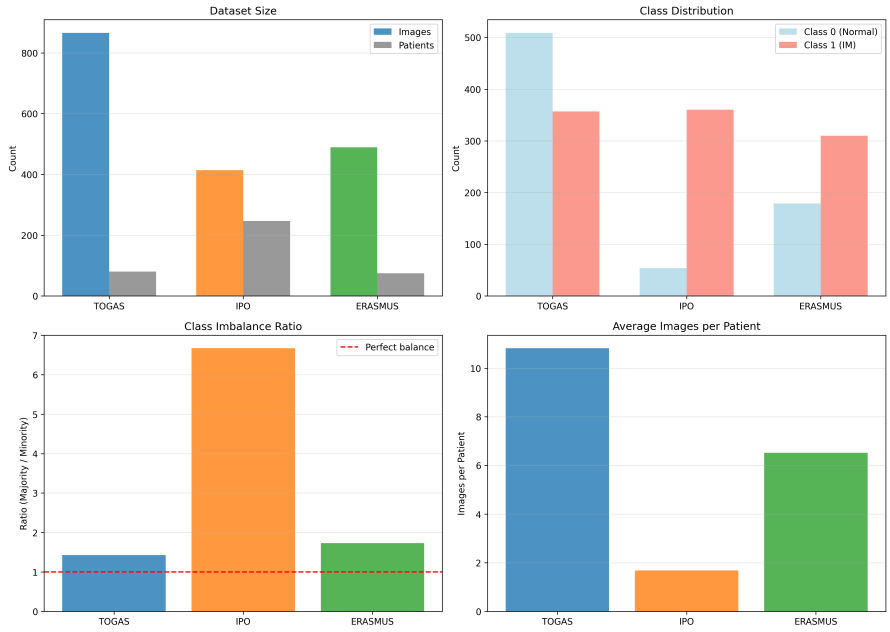


Figure 5.3: Binary class distribution for metaplasia.

When performing domain adaptation from TOGAS (source) to IPO (target) for three-class classification, the model is therefore required to adapt to a substantially different class prior distribution. In particular, IPO presents a strong class imbalance, with class 2 dominating the dataset, whereas TOGAS exhibits a more moderate imbalance. This mismatch can bias the decision boundary learned on the source domain, potentially leading to overestimation of majority classes or under-detection of minority ones in the target domain.

A different scenario emerges in the binary classification setting (Normal vs Intestinal Metaplasia) for the TOGAS → ERASMUS adaptation. While the imbalance ratios of TOGAS and ERASMUS are more comparable than in the IPO case, differences remain in both class proportions and dataset structure. ERASMUS shows a higher prevalence of metaplasia cases relative to normal samples compared to TOGAS. Furthermore, differences in the average number of images per patient and in the number of patients per dataset introduce additional sources of distribution shift, potentially affecting model generalization. Specifically, TOGAS contains a larger number of images per patient, which may lead the model to learn patient-specific visual patterns that are less transferable to ERASMUS, where the distribution of samples per patient differs. This structural discrepancy can introduce a hidden bias during training, particularly if the source model implicitly captures intra-patient correlations.

Overall, the presented distributions demonstrate that the investigated domain adaptation scenarios are characterized not only by potential feature-level discrepancies

due to acquisition conditions but also by significant label distribution differences. Consequently, the adaptation task involves handling both covariate shift and class imbalance shift, making these scenarios particularly challenging and representative of real-world multi-center deployment.

5.1.5 Feature distribution similarity

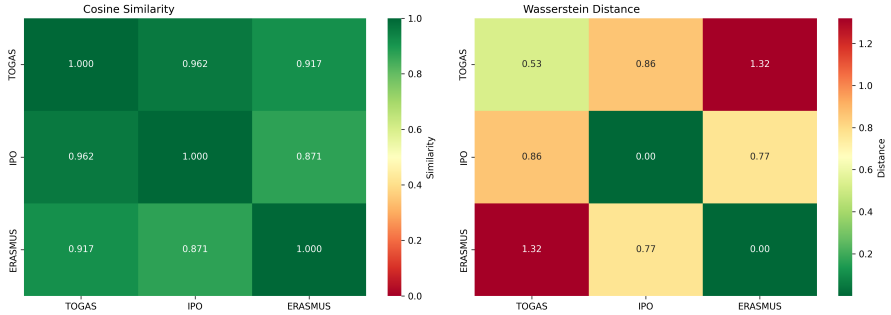


Figure 5.4: Cosine similarity and Wasserstein distance among datasets.

To quantitatively assess the discrepancy between datasets, two complementary metrics were computed: cosine similarity and Wasserstein distance. These measures provide insights into the similarity of feature distributions across domains.
Cosine similarity evaluates the angular similarity between two vectors and is defined as the normalized dot product between them. Given two feature vectors x and y, cosine similarity is computed as:

$$\text{Cosine}(x, y) = \frac{x \cdot y}{\|x\|\|y\|}$$

Values range between -1 and 1, although in practice for feature embeddings they typically fall between 0 and 1. Higher values indicate greater similarity in direction, suggesting that the global structure of the feature representations is aligned across domains.
In contrast, the Wasserstein distance (also known as Earth Mover's Distance) measures the minimum cost required to transform one probability distribution into another. Unlike cosine similarity, which captures angular alignment, Wasserstein distance reflects discrepancies in the overall distribution shape and mass displacement. Lower values indicate more similar distributions, whereas higher values reflect stronger distributional shift. As shown in Figure 5.4, TOGAS and IPO exhibit a high cosine similarity (0.962), indicating that their global feature representations are relatively aligned. This suggests that the overall embedding structure is similar between the two datasets. However, the Wasserstein distance between TOGAS and IPO (0.86) reveals a non-negligible distributional discrepancy.

This combination indicates a scenario where features are directionally similar but differ in their statistical distribution, reflecting a moderate covariate shift.
A more pronounced discrepancy is observed between TOGAS and ERASMUS. Although the cosine similarity remains relatively high (0.917), the Wasserstein distance reaches 1.32, which is the largest among all dataset pairs. This indicates that while the feature embeddings may lie in a similar subspace, the underlying distributions differ substantially. Such behavior is consistent with structural differences between the datasets, including acquisition conditions, class proportions, and dataset composition.
Interestingly, IPO and ERASMUS show a lower Wasserstein distance (0.77) compared to TOGAS–ERASMUS, suggesting that IPO may be distributionally closer to ERASMUS than TOGAS is, despite TOGAS being used as the source domain in both adaptation scenarios.
From a domain adaptation perspective, these results provide quantitative evidence that the TOGAS $\rightarrow$ IPO scenario corresponds to a moderate domain shift, characterized primarily by differences in class priors and moderate feature distribution misalignment. In contrast, the TOGAS $\rightarrow$ ERASMUS scenario represents a stronger domain shift, as reflected by the larger Wasserstein distance, implying greater distributional divergence.
Therefore, the adaptation task in the three-class TOGAS $\rightarrow$ IPO setting can be interpreted as addressing a combined covariate and label shift problem, whereas the binary TOGAS $\rightarrow$ ERASMUS setup involves a more substantial distributional mismatch that may require stronger feature alignment mechanisms. These findings support the need for explicit domain adaptation strategies in both scenarios, particularly for cross-center deployment.

5.2 Preprocessing and Data Augmentation

A consistent preprocessing strategy was adopted to ensure structural comparability across datasets while preserving diagnostically relevant visual information. All images were resized to 224×224 pixels to match the input requirements of the backbone architectures used in this study.
For the TOGAS and IPO datasets, bounding box annotations identifying the region of interest were available and used to crop each image before resizing. Cropping restricts the model input to the annotated mucosal area, reducing the influence of peripheral regions and acquisition-related artifacts. This encourages the model to focus on pathology-related patterns rather than background elements that may vary across centers.
The ERASMUS dataset does not include bounding box annotations. To maintain consistency in spatial processing, a central crop was applied prior to resizing.

Although less precise than bounding box-based cropping, this approach ensures comparable image scaling and limits the impact of irrelevant borders.
After cropping and resizing, per-image standardization was applied to all datasets. Each image was independently normalized to reduce variations in brightness and contrast caused by differences in endoscopic devices, illumination conditions, and acquisition protocols.
Endoscopic imaging is particularly sensitive to lighting intensity and color balance, which may differ substantially across centers. Per-image normalization attenuates these low-level intensity discrepancies, allowing the model to focus more on structural and textural characteristics rather than absolute brightness values.
For the three-class EGGIM classification task, original labels (0, 1, 2) were preserved for both TOGAS and IPO datasets.
For the binary classification task (Normal vs Intestinal Metaplasia), labels were remapped such that images annotated with EGGIM score 0 were assigned to class 0 (normal), while images annotated with score 1 or 2 were grouped into class 1 (metaplasia). This aggregation reflects a clinically meaningful distinction between absence and presence of metaplastic changes.
In the ERASMUS dataset, images labeled as "normal" were assigned to class 0 and those labeled as "metaplasia" to class 1. Images associated with other findings, such as atrophic gastritis, or uncertain diagnoses were excluded from the analysis to maintain semantic consistency between source and target domains.
Data augmentation was intentionally limited to simple geometric transformations to preserve the anatomical and morphological integrity of endoscopic structures. Specifically, random horizontal and vertical flips were applied during training. These transformations increase sample variability without altering mucosal texture or pathological patterns.
Augmentation was applied exclusively to the training set, while validation and test sets remained unchanged to ensure an unbiased performance evaluation.
The datasets exhibit varying degrees of class imbalance, particularly in the three-class EGGIM setting. To counteract the effect of skewed class distributions during optimization, class weights were computed inversely proportional to class frequency in the training set. For each class i, the weight was defined as:

$$w_i = \frac{N}{K \cdot n_i}$$

where N is the total number of training samples, K is the number of classes, and n_i is the number of samples belonging to class i. For the EGGIM classification task, these weights were incorporated into a weighted categorical cross-entropy loss function. In the binary classification setting, a binary cross-entropy loss was employed, with class weights computed using the same formulation. This weighting scheme increases the contribution of minority classes to the overall loss, reducing

bias toward majority classes and improving training stability.

Chapter 6

Methodology

6.1 General Setting

6.1.1 Baseline Architecture

The ResNet50 architecture was chosen as the feature extractor due to its proven effectiveness in a wide range of image classification tasks [18]. It was pre-trained on ImageNet, which provides a robust set of general features applicable across different domains. This pre-trained model offers a solid starting point, especially in the context of gastric endoscopic image analysis, where feature patterns like textures and contours are shared with general image datasets. Utilizing ResNet50 helps mitigate the challenge of limited labeled data often encountered in medical imaging.

ResNet50's residual connections enable the model to train deep networks without suffering from vanishing gradients, allowing the network to learn complex hierarchical features effectively. The top classification layers were removed, and custom layers were added to adapt the model to the specific task at hand. The model was initialized with pre-trained weights, and during the initial training phase, the weights of the ResNet50 base model were frozen. This helps prevent overfitting and ensures that the model learns task-specific representations through the custom layers.

For classification, the model uses a custom head designed to accommodate both binary and multi-class tasks. In the case of binary classification, a sigmoid activation function was applied with a single output neuron. For multi-class classification (three classes in this case), a softmax activation function was used with one output neuron per class. To process the feature map output from ResNet50, a Global Average Pooling (GAP) layer was added to reduce dimensionality while preserving spatial information. This was followed by a dense layer with 1024 units and ReLU activation, which captures more complex relationships in the data. L2 regularization

was applied to this layer to avoid overfitting, which is particularly important due to the small size of the medical datasets. Dropout layers with rates of 0.3 and 0.5 were included to further prevent overfitting by randomly deactivating neurons during training, thus encouraging the model to generalize better to unseen data.
By combining ResNet50's pre-learned feature extraction capabilities with custom layers tailored to the gastric endoscopic task, this architecture strikes a balance between using robust general features and learning specific patterns required for the classification task. This approach ensures efficient training with limited data, which is crucial in the medical imaging domain, where large labeled datasets are often unavailable.

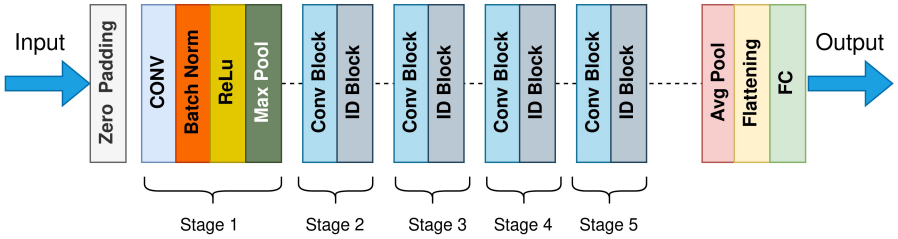


Figure 6.1: ResNet50 architecture

6.1.2 Training Settings

The training process was implemented using the TensorFlow framework, with the Keras API providing a flexible interface for model definition and training. For efficient data handling, the dataset was processed using TensorFlow's 'tf.data' API. This API allows the data to be pre-processed, batched, and prefetched, ensuring that the model receives data efficiently during training.
The preprocessing pipeline involved reading the images and applying necessary transformations. Data augmentation, such as random flipping, was applied during training to increase the variability of the dataset and reduce the risk of overfitting. The datasets were then batched and prefetching was enabled to optimize throughput. Prefetching allows the data pipeline to load the next batch while the current batch is being processed, which improves training efficiency.
For the optimization process, the Adam optimizer was chosen due to its adaptive learning rate, which accelerates convergence and helps prevent the model from getting stuck in local minima. The learning rate was adjusted dynamically during training using the ReduceLROnPlateau callback. This callback reduces the learning rate by a factor when the validation loss stagnates, enabling more refined adjustments as the model converges.
To address class imbalance, weighted loss functions were used. For multi-class classification tasks, weighted categorical cross-entropy was applied, where the class

weights were computed inversely proportional to the frequency of each class in the training set. This ensures that underrepresented classes are given more importance during training. The class weights were computed using the formula:

$$w_i = \frac{N}{K \cdot n_i}$$

where N is the total number of training samples, K is the number of classes, and n_i is the number of samples in class i. For binary classification tasks, a similar approach was used with binary cross-entropy, adjusting for class imbalance.
Several callbacks were integrated to optimize the training process. The ModelCheckpoint callback saved the model weights with the best validation loss, ensuring that the model with the best performance was retained. The EarlyStopping callback was used to halt training if there was no improvement in the validation loss for a set number of epochs, preventing overfitting.
The final model was trained using the specified callbacks, and the training history was recorded for further analysis. The best model weights, based on validation performance, were saved during training to ensure that the model that generalized best to the validation data was retained.

6.1.3 Evaluation Metrics

The performance of the model was evaluated using several standard classification metrics to ensure a comprehensive assessment of its effectiveness. These metrics provide insights into the model's ability to correctly classify images, minimize misclassifications, and handle class imbalance.
Accuracy was used as the primary metric to measure the proportion of correct predictions across all classes. While accuracy provides a general measure of performance, it may be misleading in the case of imbalanced datasets, as it can be dominated by the majority class.
Precision and recall were computed for each class to assess the model's ability to correctly identify positive instances and avoid false positives. Precision evaluates the proportion of true positive predictions among all positive predictions, while recall measures the proportion of actual positive instances correctly identified by the model.
The F1-score was used as a balanced performance metric, combining both precision and recall into a single value. The F1-score is particularly useful when there is an imbalance between precision and recall, as it gives a better sense of the model's overall performance in identifying positive instances while avoiding false positives.
To further analyze the model's performance across different classes, confusion matrices were used. These matrices allow for a detailed visualization of how the model predicts each class, highlighting where misclassifications occur and providing

insight into whether the model is consistently confusing certain classes with others. In addition to the standard classification metrics, AUC (Area Under the Curve) was also used as a performance measure. AUC evaluates the model's ability to distinguish between classes, with higher values indicating better discrimination power. This metric is particularly useful in imbalanced classification tasks, as it provides a more nuanced view of the model's performance across different decision thresholds.
Overall, these metrics were chosen to provide a comprehensive view of the model's ability to classify images accurately, minimize errors, and handle imbalanced classes. By analyzing these metrics, the model's strengths and weaknesses could be identified, allowing for better tuning and refinement during the training process.

6.2 Self-Supervised Learning and Pseudo-Labeling

6.2.1 General Framework

Self-supervised learning combined with pseudo-labeling represents an effective strategy for domain adaptation when supervision is available only in the source domain during training [19]. The main objective is to progressively adapt the model to the target distribution by exploiting its own high-confidence predictions as additional supervisory signals.
The pseudo-labeling pipeline follows a structured procedure. First, a model is trained in a fully supervised manner on the labeled source dataset. Once convergence is achieved, the trained model is used to generate predictions on the target samples. For each target image, the model outputs a probability distribution over the classes, and the class associated with the maximum predicted probability is assigned as a pseudo-label.
Since pseudo-labels may contain errors, a confidence-based filtering mechanism is typically applied. Only samples whose maximum predicted probability exceeds a predefined confidence threshold are retained. This reduces the risk of propagating incorrect predictions and stabilizes the adaptation process.
The procedure can be iteratively repeated. After selecting high-confidence pseudo-labeled samples, they are incorporated into the training set and the model is retrained. The updated model is then used again to generate pseudo-labels for the remaining target samples. Through this iterative self-training loop, the feature representations progressively adapt to the target domain.

6.2.2 Implementation and Experimental Settings

Although ground-truth annotations were available for the target dataset, they were not used during adaptation. The target domain was treated as unlabeled during

the pseudo-labeling process and reserved exclusively for final evaluation.
The initial model was trained on the source domain (TOGAS) using supervised learning. After convergence, the trained model was used to generate probability predictions for the entire target dataset (IPO), which constituted the adaptation pool.
The self-training process was configured with a maximum of five iterations. At each iteration, pseudo-labels were generated for all remaining samples in the target pool. High-confidence samples were selected using a confidence threshold of 0.9 on the maximum predicted class probability. Only predictions exceeding this threshold were considered reliable enough to be incorporated into the training process.
No fixed upper limit on the number of samples per class was imposed; however, class-wise balancing was monitored to avoid excessive skew introduced by pseudo-label accumulation. When sufficient pseudo-labeled samples were available, a stratified split was applied to create pseudo-training and pseudo-validation subsets using an 80/20 ratio. Stratification was performed according to the pseudo-label distribution to preserve class balance.
Pseudo-labeled training samples were accumulated across iterations, progressively enlarging the adaptation dataset. Samples not meeting the confidence criterion remained in the target pool and were reconsidered in subsequent iterations. An acceptance threshold of 0.8 was introduced as a stopping criterion: if more than 80% of the target pool had been pseudo-labeled, the iterative process was terminated to prevent excessive inclusion of potentially noisy samples.
At each iteration beyond the initial one, the model was re-initialized and retrained from scratch using a combined dataset composed of labeled source samples and accumulated pseudo-labeled target samples. This strategy avoided bias due to optimizer state carry-over and ensured consistent training dynamics across iterations.
Model selection was performed exclusively based on source-domain validation performance. The model achieving the highest source test accuracy across iterations was retained as the final adapted model. The target test set was evaluated only once at the end of the entire self-training procedure.

6.2.3 Evaluation Metrics and Representation-Level Analysis

Beyond standard classification metrics, additional representation-level analyses were conducted to assess the structural integration of pseudo-labeled samples within the learned feature space.
Feature embeddings were extracted from the penultimate layer of the network. To reduce dimensionality while preserving most of the informative variance, Principal Component Analysis was first applied, retaining up to 50 principal components. The retained components typically captured a large proportion of the total variance.

Subsequently, UMAP (Uniform Manifold Approximation and Projection) was applied with 15 nearest neighbors, a minimum distance parameter of 0.1, Euclidean metric, and fixed random seed for reproducibility. UMAP provided a two-dimensional visualization of the feature space, enabling qualitative inspection of clustering behavior between manually labeled target samples and pseudo-labeled samples. A desirable adaptation scenario is characterized by overlapping, compact clusters rather than isolated pseudo-labeled regions.
To complement visual inspection with quantitative measures, Nearest-Neighbor Distance (NND) metrics were computed in the PCA-reduced feature space. Two distributions were analyzed: manual-to-manual nearest-neighbor distances (baseline) and pseudo-to-manual nearest-neighbor distances. Median and interquartile range (IQR) were computed for both distributions. The ratio between the median pseudo-to-manual and manual-to-manual distances was used as an alignment indicator. A ratio close to one suggests that pseudo-labeled samples occupy regions of the feature space comparable to genuine labeled samples.
Additionally, Local Density Ratio (LDR) analysis was performed. For each sample, local density was estimated using the inverse of the mean Euclidean distance to the 15 nearest manual neighbors. The 25th percentile of the manual density distribution was used as a reference threshold. The fraction of pseudo-labeled samples falling below this density threshold was computed. A high LDR indicates that many pseudo-labels lie in low-density regions, potentially corresponding to unreliable or structurally inconsistent predictions.
The combined use of UMAP visualization, NND statistics, and LDR analysis provided both qualitative and quantitative assessment of feature alignment and pseudo-label reliability throughout the iterative self-training process.

6.3 Domain Adversarial Neural Networks (DANN)

6.3.1 General Framework

Domain Adversarial Neural Networks (DANN) provide a principled framework for domain adaptation based on adversarial learning. The objective is to learn feature representations that are simultaneously discriminative for the main classification task and invariant with respect to the domain.

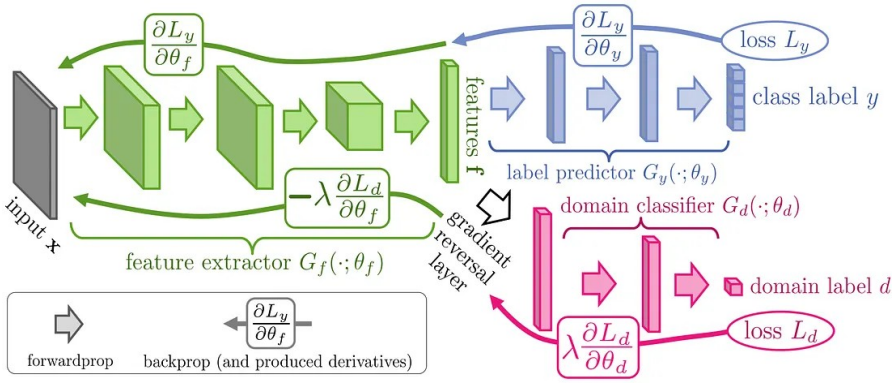


Figure 6.2: Overview of the Domain-Adversarial Neural Network (DANN) architecture. Figure adapted from [20].

The architecture is composed of three components: a feature extractor, a task classifier, and a domain discriminator. The feature extractor maps input images into a latent representation space. The task classifier is trained using supervised labels from the source domain. In parallel, the domain discriminator attempts to distinguish whether extracted features originate from the source or the target domain.
Adversarial learning is achieved through a Gradient Reversal Layer (GRL), which multiplies the gradient coming from the domain discriminator by a negative scalar during backpropagation. While the discriminator learns to maximize domain separability, the feature extractor is optimized to minimize it. This min–max formulation encourages the emergence of domain-invariant features.
The overall optimization objective combines task supervision and domain confusion:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda \mathcal{L}_{domain}$$

where $\mathcal{L}_{task}$ is the classification loss on source samples, $\mathcal{L}_{domain}$ is the binary domain discrimination loss computed on both source and target samples, and λ controls the strength of the adversarial signal.

6.3.2 Model Architecture

The implemented DANN model was built upon a ResNet50 backbone pre-trained on ImageNet. The convolutional base served as the shared feature extractor. After global average pooling, a fully connected layer with 1024 units and L2 regularization was used as a shared representation layer (feature layer) feeding both the task and domain branches.
The task head consisted of a dense output layer with softmax activation for the

three-class EGGIM problem and sigmoid activation for the binary setting. To address class imbalance, weighted categorical cross-entropy (or weighted binary cross-entropy in the binary case) was used as the task loss.
The domain discriminator branch was connected through a Gradient Reversal Layer. In the standard DANN configuration, the discriminator consisted of a dense hidden layer followed by dropout and a final sigmoid neuron for binary domain classification.
In addition to standard DANN, a Conditional Domain Adversarial Network (CDAN) variant was implemented. In this configuration, the discriminator receives as input the concatenation of feature representations and class predictions. This class-aware conditioning encourages alignment of class-conditional feature distributions across domains.

6.3.3 Training Procedure

Training was performed using mixed batches containing both source and target samples. Source samples contributed to both the task loss and the domain loss, while target samples contributed exclusively to the domain loss. This behavior was implemented through sample weighting, ensuring that classification loss was computed only for source instances.
The total loss consisted of two components: a weighted classification loss for the main task and a binary cross-entropy loss for domain discrimination. Both losses were jointly optimized using the Adam optimizer with weight decay regularization. The relative influence of the adversarial component was controlled dynamically through the GRL parameter λ.
To ensure stable optimization, the adversarial signal was not applied at full strength from the beginning of training. Instead, a progressive scheduling strategy was adopted. The GRL parameter λ was gradually increased following the classical DANN ramp function:

$$\lambda(p) = \frac{2}{1 + e^{-10p}} - 1$$

where p denotes normalized training progress. The maximum value of λ was capped to avoid excessive adversarial pressure that could degrade classification performance.
In addition to the standard DANN training, a variant was implemented by initializing the feature extractor and the task classifier head with weights obtained from the source-only training. A warm-up phase was then introduced: during this initial stage the GRL parameter was set to zero, effectively disabling domain-adversarial gradients, so that the task classifier could stabilize under supervised source training before progressively enforcing domain alignment.

At the end of the warm-up phase, the deepest convolutional block of the backbone was unfrozen to enable fine-tuning under adversarial supervision. This two-stage strategy, initial task stabilization followed by controlled adversarial adaptation, improved convergence stability and reduced oscillatory training behavior.
Early stopping and learning rate scheduling were applied based on validation label accuracy to prevent overfitting and ensure balanced optimization between task performance and domain invariance.

6.3.4 Evaluation Metrics and Domain Alignment Analysis

In addition to standard task-level metrics, domain-specific indicators were monitored to quantify the effectiveness of adversarial alignment. In particular, domain accuracy and domain AUC were tracked throughout training.
Domain accuracy measures the ability of the domain discriminator to correctly classify whether a feature representation originates from the source or the target domain. High domain accuracy indicates that the learned representations remain domain-separable, meaning that domain-specific information is still encoded in the feature space. Conversely, domain accuracy approaching 0.5 reflects increasing domain confusion, suggesting that the feature extractor has successfully reduced domain-specific cues. However, excessively low domain performance combined with degraded task accuracy may indicate unstable adversarial optimization. Therefore, domain accuracy was interpreted jointly with task metrics to assess balanced adaptation.
Domain AUC provides a threshold-independent measure of domain separability. Values close to 1 indicate strong domain discrimination, whereas values close to 0.5 indicate that source and target representations are indistinguishable. Monitoring both accuracy and AUC allowed a more robust evaluation of domain invariance beyond a single operating threshold.
To complement quantitative domain metrics, representation-level analysis was performed using UMAP projections of the shared feature embeddings. Rather than serving as a standalone metric, UMAP visualization was used to qualitatively inspect the structural organization of the feature space. Effective domain adaptation is characterized by overlapping source and target distributions with coherent class-wise clustering, while persistent domain separation suggests incomplete alignment. The joint analysis of classification performance, domain discriminator behavior, and feature-space structure enabled a comprehensive assessment of both predictive accuracy and the degree of learned domain invariance.

6.4 Margin Disparity Discrepancy (MDD)

6.4.1 Theoretical Foundations

Margin Disparity Discrepancy (MDD) is a domain adaptation framework based on classifier discrepancy rather than explicit domain discrimination. Unlike adversarial approaches relying on a domain classifier, MDD aligns domains by exploiting disagreement between two task classifiers operating on a shared feature representation. The central idea is that if two classifiers trained on the source domain produce highly divergent predictions on target samples, then the target features lie outside the source-supported decision regions. By explicitly maximizing this disagreement and subsequently minimizing it through feature adaptation, the model is encouraged to learn domain-invariant representations that preserve class discrimination.
Let C_1 and C_2 denote two classifiers sharing a common feature extractor F. The optimization follows a min–max structure:

$$\min_{F,C_1} \max_{C_2} \; \mathcal{L}_{source} - \lambda \, \mathcal{D}(C_1, C_2)$$

where $\mathcal{L}_{source}$ is the classification loss on labeled source data and $\mathcal{D}(C_1, C_2)$ is a discrepancy measure computed on target samples. The discrepancy term encourages C_2 to maximize disagreement with C_1, while the feature extractor is optimized to minimize it, thereby aligning the target distribution with the source decision boundary.

6.4.2 Model Architecture

The implemented MDD model was built upon a ResNet50 backbone pre-trained on ImageNet. The convolutional base acted as a shared feature extractor. After global average pooling, a fully connected layer served as the shared feature representation (feature layer).
Two parallel classifier heads, C_1 and C_2, were attached to the shared feature layer. Each classifier consisted of a dense output layer with softmax activation for the three-class problem (or sigmoid activation in the binary setting). Both classifiers were trained using weighted cross-entropy to address class imbalance in the source domain.
Unlike DANN, no domain discriminator was used. Instead, alignment was driven by a discrepancy loss computed between the softmax outputs of the two classifiers on target samples.

6.4.3 Discrepancy Loss and Optimization Strategy

The discrepancy between classifiers was defined as the mean absolute difference between their softmax outputs on target data:

$$\mathcal{D}(C_1, C_2) = \mathbb{E}_{x \sim T} \left[\| \text{softmax}(C_1(F(x))) - \text{softmax}(C_2(F(x))) \|_1 \right]$$

Training followed a three-part objective:
- Classifier 1 (C_1): minimized the weighted classification loss on source data. - Classifier 2 (C_2): minimized the source classification loss while maximizing the discrepancy term. - Feature extractor (F): minimized the source classification loss and minimized the discrepancy term.
This creates an adversarial game in which C_2 attempts to expose regions of target-space uncertainty, while the feature extractor learns to suppress such disagreement. To stabilize this adversarial interaction, two separate optimizers were used: one for the main network parameters and one dedicated to the second classifier. The second optimizer operated with a slightly higher learning rate to ensure sufficient adversarial pressure and prevent premature collapse of the discrepancy signal.
The discrepancy term was weighted by a hyperparameter λ_{disc} controlling the strength of alignment. Excessively high values can destabilize training, whereas too small values lead to insufficient adaptation. The parameter was selected empirically to balance classification accuracy and stable discrepancy reduction.
The backbone remained frozen during adaptation to prevent unstable feature drift and ensure controlled alignment in the shared representation space.

6.4.4 Training Procedure

Training batches contained both source and target samples. Source samples contributed to the classification loss of both classifiers. Target samples contributed exclusively to the discrepancy objective.
Optimization alternated implicitly through gradient updates:
- The adversarial classifier (C_2) was updated to increase prediction disagreement on target samples. - The feature extractor was updated to reduce such disagreement.
- The primary classifier (C_1) remained aligned with the source task objective.
Early stopping and learning rate scheduling were applied based on validation accuracy of the primary classifier. Model checkpointing was performed using validation task performance to ensure that adaptation did not degrade discriminative capability.
Final predictions on the target test set were obtained by averaging the outputs of C_1 and C_2, effectively forming an ensemble that reduces variance and improves robustness.

6.4.5 Evaluation Metrics and Representation Analysis

Task-level evaluation was conducted using accuracy, macro-averaged AUC (one-vs-rest), precision, recall, and F1-score on the target test set. Confusion matrices were generated to inspect class-wise performance.

In addition to classification metrics, the discrepancy value between the two classifiers was tracked across training epochs. A high discrepancy indicates domain misalignment, while progressive reduction of discrepancy suggests improved alignment of target features within the source decision regions. However, a trivial reduction of discrepancy accompanied by deteriorated task accuracy may indicate over-regularization or feature collapse. Therefore, discrepancy was interpreted jointly with task metrics.

To complement quantitative evaluation, embeddings extracted from the shared feature layer were visualized using UMAP. Source and target samples were projected into a two-dimensional space to qualitatively inspect structural alignment. Successful MDD adaptation is characterized by improved overlap between domains while maintaining coherent class-wise organization.

The combined analysis of classification performance, classifier discrepancy dynamics, and feature-space structure provided a comprehensive assessment of adaptation effectiveness.

Chapter 7

Experiments

7.1 EGGIM Three-Class Classification

7.1.1 Upper Bound

To estimate the upper bound performance achievable when training directly on the target domain, a patient-wise cross-validation protocol was adopted using only the IPO dataset. This evaluation represents an optimistic scenario in which labeled target data are fully available for supervised learning.
The dataset was partitioned using a five-fold patient-wise cross-validation scheme to ensure that images from the same patient did not appear in multiple splits. For each fold, the data were divided into training, validation, and test subsets while preserving patient-level separation.
For every fold, a new model was initialized and trained independently using the architecture described in the previous chapter. Training was performed with a batch size of 32 and an initial learning rate of 1×10^{-4}. Gradient clipping with a maximum norm of 1.0 was applied to stabilize the optimization process.
Class weights were recomputed separately for each fold using only the training subset in order to account for potential variations in class distribution across folds. These weights were incorporated into the loss function to maintain balanced learning between the three EGGIM classes.
Each fold was trained independently, and the best model weights were selected according to validation performance through model checkpointing. After training, the selected model was evaluated on the corresponding held-out test subset. The final upper bound performance was obtained by aggregating the results across all folds.
The performance obtained through the five-fold patient-wise cross-validation on the IPO dataset is reported in Table 7.1. Results are presented as mean and standard deviation across folds.

Class	Precision	Recall	F1-score	Support
0	0.9080 ± 0.1047	0.6882 ± 0.2782	0.7594 ± 0.2230	54
1	0.7078 ± 0.2174	0.7621 ± 0.2567	0.7187 ± 0.2064	55
2	0.9261 ± 0.0576	0.9431 ± 0.0518	0.9330 ± 0.0371	305
Macro Avg	0.8473 ± 0.1090	0.7978 ± 0.1544	0.8037 ± 0.1414	414
Weighted Avg	0.8949 ± 0.0693	0.8895 ± 0.0676	0.8838 ± 0.0755	414
Accuracy		0.8895 ± 0.0676		414

Table 7.1: Upper bound performance obtained using five-fold patient-wise cross-validation on the IPO dataset.

The upper bound experiment achieved an overall accuracy of 0.8895 ± 0.0676, providing an estimate of the best achievable performance when the model is trained directly on the target domain.
The results show strong performance across the dataset, particularly for class 2, which achieves the highest precision, recall, and F1-score. This behavior is consistent with the class distribution, where class 2 represents the majority of samples (73.7% of the IPO dataset) and therefore benefits from a larger number of training examples. The model can reliably learn the visual patterns associated with severe intestinal metaplasia when sufficient labeled data are available.
Classes 0 and 1 exhibit lower recall and higher variability across folds, as reflected by the larger standard deviations. This indicates that the model has greater difficulty consistently identifying minority classes, which is expected given their limited representation in the dataset (54 and 55 samples respectively). The high variability across folds for these classes is a direct consequence of the small sample sizes: slight changes in fold composition can substantially affect the class-conditional decision boundaries learned by the model, resulting in unstable performance estimates. This observation is important context for interpreting the domain adaptation results: even under ideal supervised conditions, the classification of minority classes remains inherently unreliable due to data scarcity.
The difference between macro-average and weighted-average metrics further highlights the influence of class imbalance. While the weighted averages remain close to the overall accuracy, the lower macro-average values indicate that performance is not uniform across classes.
Overall, these results provide a strong reference performance level for the subsequent domain adaptation experiments. Since the model is trained and evaluated directly on the IPO dataset, this configuration represents the maximum achievable performance under fully supervised target-domain conditions.

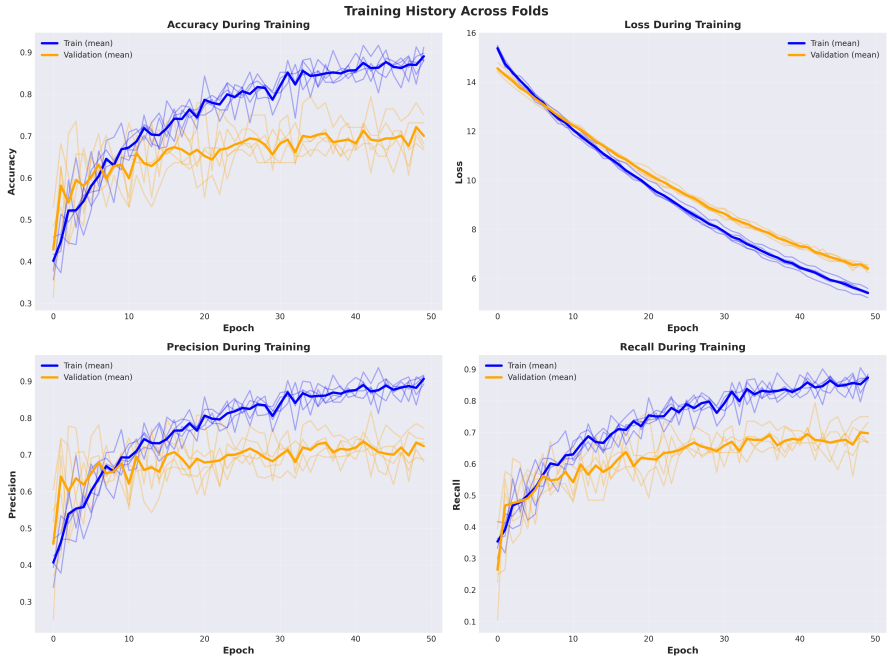


Figure 7.1: Average training dynamics across the cross-validation folds.

Figure 7.1 illustrates the average training behavior across folds. Training metrics increase steadily during the first epochs while the loss decreases consistently, indicating stable convergence of the optimization process. Validation metrics follow a similar trend but remain lower than the training curves, which is expected given the limited dataset size and the variability across folds. Overall, the curves suggest that the model converges smoothly without severe overfitting.

7.1.2 Zero-Shot Baseline

The zero-shot baseline evaluates the performance of the model trained exclusively on the source domain (TOGAS) and directly applied to the target domain (IPO) without any form of domain adaptation. In this configuration, the model architecture and training strategy described in the previous chapter were used without modifications.

The model was trained on the full TOGAS dataset and evaluated directly on the IPO dataset. No samples from the target domain were used during training. This experimental setup simulates a realistic deployment scenario in which a model trained on one clinical dataset is applied to data acquired from a different medical center without further adaptation.

The results of the zero-shot evaluation on the IPO dataset are reported in Table 7.2.

Class	Precision	Recall	F1-score	Support
0	0.2361	0.9444	0.3778	54
1	0.1707	0.1273	0.1458	55
2	0.9236	0.4754	0.6277	305
Macro Avg	0.4435	0.5157	0.3838	414
Weighted Avg	0.7339	0.4903	0.5311	414
Accuracy		0.4903		414

Table 7.2: Zero-shot performance obtained by training the model on the TOGAS dataset and directly evaluating it on the IPO dataset without domain adaptation.

The zero-shot experiment achieved an overall accuracy of 0.4903 on the IPO dataset. Compared with the upper bound performance obtained through cross-validation on the target domain, a substantial performance drop can be observed. This gap highlights the presence of a significant domain shift between the TOGAS and IPO datasets.
Class-wise results reveal a highly unbalanced behavior that is particularly informative about the nature of the domain shift. Class 0 exhibits very high recall (0.94) but extremely low precision (0.24), indicating that the model massively over-predicts this class on the target domain. This is consistent with the label distribution mismatch discussed in Chapter 5: TOGAS is dominated by class 0 samples (58.8%), so the source-trained model inherits a strong prior toward predicting normal mucosa. When applied to IPO, where class 0 represents only 13% of samples, this prior becomes a systematic source of error. Class 1 shows the lowest performance across all metrics, which is expected as it is a minority class in both datasets and the one most likely to be visually confused with the adjacent grades. Class 2 maintains relatively high precision but suffers from reduced recall, indicating that a large portion of true class 2 samples are misclassified, likely as class 0 due to the model's source-domain bias.
These results clearly demonstrate the difficulty of directly transferring a model trained on the source domain to the target dataset. The large performance gap between the zero-shot baseline and the upper bound motivates the use of domain adaptation strategies to improve generalization across datasets.

7.1.3 Self-Supervised Pseudo-Labeling

In addition to the methodological description provided in the previous chapter, a few implementation details were defined for the experimental setup. Training was performed with a batch size of 32 using the Adam optimizer with an initial learning rate of 10^{-4}. The first training phase on the source dataset was conducted for a

longer training schedule in order to obtain a stable initialization before starting the self-training procedure, while subsequent retraining phases were shorter as the model progressively incorporated pseudo-labeled samples.
The results obtained with the self-training pseudo-labeling strategy are reported in Table 7.3. The model achieved an overall accuracy of 0.7053 on the IPO dataset.

Class	Precision	Recall	F1-score	Support
0	0.3763	0.6481	0.4762	54
1	0.3000	0.1636	0.2118	55
2	0.8522	0.8131	0.8322	305
Macro Avg	0.5095	0.5416	0.5067	414
Weighted Avg	0.7168	0.7053	0.7033	414
Accuracy		0.7053		414

Table 7.3: Performance obtained using the self-supervised pseudo-labeling strategy on the IPO dataset.

Compared to the zero-shot baseline, the self-training approach substantially improves performance, increasing the overall accuracy from 0.49 to 0.71. This result indicates that pseudo-labeling effectively allows the model to exploit information from the target dataset and partially adapt to the target distribution.
Class-wise analysis shows that class 2 maintains strong performance with high precision and recall, reflecting the large number of available samples for this category. The iterative pseudo-labeling is particularly effective for class 2 precisely because it is the majority class in IPO: high-confidence predictions on this class are generated early and reliably, providing a stable signal for adaptation from the first iterations. Class 0 benefits from the self-training process, achieving a notable increase in recall compared to the zero-shot scenario, as the model progressively corrects its source-domain bias toward over-predicting normal mucosa. However, class 1 remains challenging, with relatively low recall and F1-score. This can be explained by a compounding of two factors: first, class 1 is a minority class in both datasets; second, it occupies a visually intermediate position between class 0 and class 2, making confident predictions unlikely and therefore reducing its representation in the pseudo-labeled pool. The confidence threshold of 0.9 used in this work effectively filters out ambiguous class 1 samples, inadvertently limiting the adaptation signal available for this category.
To further analyze the effect of pseudo-labeling on the feature representation, a UMAP projection of the learned embeddings was computed, as shown in Figure 7.2. The visualization compares manually labeled target samples with pseudo-labeled samples generated during the self-training process.

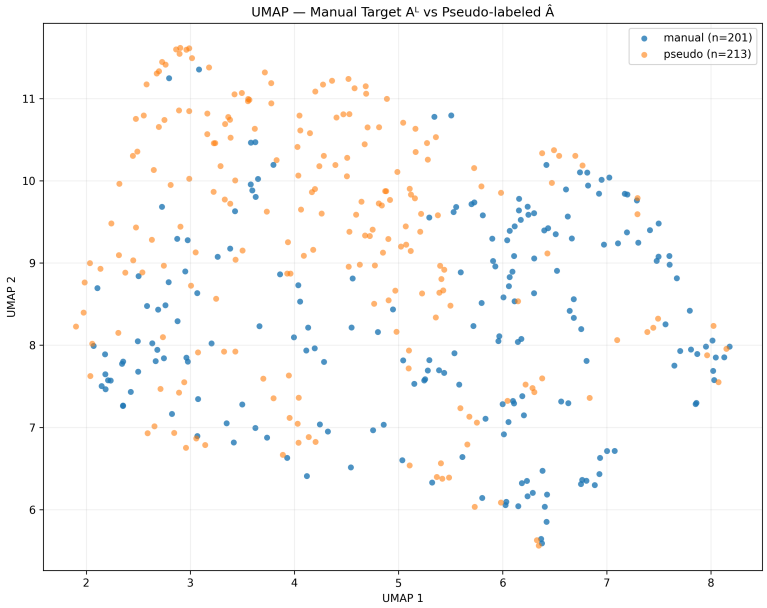


Figure 7.2: UMAP visualization of feature embeddings for manual target samples and pseudo-labeled samples obtained during self-training.

The projection reveals a partial overlap between manually labeled and pseudo-labeled samples, indicating that the pseudo-labeling process is able to capture relevant regions of the target feature space. However, pseudo-labeled samples appear more dispersed, suggesting that some predictions may still lie in regions of lower feature density.
This observation is supported by the quantitative neighborhood analysis. The median nearest-neighbor distance between pseudo-labeled samples and manual target samples (Pseudo→Manual NND = 21.32) is higher than the distance between manual samples themselves (Manual→Manual NND = 18.53), resulting in an NND ratio of approximately 1.15. This indicates that pseudo-labeled samples tend to lie slightly further from the core regions of the target distribution.
Additionally, the Local Density Ratio (LDR) analysis shows that approximately 66% of pseudo-labeled samples fall within low-density regions of the feature space. This suggests that a substantial portion of pseudo-labels are located near the boundaries of the learned representation, which may introduce noise into the training process.
Overall, the self-training strategy significantly reduces the domain gap observed in the zero-shot setting, improving classification performance while partially aligning the target feature distribution. However, the feature-space analysis indicates that pseudo-label quality remains imperfect, particularly for minority classes, leaving room for further improvements through more advanced domain adaptation strategies.

7.1.4 Domain-Adversarial Neural Network (DANN)

In order to evaluate adversarial domain adaptation, experiments were conducted using a Domain-Adversarial Neural Network (DANN). Two configurations were considered: a standard DANN training procedure and a variant initialized with a pretrained task classifier.
The general training configuration remained consistent with the settings described in the previous chapter. In particular, the same architecture, optimizer, batch size, and loss functions were used. The primary difference between the two experiments concerned the initialization of the task classifier and the scheduling of the Gradient Reversal Layer (GRL).
In the standard DANN configuration, the network was trained directly from ImageNet-initialized weights. The GRL coefficient was progressively increased during training through a scheduler, allowing the domain adversarial signal to gradually influence feature learning.
In the pretrained configuration, the task classifier was first initialized using weights obtained from the source-only training stage. To stabilize adversarial training, a warm-up phase was introduced in which the domain alignment signal was progressively activated over the first training epochs before reaching its maximum strength.
The total training duration was slightly reduced in the pretrained experiment, as the model already started from a task-specialized representation.
Table 7.4 reports the classification performance obtained using the standard DANN configuration.

Class	Precision	Recall	F1-score	Support
0	0.3303	0.6667	0.4417	54
1	0.2556	0.4182	0.3172	55
2	0.9163	0.6459	0.7577	305
Macro Avg	0.5007	0.5769	0.5055	414
Weighted Avg	0.7521	0.6184	0.6580	414
Accuracy		0.6184		414
AUC		0.7464		

Table 7.4: Classification performance obtained using the standard DANN configuration.

In addition to classification metrics, domain alignment was monitored through the domain classifier performance. The evolution of domain accuracy and domain AUC during training is shown in Figure 7.3.

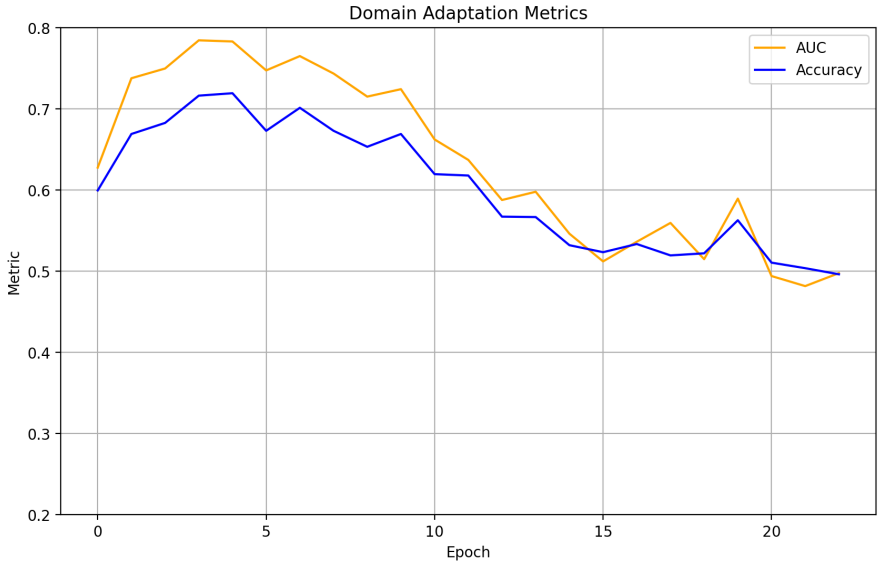
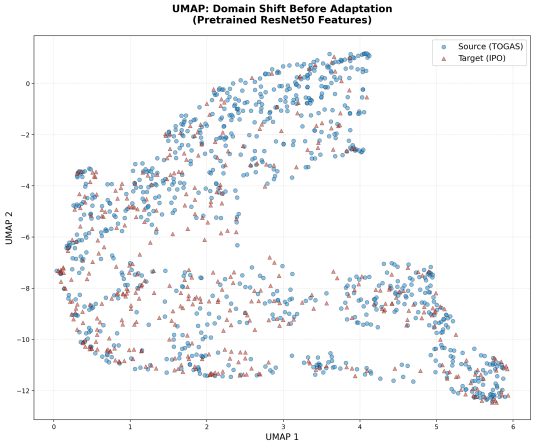
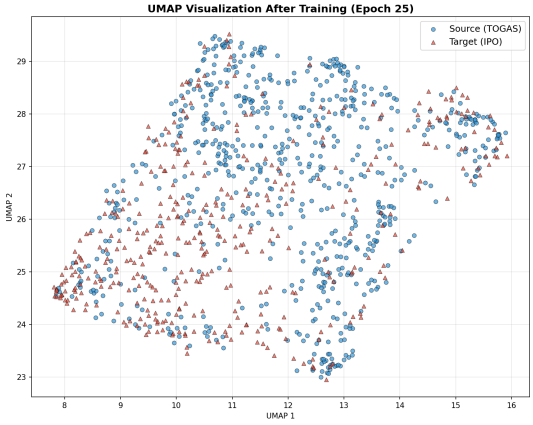


Figure 7.3: Evolution of domain classifier accuracy and AUC during standard DANN training.

The domain classifier converges toward chance-level performance (approximately 0.5), indicating that the learned feature representations become increasingly domain-invariant. This suggests that the adversarial training process successfully reduces domain-specific information in the shared feature space.
Despite this alignment, the overall classification accuracy remains lower than the upper bound obtained using fully supervised target training. This is a fundamental limitation of purely feature-level alignment: DANN optimizes for domain confusion, not for class discriminability in the target domain. When the label distributions differ substantially between source and target, as is the case here, with IPO heavily dominated by class 2, the domain-invariant features learned by DANN may not reflect the class boundaries that are relevant for the target distribution. In other words, the model learns to produce representations that look similar across domains, but the source-informed decision boundary may still be misaligned with the target class structure. Performance improvements compared to the zero-shot setting are nevertheless observed, indicating that adversarial alignment partially mitigates the domain shift between TOGAS and IPO datasets.
To further analyze the effect of adversarial training on the learned feature representations, a qualitative inspection of the feature space was performed using UMAP projections. The embeddings extracted from the shared feature layer were projected into a two-dimensional space both before and after DANN training.



(a) Feature distribution before adaptation



(b) Feature distribution after DANN training

Figure 7.4: UMAP visualization of the feature representations before and after adversarial domain adaptation. Source samples (TOGAS) are shown in blue, while target samples (IPO) are shown in red.

Before adaptation, the feature representations extracted from the pretrained ResNet50 backbone exhibit a noticeable distribution shift between source and target samples. Although partial overlap exists, several regions of the feature space remain dominated by one domain, indicating that the pretrained representation still retains domain-specific characteristics.
After DANN training, the separation between source and target distributions becomes less pronounced. Target samples appear more interleaved with the source clusters, suggesting that adversarial training encourages the feature extractor to learn representations that are less domain-dependent. This behavior is consistent with the domain classifier metrics observed during training, where domain accuracy approaches chance level.
However, the overlap between domains is not complete, and some regions of the feature space still exhibit partial separation. This residual structure may explain why the classification performance, although improved compared to the zero-shot baseline, remains below the upper bound obtained using fully supervised target training. Notably, domain confusion does not imply class alignment: even when source and target samples are indistinguishable to the domain discriminator, the class-conditional distributions may still differ, particularly for minority classes with few samples in the target domain.
The second experiment was conducted by initializing the task classifier with weights obtained from the source-only training stage before performing adversarial adaptation. The resulting classification performance is reported in Table 7.5.

Class	Precision	Recall	F1-score	Support
0	0.3139	0.7963	0.4503	54
1	0.1957	0.3273	0.2449	55
2	0.9459	0.5738	0.7143	305
Macro Avg	0.4852	0.5658	0.4698	414
Weighted Avg	0.7638	0.5700	0.6175	414
Accuracy		0.57		414
AUC		0.7557		

Table 7.5: Classification performance obtained using DANN with pretrained task head initialization.

The domain alignment behavior observed during training is shown in Figure 7.5.

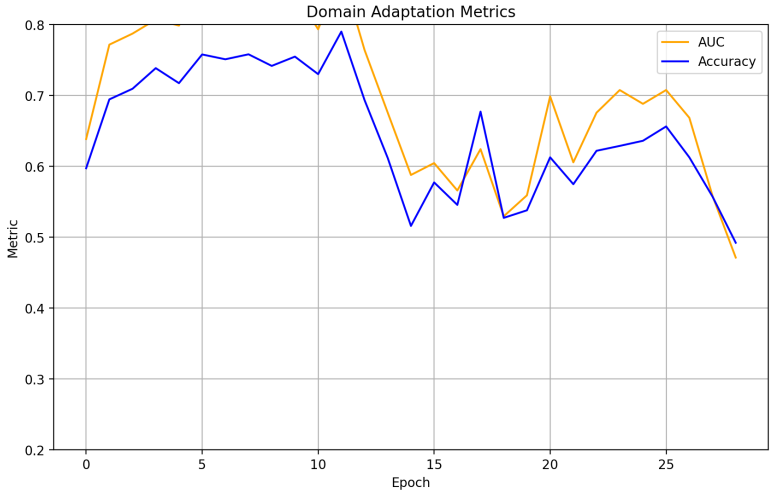


Figure 7.5: Domain classifier metrics during DANN training with pretrained task head initialization.

Interestingly, the pretrained configuration does not lead to improved performance compared to the standard DANN setup. Although the classifier starts from a task-specialized representation, this initialization may reduce the flexibility of the feature extractor during adversarial training. The pretrained task head already encodes strong source-domain decision boundaries, which may limit the ability of the adversarial process to reshape the representation space in order to better accommodate the target distribution. This creates a tension between two competing objectives: the task head pushes the feature extractor to preserve source-domain discriminative structure, while the gradient reversal layer pushes it toward domain invariance. When the task head is already strongly tuned to the source domain, this

tension is resolved in favor of the task objective, effectively reducing the influence of the adversarial signal. The result is a representation that is simultaneously less domain-invariant than the standard DANN variant and less class-discriminative on the target domain, leading to overall degraded performance.
As a result, while domain confusion remains close to chance level, the model may converge toward representations that remain partially biased toward the source domain. This phenomenon can limit the effectiveness of domain alignment and explain why the pretrained configuration does not outperform the standard adversarial training strategy.

7.1.5 Margin Disparity Discrepancy (MDD)

In addition to the methodological description provided in the previous chapter, a few experiment-specific settings were defined for the MDD experiments. Training was performed with a batch size of 32 using the Adam optimizer with an initial learning rate of 10^{-4}.
The discrepancy regularization parameter was set to $\lambda_{disc} = 1.0$, balancing the classification objective on the source domain with the adversarial discrepancy minimization on the target domain. Two optimizers were used during training: one for the main network parameters and one dedicated to the adversarial classifier responsible for maximizing the discrepancy. The second optimizer operated with a slightly higher learning rate to maintain a strong adversarial signal during optimization.
Each training batch contained a balanced mixture of source and target samples. Source samples contributed to the classification loss, while target samples contributed only to the discrepancy objective. The backbone feature extractor remained frozen during training to ensure stable optimization and prevent uncontrolled feature drift during the adversarial learning process.
Model selection was performed based on validation accuracy of the primary classifier, and final predictions on the target dataset were obtained by averaging the outputs of the two classifiers.
The classification results obtained using the MDD approach are reported in Table 7.6. The model achieved an overall accuracy of 0.6691 on the IPO dataset.

Class	Precision	Recall	F1-score	Support
0	0.3875	0.5741	0.4627	54
1	0.2740	0.3636	0.3125	55
2	0.8659	0.7410	0.7986	305
Macro Avg	0.5091	0.5596	0.5246	414
Weighted Avg	0.7249	0.6691	0.6902	414
Accuracy		0.6691		414
AUC		0.7654		
Final Discrepancy		0.3644		

Table 7.6: Classification performance obtained using the MDD domain adaptation framework.

Compared to the zero-shot baseline, the MDD approach significantly improves classification performance, demonstrating the effectiveness of discrepancy-based domain adaptation. The model achieves strong performance on class 2, which remains the dominant class in the dataset. Classes 0 and 1 show moderate improvements compared to the zero-shot scenario, although they remain more challenging due to their smaller sample sizes.
To further analyze the adaptation process, the evolution of the classifier discrepancy on target samples was monitored during training. The discrepancy curve is shown in Figure 7.6.

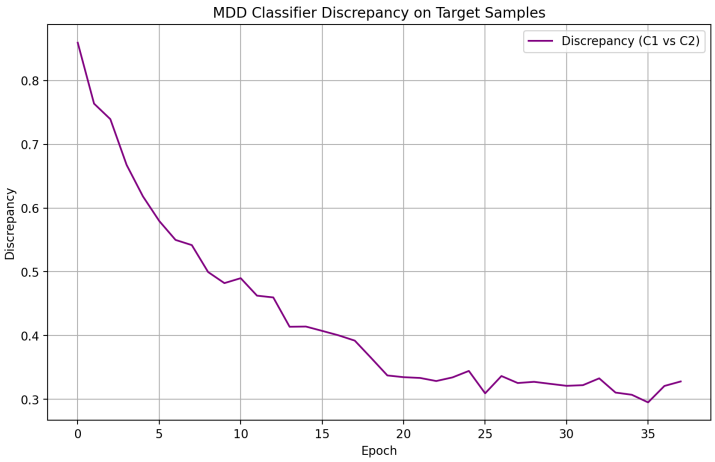
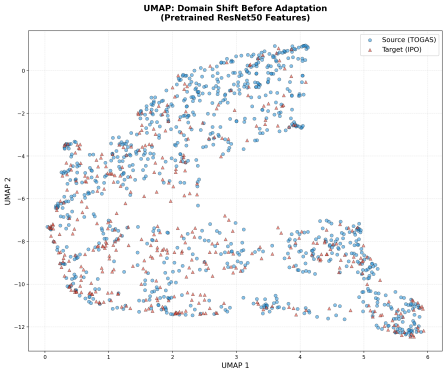


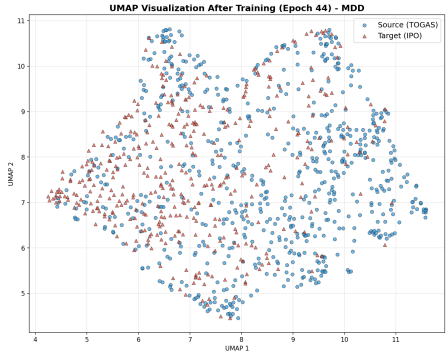
Figure 7.6: Evolution of classifier discrepancy on target samples during MDD training.

At the beginning of training, the discrepancy between the two classifiers is high, indicating that the target samples lie outside the decision regions supported by

the source-trained classifiers. As training progresses, the discrepancy gradually decreases, reflecting the ability of the feature extractor to adapt the representation space so that target samples become more consistent with the source decision boundaries.
The final discrepancy value stabilizes around 0.36, suggesting that the adversarial optimization successfully reduces disagreement between classifiers while preserving classification performance. This behavior confirms that the MDD training procedure effectively encourages domain-invariant representations. Notably, the non-zero residual discrepancy is not a failure of the method but rather an expected equilibrium: a perfect discrepancy of zero would indicate that the two classifiers produce identical predictions on all target samples, which would in turn suppress the adversarial gradient entirely and prevent further adaptation. The stabilization at 0.36 reflects a balanced regime in which the feature extractor has moved target representations toward the source decision boundary without fully collapsing the class structure.



(a) Feature distribution before adaptation



(b) Feature distribution after MDD training

Figure 7.7: UMAP visualization of the feature representations before and after MDD. Source samples (TOGAS) are shown in blue, while target samples (IPO) are shown in red.

Overall, the results indicate that discrepancy-based adaptation provides a stable mechanism for aligning feature distributions across domains. Although the achieved accuracy remains below the fully supervised upper bound, MDD successfully reduces the domain gap and improves performance compared to the zero-shot baseline. The slightly higher accuracy of MDD compared to standard DANN (0.6691 vs. 0.6184) suggests that operating directly on class-level disagreement provides a more task-aware alignment signal than domain-level confusion alone, which is agnostic to the class structure.

7.1.6 Comparison and considerations

Method	Accuracy	Precision	Recall	F1-score
Upper Bound	**0.8895**	**0.8473**	**0.7978**	**0.8037**
Zero-Shot	0.4903	0.4435	0.5157	0.3838
DANN	0.6184	0.5007	**0.5769**	0.5055
DANN + Pretrained Head	0.5700	0.4852	0.5658	0.4698
MDD	0.6691	0.5091	0.5596	**0.5246**
Self-Training	**0.7053**	**0.5095**	0.5416	0.5067

Table 7.7: Comparison of the evaluated approaches on the three-class EGGIM classification task. The upper bound represents the fully supervised performance on the target dataset. The best performing domain adaptation method is highlighted in green.

Table 7.7 summarizes the overall performance obtained by the different approaches evaluated in this study. Several relevant observations emerge from this comparison. First, the upper bound experiment clearly represents the best achievable performance when the model is trained directly on the target dataset. With an accuracy of 0.8895 and an F1-score of 0.8037, this configuration establishes a reference performance level and confirms that the classification task can be effectively solved when sufficient labeled target data are available.
In contrast, the zero-shot setting highlights the severity of the domain shift between the TOGAS and IPO datasets. When the model is trained exclusively on the source domain and evaluated directly on the target domain, accuracy drops to 0.4903 and the F1-score to 0.3838. This large performance gap is driven by two compounding factors: a covariate shift due to differences in acquisition equipment and protocols between the two clinical centers, and a label shift arising from the strongly different class distributions (TOGAS is dominated by class 0 with 58.8%, while IPO is dominated by class 2 with 73.7%). These two effects jointly produce the observed systematic misclassification patterns.
All domain adaptation methods improve over the zero-shot baseline, confirming the importance of explicit adaptation strategies. Among the adversarial approaches, the standard DANN configuration increases accuracy to 0.6184 and F1-score to 0.5055, indicating that domain-invariant feature learning partially reduces the domain gap. However, initializing the DANN model with a pretrained task head does not lead to further improvements. In fact, performance slightly decreases, suggesting that a classifier already specialized on the source domain may limit the flexibility of the feature extractor during adversarial alignment, as discussed in the previous section.
The MDD approach achieves stronger results than DANN, reaching an accuracy

of 0.6691 and an F1-score of 0.5246. This improvement over DANN can be attributed to the fundamentally different nature of the alignment signal: while DANN optimizes for domain confusion in a class-agnostic manner, MDD directly measures classifier disagreement on target samples, making the adaptation objective inherently sensitive to the class structure of the problem. This task-aware property appears to be particularly beneficial in the presence of significant label shift, as it encourages the feature extractor to move target representations toward regions that are discriminative for the relevant classes rather than merely domain-indistinct.
The best performing domain adaptation strategy in this study is the self-training approach, which achieves an accuracy of 0.7053. The superiority of self-training over the feature-alignment methods in this specific scenario can be explained by the nature of the domain shift itself. The combination of covariate and label shift present in the TOGAS→IPO adaptation means that feature alignment alone,whether through DANN or MDD, cannot fully correct the decision boundary, because the optimal boundary for IPO is structurally different from the one learned on TOGAS. Self-training, by contrast, directly supervises the model on target-domain samples using pseudo-labels, allowing it to implicitly adjust for both shifts simultaneously. When pseudo-labels are sufficiently accurate (as they are for the dominant class 2), the model receives a direct correction signal that no feature-level method can provide without access to target labels.
Overall, the results confirm that domain adaptation significantly improves cross-domain generalization in endoscopic image analysis. However, a substantial gap remains between unsupervised adaptation methods and the fully supervised upper bound, highlighting the intrinsic difficulty of the task and the persistent differences between source and target data distributions. A portion of this gap is attributable to the limited size of the datasets, particularly the small number of class 0 and class 1 samples in IPO, which prevents any adaptation method from reliably learning the minority class boundaries without direct supervision.

7.2 Binary Metaplasia Classification

7.2.1 Upper Bound

As in the previous experiments, an upper bound experiment was conducted to estimate the maximum achievable performance when the model is trained directly on the target domain. The training configuration remained identical to the one described in the previous chapter, with the only difference being the binary formulation of the classification task (Normal vs Metaplasia).
A patient-wise cross-validation strategy was applied to the target dataset to obtain a robust estimate of the achievable performance. Results are reported as mean and standard deviation across folds.

Class	Precision	Recall	F1-score	Support
0 (Normal)	0.5599 ± 0.1035	0.5247 ± 0.1117	0.5326 ± 0.0728	179
1 (Metaplasia)	0.7357 ± 0.0614	0.7555 ± 0.1049	0.7407 ± 0.0603	310
Macro Avg	0.6478 ± 0.0647	0.6401 ± 0.0569	0.6366 ± 0.0628	489
Weighted Avg	0.6739 ± 0.0606	0.6677 ± 0.0649	0.6642 ± 0.0631	489
Accuracy		0.6677 ± 0.0649		489

Table 7.8: Upper bound performance for the binary gastric metaplasia classification task obtained using cross-validation on the target dataset.

The upper bound experiment establishes the maximum achievable performance when the model is trained directly on the target domain. The model achieves an average accuracy of approximately 0.67, which is noticeably lower than the performance observed in the three-class EGGIM classification task (0.89).
This difference is not paradoxical despite binary being nominally a simpler problem: it reflects the fact that the EGGIM three-class task is dominated by a single class (class 2 accounts for 73.7% of IPO), making its accuracy inflated by correct majority-class predictions. The binary task, where the class distribution is more balanced (179 normal vs 310 metaplasia) and the visual distinction is more subtle, provides a more challenging and arguably more clinically informative evaluation scenario. The distinction between normal mucosa and intestinal metaplasia involves fine-grained textural and vascular pattern differences that are difficult even for experienced endoscopists, and this inherent visual ambiguity is reflected in the relatively modest upper bound.
The results also highlight an imbalance in class performance. The metaplasia class achieves substantially higher precision and recall compared to the normal class. This behavior may be partially explained by the higher number of metaplasia samples in the dataset as well as the presence of more distinctive visual patterns associated with metaplastic tissue, such as the characteristic light-blue crest pattern visible under NBI imaging.

7.2.2 Zero-Shot Baseline

To evaluate the magnitude of the domain shift between the source and target datasets, a zero-shot baseline was computed by training the model exclusively on the source dataset and directly evaluating it on the target dataset without any adaptation.
The same architecture and training configuration described previously were used, with the only difference being the binary classification objective.

Class	Precision	Recall	F1-score	Support
0 (Normal)	0.4423	0.8944	0.5919	180
1 (Metaplasia)	0.8603	0.3656	0.5132	320
Macro Avg	0.6513	0.6300	0.5525	500
Weighted Avg	0.7098	0.5560	0.5415	500
Accuracy		0.556		500
Balanced Accuracy		0.630		

Table 7.9: Zero-shot performance for the binary gastric metaplasia classification task.

The zero-shot baseline highlights the presence of a significant domain shift between the source and target datasets. When the model is trained exclusively on the source domain, the overall accuracy decreases to approximately 0.56.
The confusion matrix reveals a strong bias toward predicting the normal class. In particular, the model achieves very high recall for the normal class (0.89) but substantially lower recall for the metaplasia class (0.37). This asymmetric behavior is consistent with the domain shift between TOGAS and ERASMUS: the ERASMUS dataset was acquired using FUJIFILM ELUXEO endoscopes with blue-light imaging modality, which produces images with a substantially different color profile and texture appearance compared to the Olympus NBI system used for TOGAS. As a result, metaplasia patterns in ERASMUS may not match the visual features the model learned to associate with metaplasia in the source domain, causing the model to default to the majority source-domain class. This is also reflected in the Wasserstein distance of 1.32 between TOGAS and ERASMUS, the largest among all dataset pairs, confirming the stronger distributional divergence of this adaptation scenario.
Compared to the upper bound experiment, the performance gap confirms that domain adaptation is necessary to improve cross-dataset generalization in this binary classification setting.

7.2.3 Self-Supervised Pseudo-Labeling

A self-training strategy based on iterative pseudo-labeling was also evaluated for the binary classification task. The same procedure described previously for the multi-class EGGIM problem was applied, with the only difference being the binary formulation of the classification objective.
The model was initially trained on the source dataset and subsequently used to generate predictions on the target domain. At each iteration, only high-confidence predictions were selected and incorporated into the training set as pseudo-labeled

samples. This process progressively expands the training dataset with target-domain data, allowing the model to adapt its decision boundaries to the target distribution.

Class	Precision	Recall	F1-score	Support
0 (Normal)	0.4561	0.8944	0.6041	180
1 (Intestinal Metaplasia)	0.8707	0.4000	0.5482	320
Macro Avg	0.6634	0.6472	0.5762	500
Weighted Avg	0.7215	0.5780	0.5683	500
Accuracy		0.578		
Balanced Accuracy		0.6472		
AUC		0.6766		

Table 7.10: Binary classification performance obtained using the self-training approach.

To analyze how pseudo-labeled samples are distributed within the target feature space, a UMAP visualization was generated comparing manually labeled target samples and pseudo-labeled samples (Figure 7.8).

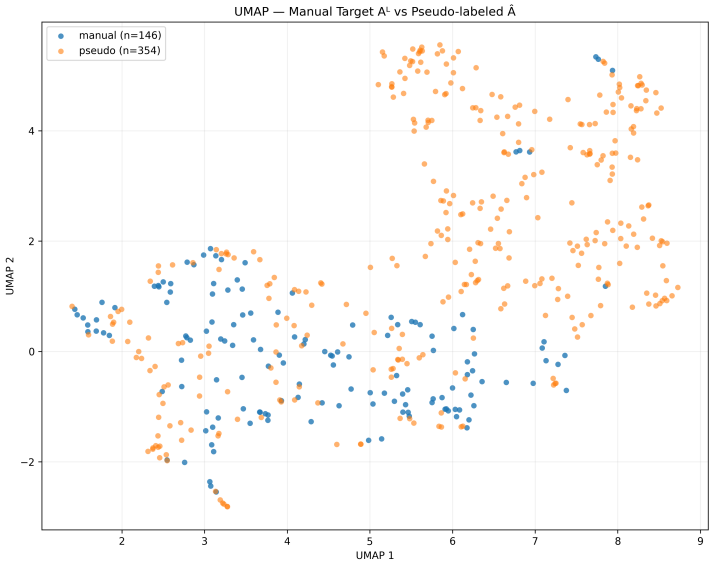


Figure 7.8: UMAP visualization comparing manually labeled target samples and pseudo-labeled samples generated during the self-training process.

The projection shows that pseudo-labeled samples occupy a broader portion of the feature space compared to the manually labeled target samples. While a certain

degree of overlap is present, many pseudo-labeled samples appear in regions where the density of manually labeled samples is relatively low.
To further investigate this phenomenon, nearest-neighbor distance (NND) statistics were computed. The median distance between manually labeled samples is approximately 18.70, whereas the median distance between pseudo-labeled samples and manually labeled samples increases to 26.01. The resulting NND ratio of 1.39 indicates that pseudo-labeled samples tend to lie farther from dense regions of the manual target distribution.
This observation is reinforced by the low-density ratio (LDR), which shows that approximately 84.7% of pseudo-labeled samples are located in low-density regions of the feature space. The LDR for the binary task (84.7%) is markedly higher than that observed for the three-class task (66%), which provides a structural explanation for the contrasting performance of self-training across the two scenarios. In the TOGAS→IPO setting, the strong representation of class 2 in the target domain allowed the pseudo-labeling to accumulate a large volume of high-quality samples for the dominant class, generating a reliable adaptation signal. In the TOGAS→ERASMUS setting, the stronger distributional shift (Wasserstein distance of 1.32 vs. 0.86) means that the source model produces fewer high-confidence predictions on target samples overall, and those it does produce are disproportionately concentrated near the normal class — which the model already biases toward. As a result, pseudo-labeling provides little additional benefit for the metaplasia class, where adaptation is most needed.
When comparing these results with the previous experiments, the self-training strategy slightly improves over the zero-shot baseline. The accuracy increases from 0.556 in the zero-shot setting to 0.578 with pseudo-labeling, indicating that incorporating target-domain samples can provide additional useful information for the model.
However, the achieved performance remains below the upper bound obtained when training directly on the target dataset, which reaches approximately 0.668 accuracy. This remaining gap suggests that the quality of pseudo-labels is not sufficient to fully replicate the benefits of supervised training on the target domain.
Overall, the results indicate that self-training provides a modest improvement over the zero-shot baseline by exploiting additional target-domain samples. Nevertheless, the presence of pseudo-label noise and the tendency of pseudo-labeled samples to occupy low-density regions of the feature space limit the overall effectiveness of this strategy, particularly when the domain shift is large and the source model's predictions on target samples are unreliable.

7.2.4 Domain Adversarial Neural Network (DANN)

For the binary classification task, the same DANN architecture described in the previous section was employed. The overall training configuration remained largely unchanged, with the exception of two modifications aimed at improving generalization and stabilizing adversarial training.
First, an online data augmentation pipeline was introduced during training. Instead of using a fixed augmented dataset, transformations were applied dynamically at each epoch. This strategy allows the model to observe different variants of the same image during training, which helps reduce overfitting and encourages the learning of more robust representations.
Importantly, the same augmentation pipeline was applied to both source and target samples (`augment_source=True` and `augment_target=True`). This design choice ensures that the augmentation process does not artificially increase the visual differences between the two domains. The applied transformations were intentionally mild, including small brightness variations (± 0.08), contrast adjustments (0.85–1.15), and limited hue perturbations (± 0.02). These conservative transformations were selected to preserve the medical visual characteristics of the images while still introducing useful variability.
Additionally, the batch size was increased to 64. Larger batches were found to stabilize adversarial training by providing more consistent domain gradients for the domain discriminator, which is particularly important when optimizing the gradient reversal layer.
Table 7.11 reports the results obtained using the standard DANN configuration.

Class	Precision	Recall	F1-score	Support
0 (Normal)	0.4894	0.7722	0.5991	180
1 (Metaplasia)	0.8102	0.5469	0.6530	320
Macro Avg	0.6498	0.6595	0.6261	500
Weighted Avg	0.6947	0.6280	0.6336	500
Accuracy		0.628		
Balanced Accuracy		0.6595		
AUC		0.6826		

Table 7.11: Binary classification performance obtained using the standard DANN configuration.

The standard DANN configuration achieves an image-level accuracy of 0.628, which represents a clear improvement over the zero-shot baseline (0.556). This improvement confirms that adversarial domain adaptation is able to partially mitigate the distribution shift between the source and target datasets. By encouraging the

feature extractor to learn domain-invariant representations through the gradient reversal mechanism, the model becomes less sensitive to dataset-specific visual characteristics.
At the same time, the achieved performance remains below the upper bound accuracy of approximately 0.668 obtained when training directly on the target dataset. This remaining performance gap highlights that, although the adversarial alignment reduces the discrepancy between domains, the learned features still retain some residual domain-specific information that limits full transferability.
A more detailed inspection of the class-wise metrics provides additional insights. The model achieves a high recall for the normal class (0.77) but a lower recall for the metaplasia class (0.55). This imbalance suggests that the decision boundary learned during adaptation still favors the majority patterns observed in the source domain, leading to a tendency to over-predict the normal class. This is a structural limitation of domain-level adversarial alignment: by enforcing indistinguishability between domains without conditioning on class labels, DANN may inadvertently align regions of the feature space corresponding to different classes, a phenomenon known as negative transfer or class misalignment. A class-conditional variant such as CDAN could mitigate this effect by ensuring that features from the same class are aligned across domains rather than features from the same domain regardless of class.
Interestingly, the balanced accuracy (0.6595) approaches the upper bound performance more closely than the raw accuracy metric. This indicates that, despite the moderate overall accuracy improvement, the adversarial training helps maintain a more balanced discrimination between the two classes.
Overall, these results suggest that the DANN framework successfully reduces the domain gap, but does not completely eliminate it. The model benefits from adversarial alignment, yet the intrinsic differences between the source and target datasets still limit the achievable performance compared to the fully supervised target-domain training scenario.
Figure 7.9 illustrates the evolution of the domain classifier metrics during training.

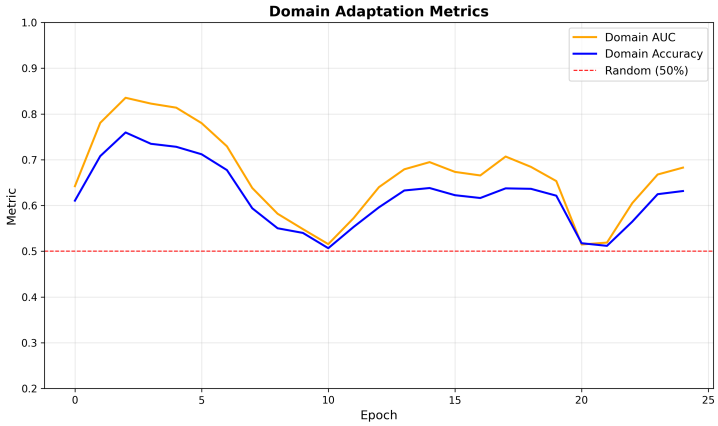


Figure 7.9: Domain classifier accuracy and AUC during DANN training for the binary classification task.

Ideally, successful domain adaptation should push the domain classifier toward chance-level performance (approximately 0.5), indicating that the feature extractor is producing domain-invariant representations. In this experiment, the domain accuracy initially rises above 0.7 during the early epochs, suggesting that the domain discriminator can clearly distinguish between the source and target domains. As training progresses, the domain accuracy gradually decreases and approaches values closer to 0.5, indicating partial domain confusion.
Although perfect domain invariance is not achieved, this reduction suggests that the adversarial training process encourages the feature extractor to reduce domain-specific cues in the learned representations.
To further analyze the effect of domain adaptation, UMAP visualizations were generated before and after training, as shown in Figure 7.10.

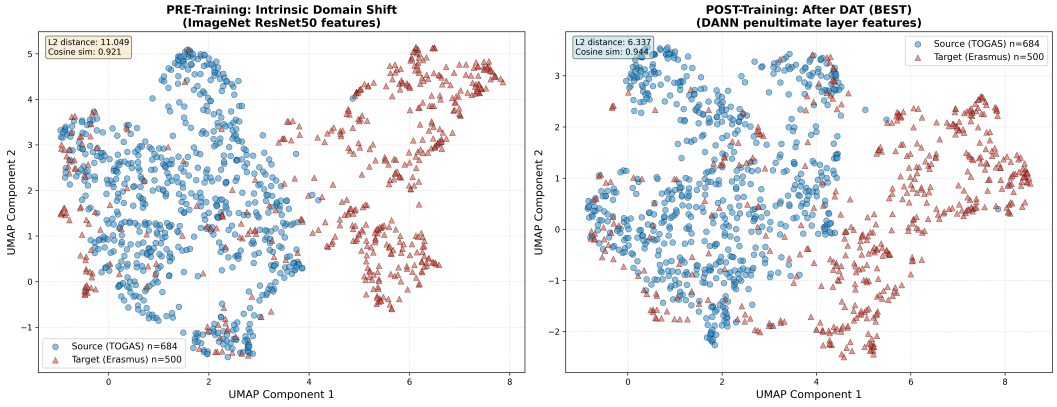


Figure 7.10: UMAP visualization of the feature space before and after DANN training.

The pre-training visualization clearly reveals a strong separation between source and target samples, confirming the presence of a significant domain shift. This is also reflected by the large L2 distance between the domain centroids (11.049), despite the relatively high cosine similarity (0.921), which indicates that the global orientation of the feature spaces is similar but spatially shifted. This pattern is consistent with the Wasserstein distance analysis reported in Chapter 5: TOGAS and ERASMUS share a similar embedding direction (high cosine similarity) but differ substantially in their distributional mass (high Wasserstein distance), suggesting that the domain shift is primarily spatial rather than structural.
After adversarial training, the L2 distance decreases substantially to 6.337, while the cosine similarity increases slightly to 0.944. This behavior suggests that the DANN training process successfully reduces the spatial separation between the two domains while maintaining a consistent feature orientation. Visually, this is reflected in the increased overlap between source and target clusters in the feature space.
A second experiment was conducted by initializing the task classifier with pretrained weights before adversarial training. The goal of this configuration was to provide a stronger task-specific initialization that could potentially stabilize the early stages of training.

Class	Precision	Recall	F1-score	Support
0 (Normal)	0.4613	0.8611	0.6008	180
1 (Metaplasia)	0.8476	0.4344	0.5744	320
Macro Avg	0.6544	0.6477	0.5876	500
Weighted Avg	0.7085	0.5880	0.5839	500
Accuracy		0.588		
Balanced Accuracy		0.6477		
AUC		0.6812		

Table 7.12: Binary classification performance using DANN with pretrained task head initialization.

Overall, the standard DANN configuration achieves slightly better performance than the pretrained variant. While the pretrained model provides a strong initial task representation, it may also constrain the flexibility of the feature extractor during adversarial alignment. This result is consistent with the analogous finding on the three-class task, suggesting that the negative effect of pretrained task head initialization is not specific to one task but reflects a more general tension between source-domain specialization and domain-adaptive flexibility.
When the classifier head is already specialized for the source domain, the adversarial

training process may struggle to significantly modify the learned decision boundaries, leading to weaker domain adaptation. In contrast, training the classifier jointly with the adversarial objective allows the feature extractor to adapt more freely to the target domain.
These results suggest that, in this setting, allowing the task head to co-evolve with the adversarial training process provides a more effective adaptation strategy than relying on a fixed pretrained initialization.

7.2.5 Margin Disparity Discrepancy (MDD)

In addition to adversarial domain confusion, a discrepancy-based domain adaptation method was also evaluated using the Margin Disparity Discrepancy (MDD) framework. As described in the methodology chapter, this approach relies on two classifiers that compete over target samples: one classifier attempts to maximize the prediction discrepancy on target data, while the feature extractor learns to minimize this disagreement.
This adversarial interaction allows the model to identify regions of uncertainty in the target domain and progressively align the feature representation across domains.

Class	Precision	Recall	F1-score	Support
0 (Normal)	0.4654	0.8222	0.5944	180
1 (Intestinal Metaplasia)	0.8242	0.4688	0.5976	320
Macro Avg	0.6448	0.6455	0.5960	500
Weighted Avg	0.6950	0.5960	0.5964	500
Accuracy		0.596		
Balanced Accuracy		0.6455		
AUC		0.6671		

Table 7.13: Binary classification performance obtained using the MDD domain adaptation framework.

Figure 7.11 illustrates the evolution of the classifier discrepancy on the target domain during training.

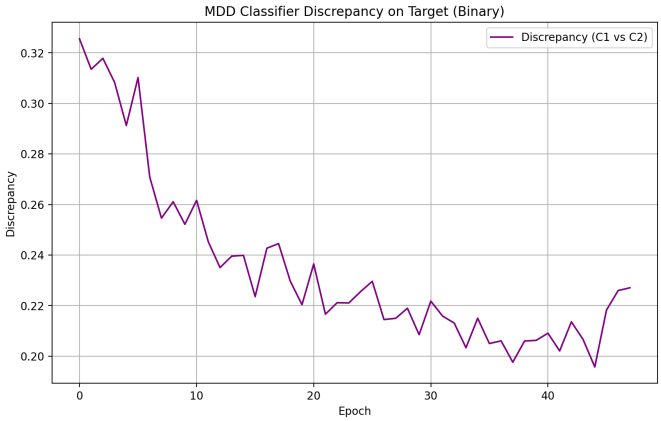


Figure 7.11: Evolution of classifier discrepancy during MDD training on the binary classification task.

At the beginning of training, the discrepancy between the two classifiers is relatively high, indicating that the two models disagree substantially on target samples. As training progresses, the discrepancy gradually decreases and stabilizes around a value of approximately 0.227. This reduction indicates that the feature extractor successfully learns to minimize classifier disagreement, producing a more consistent representation for target samples.
To further investigate the effect of discrepancy-based alignment, UMAP visualizations were generated before and after MDD training (Figure 7.12).

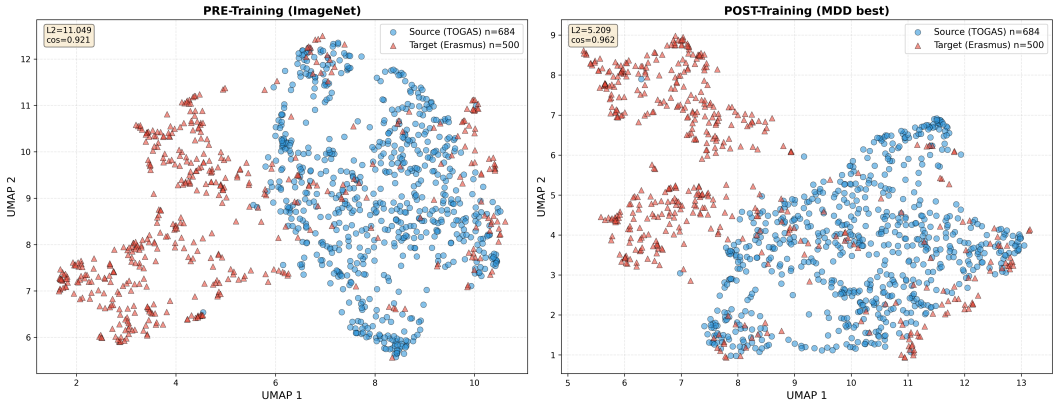


Figure 7.12: UMAP visualization of the feature space before and after MDD training.

Before training, the feature representations extracted from the pretrained backbone show a strong separation between source and target samples. This separation is reflected by a large L2 distance between the domain centroids (11.049), despite the

relatively high cosine similarity (0.921), which indicates that the two domains share a similar global orientation but remain spatially separated in the feature space.
After MDD training, the L2 distance decreases significantly to 5.209 while the cosine similarity increases to 0.962. This indicates that the discrepancy-based optimization successfully reduces the spatial gap between the two domains and aligns their representations more closely. Notably, MDD achieves a greater reduction in L2 distance than DANN (5.209 vs. 6.337), suggesting that discrepancy minimization produces a stronger geometric alignment of the feature spaces. However, this more aggressive spatial alignment does not translate into better classification performance, indicating that geometric proximity in feature space is a necessary but not sufficient condition for effective domain adaptation: the class-conditional alignment also matters.
Visually, the UMAP projection shows that the source and target samples become more interleaved, suggesting that the feature extractor has learned more domain-invariant representations.
When comparing the obtained results with the previous experiments, several observations can be made.
First, the MDD model improves over the zero-shot baseline, increasing the image-level accuracy from 0.556 to 0.596. This confirms that discrepancy-based adaptation is capable of reducing the domain gap between the source and target datasets.
However, the improvement remains smaller than the one observed with the DANN approach, which achieved an accuracy of 0.628. The reversal of the MDD vs. DANN ranking between the two tasks is a noteworthy finding: MDD outperforms DANN on the three-class task (0.6691 vs. 0.6184) but underperforms it on the binary task (0.596 vs. 0.628). This inversion can be interpreted in light of the different nature of the two problems. In the three-class task, the class-sensitive alignment of MDD is advantageous because the label shift requires the model to distinguish among three categories with very different prevalences. In the binary task, the simpler decision boundary means that domain-level confusion enforced by DANN is sufficient to realign the feature spaces effectively, while the discrepancy signal in MDD may be noisier in the binary setting due to the higher ambiguity between the two classes.
At the same time, both methods remain below the upper bound performance obtained when training directly on the target dataset (approximately 0.668). This remaining gap indicates that, although domain adaptation reduces distribution mismatch, the source and target datasets still exhibit intrinsic differences that cannot be fully compensated without access to labeled target data.
Interestingly, the class-wise metrics reveal a behavior similar to the one observed with DANN. The model achieves very high recall for the normal class (0.82) but significantly lower recall for the metaplasia class (0.47). This pattern suggests that the classifier still tends to favor predictions toward the normal class, reflecting residual bias introduced by the source-domain distribution.

Overall, these results confirm that discrepancy-based domain adaptation successfully reduces the domain shift but does not completely eliminate it. While the MDD framework effectively aligns feature distributions, adversarial domain confusion appears slightly more effective for this specific binary classification task.

7.2.6 Comparison and considerations

Method	Accuracy	Precision	Recall	F1-score
Upper Bound	**0.6677**	**0.6478**	**0.6401**	**0.6366**
Zero-Shot	0.556	0.6513	0.6300	0.5525
DANN	**0.628**	0.6498	**0.6595**	**0.6261**
MDD	0.596	0.6448	0.6455	0.5960
Self-Training	0.578	**0.6634**	0.6472	0.5762

Table 7.14: Comparison of all evaluated methods for the binary classification task. The upper bound represents the fully supervised performance on the target dataset, while the best domain adaptation method is highlighted in green.

Table 7.14 summarizes the overall performance obtained by all evaluated methods for the binary classification task.
As expected, the upper bound experiment provides the best achievable performance, reaching an accuracy of 0.6677 and a macro F1-score of 0.6366. This result represents the reference performance when the model is trained directly on labeled target-domain data. The relatively modest upper bound itself (0.67) is informative: it confirms that the visual distinction between normal mucosa and intestinal metaplasia in white-light and blue-light endoscopy is intrinsically ambiguous, and that performance limitations are not solely attributable to domain shift but also to the difficulty of the underlying visual task.
The zero-shot baseline highlights the presence of a domain shift between the source and target datasets. When the model is trained only on the source domain and evaluated directly on the target domain, the accuracy drops to 0.556 and the F1-score to 0.5525. This gap confirms that models trained exclusively on the source domain do not generalize optimally to the target dataset.
Among the domain adaptation approaches, the DANN framework achieves the best overall performance, reaching an accuracy of 0.628 and the highest macro F1-score among the adaptation methods (0.6261). The superiority of DANN in this scenario, compared to its underperformance relative to MDD in the three-class task, highlights the interaction between task complexity and adaptation strategy. For binary classification with a larger domain shift (TOGAS→ERASMUS, Wasserstein 1.32), enforcing strong domain-level feature confusion through gradient reversal is

sufficient and effective. For the three-class task with additional label shift, a more class-aware approach like MDD or direct pseudo-supervision through self-training becomes necessary to correctly realign the class-conditional distributions.
The MDD method also improves over the zero-shot baseline, achieving an accuracy of 0.596 and a macro F1-score of 0.5960. Although discrepancy-based adaptation successfully reduces domain mismatch, its performance remains slightly lower than the adversarial alignment achieved by DANN.
The self-training approach produces the highest precision among the domain adaptation methods (0.6634). This behavior suggests that pseudo-labeling helps the model make more confident predictions for the classes it does correctly classify. However, the overall accuracy and F1-score remain lower than those achieved by DANN and MDD. This is a direct consequence of the higher LDR observed in this task (84.7% of pseudo-labeled samples in low-density regions), which indicates that the pseudo-label quality in the TOGAS→ERASMUS scenario is substantially lower than in the TOGAS→IPO setting. When the source model cannot produce reliable predictions on target samples, as happens when the domain shift is large, self-training amplifies errors rather than correcting them.
Overall, these results indicate that adversarial domain adaptation provides the most effective strategy for this binary classification problem. The consistent ordering of methods across both tasks (all adaptation methods outperform zero-shot, all remain below upper bound) supports the validity of the experimental setup and the robustness of the findings. The differential behavior of individual methods across the two tasks reflects not arbitrary variance but principled interactions between task structure, domain shift type, and adaptation mechanism.

Chapter 8

Discussion

The experiments presented in the previous chapter provide several insights into the effectiveness of different domain adaptation strategies for gastric endoscopic image classification. In particular, the results highlight how the behavior of the evaluated methods varies depending on the task formulation and the characteristics of the underlying data distributions. A central contribution of this work is the systematic evaluation of multiple adaptation strategies across two structurally different scenarios, enabling cross-task comparisons that would not be possible with a single benchmark.

Across both tasks, all domain adaptation methods consistently improved performance compared to the zero-shot baseline. This confirms the presence of a significant domain shift between the source dataset (TOGAS) and the target datasets (IPO and ERASMUS), and demonstrates the necessity of explicit adaptation mechanisms when transferring models across clinical sites that differ in acquisition equipment, imaging modality, and patient population.

In the three-class EGGIM classification task, the self-training approach achieved the best overall performance among the domain adaptation methods. The iterative pseudo-labeling strategy allowed the model to progressively incorporate target-domain samples into the training process, resulting in a substantial reduction of the domain gap. The ability to leverage high-confidence pseudo-labels enabled the model to adapt its decision boundaries more effectively to the target distribution. Crucially, this task is characterized by a combined covariate and label shift: the class prior distributions of TOGAS and IPO are markedly different (58.8% vs. 13.0% for class 0; 19.5% vs. 73.7% for class 2). Feature-level alignment methods such as DANN and MDD can reduce the covariate shift component, but they cannot directly correct for label shift, as they lack access to target-domain class information. Self-training, by contrast, implicitly corrects for label shift by providing the model with pseudo-supervised signals that reflect the actual class distribution of the target domain, provided the pseudo-labels are sufficiently accurate.

In contrast, for the binary classification task, the DANN framework achieved the strongest performance. Adversarial domain alignment allowed the model to learn domain-invariant representations, which proved particularly effective for the simpler two-class decision boundary. In this scenario, the domain shift is predominantly of covariate type rather than label shift type: both TOGAS and ERASMUS exhibit a moderate imbalance in the same direction (metaplasia more frequent than normal), so the main challenge is adapting to the different visual appearance of the images rather than correcting for different class priors. This structural difference explains why feature-level alignment through DANN suffices in the binary case: by pushing source and target representations into a shared domain-invariant space, the model correctly adapts its boundary to the new visual distribution without needing access to target-domain labels. While discrepancy-based adaptation (MDD) also reduced the domain gap, its performance remained slightly below that of the adversarial approach, consistent with the observation that the simpler binary boundary benefits less from the class-sensitive alignment that MDD provides.

These results offer a principled interpretation of when each method is most appropriate: self-training and pseudo-labeling are advantageous when label shift is present and pseudo-label quality is high, which requires moderate domain shift and a dominant target class; feature-level adversarial alignment is effective when covariate shift is the primary concern and the decision boundary structure is relatively consistent between domains; discrepancy-based methods occupy an intermediate position, being particularly useful in multi-class scenarios with moderate label shift, where their class-aware alignment signal provides an advantage over purely domain-level confusion.

A direct comparison between the two tasks reveals additional insights into the nature of the domain shift.

For the three-class EGGIM classification problem, the gap between the zero-shot baseline (0.49) and the upper bound (0.89) was particularly large, indicating that the domain shift strongly affected model generalization. The magnitude of this gap is driven not only by the covariate shift between TOGAS and IPO (Wasserstein distance 0.86) but primarily by the severe label shift. The source model, having learned a strong prior toward class 0 from TOGAS, produces predictions that are systematically misaligned with the class distribution of IPO, where class 2 represents nearly three quarters of the samples. This structural mismatch means that any adaptation method that does not directly account for the target class distribution will be limited in how much it can close the gap. The fact that self-training reduces the gap by approximately 43% (from 0.49 to 0.71) is substantial, but the remaining distance to the upper bound reflects the inherent difficulty of correcting label shift without direct target supervision, especially for the minority classes (54 and 55 samples in IPO for classes 0 and 1 respectively).

In contrast, the binary classification task exhibited a smaller absolute domain gap

between the zero-shot (0.56) and upper bound (0.67) settings. However, the overall performance remained lower even in the fully supervised scenario, suggesting that the binary classification problem itself is intrinsically more challenging. The visual distinction between normal mucosa and intestinal metaplasia is more subtle than the graded differences captured by the EGGIM score, and the use of heterogeneous imaging modalities (NBI in TOGAS vs. BLI in ERASMUS) further complicates the extraction of modality-invariant pathological features. The smaller absolute gap in this scenario reflects the modest upper bound rather than a smaller domain shift.

Another notable observation is the recurring class imbalance in recall values. Across several experiments and both tasks, the models tended to achieve high recall for the normal class while struggling to correctly identify metaplasia samples. This is not merely a consequence of class weighting choices but reflects a deeper structural issue: the source domain (TOGAS) is dominated by normal mucosa samples in the binary formulation, so the source model develops a strong prior toward the normal prediction. When transferred to the target domain, this prior persists unless directly corrected by target-domain supervision or by effective domain alignment. The fact that even DANN, which achieves the best binary performance, shows a recall of 0.77 for normal vs. 0.55 for metaplasia suggests that complete correction of this bias requires labeled target-domain data.

8.1 Feature Space Alignment

The representation-level analyses performed using UMAP visualizations further support the quantitative results. In several experiments, the feature spaces of the source and target domains initially exhibited a clear separation, confirming the presence of a domain shift. The pre-adaptation L2 centroid distance of 11.049 in the binary task reflects the stronger distributional divergence between TOGAS and ERASMUS, while the more moderate shift in the three-class task is consistent with the lower Wasserstein distance of 0.86 between TOGAS and IPO.

After domain adaptation, both adversarial and discrepancy-based methods reduced this separation, as reflected by the decrease in L2 distance between domain centroids and the increase in cosine similarity. This indicates that the learned feature representations became more aligned across domains.

However, complete overlap between the two domains was rarely achieved. This observation suggests that certain dataset-specific characteristics, such as differences in imaging equipment, acquisition protocols, or patient populations, introduce variations that are difficult to fully eliminate through representation alignment alone. An important nuance is that domain overlap in the UMAP projection does not guarantee class-conditional alignment: two domains can be geometrically

close in feature space while still having different class-conditional distributions, particularly when the classes themselves are visually ambiguous or when the sample sizes per class are small. This distinction helps explain why MDD achieves lower L2 distances after adaptation than DANN in the binary task (5.209 vs. 6.337) but yields lower classification performance: stronger geometric alignment does not imply better class discrimination when the alignment is not class-aware.

8.2 Limitations

Despite the improvements obtained through domain adaptation, several limitations remain.

First, the overall dataset size remains relatively limited. The IPO dataset contains only 414 samples distributed across 247 patients, with minority classes represented by just 54 and 55 samples. This scarcity directly limits the reliability of class-level performance estimates, as evidenced by the high standard deviations across folds in the upper bound experiment, and constrains the ability of adaptation methods to learn robust representations for underrepresented categories. It is worth noting that this is not a methodological weakness but a reflection of the real-world difficulty of collecting and annotating multi-center endoscopic data: obtaining high-quality images with expert EGGIM annotations from clinical practice is intrinsically expensive and time-consuming. The limited dataset sizes therefore represent an honest characterization of the challenge rather than a limitation of the experimental design.

Second, pseudo-labeling approaches inherently introduce the risk of label noise. As observed in the feature space analysis, many pseudo-labeled samples were located in low-density regions of the representation space, particularly in the binary task where the LDR reached 84.7%. This noise propagation effect is amplified when the source model is already biased toward certain classes, as it will tend to generate high-confidence pseudo-labels preferentially for the dominant source-domain class, inadvertently reinforcing the existing bias rather than correcting it.

Third, the experiments were conducted using a single backbone architecture (ResNet50). Although this architecture is widely used and provides strong baseline performance, different architectures or more recent vision models may exhibit different adaptation behavior. Vision Transformers in particular have been shown to learn representations that are more robust to domain shift, and their evaluation in this context would constitute a natural extension of this work.

Finally, domain adaptation remains fundamentally challenging in medical imaging due to the complexity of visual patterns and the variability of clinical acquisition settings. Even with adaptation mechanisms, a noticeable gap remains between unsupervised adaptation performance and the fully supervised upper bound. This

gap is partially intrinsic to the problem: the visual distinction between healthy and metaplastic mucosa is challenging even for experienced clinicians, and no adaptation algorithm can recover information that is simply not present in unlabeled target-domain images.

8.3 Challenges Encountered During Development

During the development of the experiments, several practical challenges were encountered.
One major challenge was ensuring stable training dynamics for adversarial methods such as DANN. The balance between the task classifier and the domain discriminator required careful tuning of hyperparameters, including the gradient reversal strength and learning rate schedules. In particular, an excessively strong adversarial signal early in training was found to destabilize the task classifier before it could learn meaningful representations, while a too-weak signal failed to produce domain confusion. The progressive scheduling of the GRL coefficient was essential to navigate this trade-off.
Additionally, managing pseudo-label accumulation in the self-training framework required careful control to prevent the introduction of excessive noise. Confidence thresholds and stopping criteria were therefore necessary to ensure that only reliable pseudo-labels were incorporated into the training process. The interaction between confidence threshold and class imbalance was particularly delicate: a high threshold preserves quality but reduces coverage, especially for minority classes; a low threshold increases coverage but risks accumulating noisy labels for visually ambiguous samples.
Finally, computational constraints also played a role in shaping the experimental setup. Iterative pseudo-labeling and domain adaptation procedures can be computationally expensive, particularly when working with large image datasets. The re-initialization and full retraining of the model at each pseudo-labeling iteration, while important for avoiding optimizer state bias, multiplied the computational cost of the self-training procedure.

8.4 Future Work

Several directions can be explored to extend this work in future research.
First, more advanced self-supervised or contrastive representation learning methods could be investigated to improve feature robustness before performing domain adaptation. Pretraining models using domain-agnostic objectives may reduce the magnitude of the domain shift and provide a better starting point for adaptation.

Second, semi-supervised domain adaptation methods could be explored by incorporating a small number of labeled target samples. The results of this work suggest that even a few labeled target samples, particularly for the minority classes, could dramatically improve adaptation performance, given that the main bottleneck appears to be the difficulty of correcting label shift without direct target supervision.
Another promising direction is the use of modern vision architectures such as Vision Transformers or hybrid CNN-transformer models, which have demonstrated strong performance in many medical imaging tasks and have been reported to exhibit increased robustness to domain shift.
Additionally, class-conditional domain adaptation methods such as CDAN or ALDA could be investigated to address the class misalignment issue observed in standard DANN. By conditioning the adversarial objective on class predictions, these methods encourage alignment of class-conditional rather than marginal distributions, which may be particularly beneficial in the presence of label shift.
Finally, future work could explore multi-center datasets with more diverse clinical environments and a larger number of patients per center to better understand how domain adaptation methods scale. In particular, the availability of a larger target dataset with more balanced class distributions would enable a more robust evaluation of adaptation performance on minority classes, which currently represents the main bottleneck of all the methods evaluated in this work.
Overall, the results presented in this work demonstrate that domain adaptation techniques can significantly improve cross-dataset generalization in gastric endoscopic image classification. The comparative analysis across two structurally different tasks, a multi-class scenario with combined covariate and label shift, and a binary scenario with predominantly covariate shift, reveals that the effectiveness of each adaptation strategy is task-dependent, and that no single method universally dominates. This finding has practical implications for the deployment of AI-based tools in clinical endoscopy: adaptation strategies should be selected based on the nature of the distributional mismatch between clinical sites, with pseudo-labeling favored when target-domain class structure differs from the source and adversarial alignment favored when the shift is primarily in the visual appearance of the images.

Chapter 9

Conclusion

This thesis investigated the problem of cross-domain generalization for gastric endoscopic image classification. In particular, the work focused on the challenge of transferring models trained on one dataset to another dataset collected under different acquisition conditions, a common scenario in medical imaging applications. The presence of domain shift between datasets can significantly degrade model performance, making domain adaptation strategies essential for reliable deployment in clinical environments.

To address this problem, several unsupervised domain adaptation approaches were implemented and evaluated. The study considered multiple adaptation strategies, including adversarial feature alignment through Domain-Adversarial Neural Networks (DANN), discrepancy-based alignment through Maximum Classifier Discrepancy (MDD), and iterative pseudo-labeling through a self-training framework. These methods were systematically compared against two reference baselines: a zero-shot setting, where the model trained on the source dataset was directly applied to the target dataset, and an upper bound scenario where training was performed directly on labeled target data.

The experimental analysis was conducted on two related classification tasks derived from gastric endoscopy data. The first task involved the multi-class EGGIM classification problem, while the second considered a simplified binary classification scenario distinguishing normal mucosa from intestinal metaplasia. This dual evaluation allowed a comprehensive analysis of how domain adaptation methods behave under different levels of task complexity.

The results consistently confirmed the presence of a significant domain gap between the source and target datasets. Models trained solely on the source domain exhibited a clear performance drop when applied directly to the target dataset, highlighting the limitations of naive cross-dataset transfer.

Across the evaluated approaches, domain adaptation methods were able to partially bridge this gap and improve generalization performance. However, the effectiveness

of each strategy varied depending on the task. In the multi-class EGGIM classification setting, the self-training approach achieved the strongest results among the domain adaptation methods, demonstrating the value of progressively incorporating high-confidence pseudo-labeled target samples into the training process. In contrast, for the binary classification task, the adversarial alignment strategy implemented through DANN provided the best performance, indicating that feature-level alignment can be particularly effective when the classification boundary is simpler. Feature space analyses using UMAP visualizations further confirmed that domain adaptation techniques reduce the separation between source and target representations. Metrics such as L2 centroid distance and cosine similarity indicated that the learned representations become more aligned after adaptation, although a complete overlap between the domains was not achieved. This suggests that certain dataset-specific characteristics remain difficult to eliminate entirely through representation learning alone.

Overall, the findings of this work demonstrate that domain adaptation techniques can significantly improve cross-dataset robustness in endoscopic image classification. At the same time, the results highlight that no single adaptation strategy universally dominates across all scenarios. Instead, the effectiveness of each method depends on the structure of the classification task and the nature of the domain shift between datasets.

These results emphasize the importance of carefully selecting and evaluating adaptation strategies when developing machine learning models for medical imaging applications, particularly in settings where models must generalize across institutions, devices, or acquisition protocols.

Bibliography

[1] S. Ali et al. «Where do we stand in AI for endoscopic image analysis?» In: *npj Digital Medicine* (2022). URL: ... (cit. on p. 1).

[2] Jaehoon Kim et al. «The Advent of Domain Adaptation into Artificial Intelligence for Gastrointestinal Endoscopy and Medical Imaging». In: *Clinical Endoscopy* (2023) (cit. on pp. 3, 6, 7).

[3] T. Hirasawa et al. «Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images». In: *Gastroenterology* 155.4 (2018), pp. 1066–1074 (cit. on p. 4).

[4] Irene Ligato, Giorgio De Magistris, et al. «Convolutional Neural Network Model for Intestinal Metaplasia Recognition in Gastric Corpus Using Endoscopic Image Patches». In: *Frontiers in Medicine* (2024). URL: `https://pmc.ncbi.nlm.nih.gov/articles/PMC11241412/` (cit. on pp. 4, 5).

[5] Y. Xin, Q. Zhang, X. Liu, B. Li, T. Mao, and X. Li. «Application of artificial intelligence in endoscopic gastrointestinal tumors». In: *Journal of Clinical Medicine* (2024). URL: `https://pmc.ncbi.nlm.nih.gov/articles/PMC10747923/` (cit. on p. 5).

[6] Jovita Relasha Lewis, Sameena Pathan, Preetham Kumar, and Cifha Crecil Dias. «AI in Endoscopic Gastrointestinal Diagnosis: A Systematic Review of Deep Learning and Machine Learning Techniques». In: *IEEE Access* 12 (2024), pp. 163764–163786. DOI: `10.1109/ACCESS.2024.3483432` (cit. on p. 5).

[7] Giovanni Esposito et al. «Endoscopic grading of gastric intestinal metaplasia: the EGGIM score». In: *Endoscopy International Open* (2019) (cit. on pp. 5, 14).

[8] M. L. Martins, R. Delas, E. Almeida, D. Marques, D. Libanio, M. Dinis-Ribeiro, F. Renna, and M. T. Coimbra. «Predicting Endoscopic Grading of Gastric Intestinal Metaplasia using Small Patches». In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and*

Biology Society (EMBC). 2025. DOI: 10.1109/EMBC58623.2025.11253452 (cit. on p. 5).

[9] Changzheng Ma, Peng Zhang, et al. *Prediction of the gastric precancerous risk based on deep learning of multimodal medical images*. Research Square preprint. 2024. URL: https://www.researchsquare.com/article/rs-4747833/v1 (cit. on p. 5).

[10] Jaehyung Yoon et al. «Domain Generalization for Medical Image Analysis: A Review». In: *arXiv preprint arXiv:2310.08598* (2023) (cit. on pp. 6, 15, 19).

[11] Federico Matta et al. «A Systematic Review of Generalization Research in Medical Image Classification». In: *Medical Image Analysis* (2024) (cit. on pp. 6, 7, 15).

[12] Yaroslav Ganin and Victor Lempitsky. «Domain-Adversarial Training of Neural Networks». In: *Journal of Machine Learning Research* (2016) (cit. on pp. 7, 17).

[13] Xinyu Liu and Yixuan Yuan. «A Source-Free Domain Adaptive Polyp Detection Framework With Style Diversification Flow». In: *IEEE Transactions on Medical Imaging* 41.7 (2022), pp. 1897–1908. DOI: 10.1109/TMI.2022.3150435. URL: https://pubmed.ncbi.nlm.nih.gov/35139013/ (cit. on p. 7).

[14] Farhad Alijani et al. «Vision Transformers in Domain Adaptation and Domain Generalization: A Study of Robustness». In: *Pattern Recognition* (2024) (cit. on p. 7).

[15] M. Muto et al. «Endoscopic diagnosis of early gastric cancer and precancerous conditions». In: *Digestive Endoscopy* (2005) (cit. on pp. 9, 14).

[16] M. Dinis-Ribeiro et al. «Management of epithelial precancerous conditions and lesions in the stomach (MAPS II): guideline update». In: *Endoscopy* (2019) (cit. on pp. 11, 14).

[17] L. Zhang et al. «Domain Generalization for Medical Image Analysis: A Survey». In: *arXiv* (2024). URL: ... (cit. on p. 19).

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. «Deep Residual Learning for Image Recognition». In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90 (cit. on p. 31).

[19] Dong-Hyun Lee. «Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks». In: *Workshop on Challenges in Representation Learning, ICML*. 2013 (cit. on p. 34).

[20] Prabhs. *Domain-Adversarial Neural Networks (DANN) Explained*. 2020. URL: https://medium.com/@prabhs./domain-adversarial-neural-networks-dann-explained-f73c9740ff49 (visited on 02/20/2026) (cit. on p. 37).