FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Domain Adaptation in the Classification of Levator Avulsion

Fábio Oliveira Rodrigues

Master in Bioengineering

Supervisor: Prof.Dr. João Pedrosa

Second Supervisor: Prof.Dr. Helena Williams

June 21, 2026

Domain Adaptation in the Classification of Levator Avulsion

Fábio Oliveira Rodrigues

Master in Bioengineering

June 21, 2026

Abstract

Artificial intelligence has evolved rapidly over time, becoming widely applied in the medical field and enabling automated diagnosis of many diseases with high accuracy. However, despite these advances, the deployment of such models in real clinical environments remains challenging due to several problems, such as the presence of heterogeneity in the different clinical settings. Variations in clinical protocols, imaging modalities, acquisition equipment and patient demographics introduce discrepancies between training and target data distributions, which can lead to significant performance degradation when models are deployed in data from unseen domains.

This dissertation addresses the problem of domain shift in medical imaging, focusing on the detection of levator ani muscle avulsion using ultrasound data. Levator ani muscle (LAM) avulsion is a common pelvic floor injury that occurs during vaginal delivery, which can significantly affect women's quality of life. The diagnosis of this injury is associated with the inherent variability of ultrasound imaging, which is strongly dependent on the operator and the equipment used, and the complex diagnosis process can lead to subjectivity depending on the clinician's practice.

To address this challenge, a model capable of classifying LAM avulsion was first developed by testing different architectures, augmentation strategies, and the effect of using a mask. Subsequently, several domain adaptation (DA) techniques were evaluated using data from two different clinical institutions: UZ Leuven and Sydney Urodynamic Centres. The methods include feature alignment approaches that use an explicit discrepancy metric (MMD, CORAL, CMD) or adversarial training (DANN and CDAN). Additionally, image translation methods that operate at the pixel level (histogram matching and Fourier domain adaptation) were employed, as well as pseudo-labeling (FixMatch) and a test-time adaptation method (TENT).

The models demonstrated good performance in classifying LAM avulsion, achieving an AUC of 0.9233 ± 0.0751 (Sydney) and 0.8051 ± 0.1476 (Leuven) using a ConvNeXtV2-atto architecture with masking and augmentation. Furthermore, DA experiments consistently outperformed the lower bound baseline across both clinical centers. On the Sydney dataset, the Global Histogram Matching approach yielded the highest AUC mean gain (0.1117 ± 0.1386). Conversely, for the Leuven dataset, CDAN was the only strategy to reach statistical significance ($p < 0.05$), effectively reducing domain shift with a mean difference of 0.1577 ± 0.0672.

Overall, this work demonstrates that all tested DA methods can improve model performance if well-tuned, effectively reducing domain discrepancies, mitigating the impact of domain shift. However, their efficiency is context-dependent, requiring extensive hyperparameter tuning.

Keywords: Artificial Intelligence, Medical Imaging, Levator Ani Muscle Avulsion, Ultrasound, Domain Adaptation, Domain Shift, Feature Alignment, Image Translation, Pseudo Labeling, Test-time Adaptation

Resumo

A inteligência artificial tem evoluído rapidamente, sendo amplamente aplicada na área médica, permitindo o diagnóstico automatizado de diversas doenças com elevada precisão. No entanto, apesar destes avanços, a sua aplicação em ambientes clínicos apresenta várias dificuldades, como a presença de heterogeneidade nos diferentes contextos clínicos. Variações nos protocolos clínicos, modalidades de imagem, equipamento de aquisição e dados demográficos dos pacientes introduzem discrepâncias entre as distribuições dos dados de treino e de teste, o que pode levar a uma degradação significativa do desempenho quando os modelos são aplicados em dados de outros contextos clínicos que o modelo não teve acesso durante o treino.

Esta dissertação aborda o problema da variação de domínio na área médica, focando-se na deteção da avulsão do músculo elevador do ânus (MEA) por ultrassom. Esta é uma lesão comum do pavimento pélvico que ocorre durante o parto vaginal e que pode afetar significativamente a qualidade de vida das mulheres. O diagnóstico desta lesão está associado à variabilidade inerente da ecografia, que é fortemente dependente do operador e do equipamento utilizado. Para além disto, o processo de diagnóstico pode estar sujeito a subjetividade devido à sua elevada complexidade.

Para resolver este problema, foram treinados modelos capazes de classificar a avulsão do MEA, testando diferentes arquiteturas, aplicação de transformações e uma máscara nos dados. Subsequentemente, várias técnicas de adaptação de domínio (DA) foram avaliadas utilizando dados de duas instituições clínicas diferentes: UZ Leuven e Centro Urodinâmico de Sydney. Os métodos incluem abordagens de alinhamento de características que utilizam uma métrica de discrepância explícita (MMD, CORAL, CMD) ou treino adversarial (DANN e CDAN). Adicionalmente, foram aplicados métodos de harmonização da imagem que operam ao nível do pixel (alinhamento de histograma e adaptação no domínio de Fourier), bem como pseudo-rotulagem (FixMatch) e um método de adaptação que atua durante a inferência (TENT).

Os modelos obtiveram um bom desempenho na classificação da avulsão do MEA, alcançando uma AUC de 0.9233 ± 0.0751 (Sydney) e 0.8051 ± 0.1476 (Leuven) utilizando a arquitetura ConvNeXtV2-atto aplicando uma máscara e transformações nos dados. Além disso, as experiências de DA superaram consistentemente a linha de base em ambos os centros clínicos. Nos dados de Sydney, a abordagem de alinhamento de histograma obteve o maior ganho médio de AUC (0.1117 ± 0.1386). Por outro lado, para os dados de Leuven, o CDAN foi a única estratégia a atingir significância estatística ($p < 0.05$), reduzindo eficazmente a variação de domínio com uma diferença média de 0.1577 ± 0.0672.

Este trabalho demonstra que todos os métodos de DA testados podem melhorar o desempenho dos modelos se bem parametrizados, reduzindo eficazmente as discrepâncias de domínio e mitigando o efeito da variação de domínio. Contudo, a sua eficiência é dependente do contexto, exigindo uma extensa otimização de hiperparâmetros.

Palavras-chave: Inteligência Artificial, Imagiologia Médica, Avulsão do Músculo Elevador do Ânus, Ultrassom, Adaptação de Domínio, Variação de Domínio, Alinhamento de Características, Harmonização de Imagem, Pseudo-rotulagem, Adaptação em Tempo de Inferência

UN Sustainable Development Goals

The United Nations Sustainable Development Goals (SDGs), defined in 2015, provide a comprehensive blueprint for achieving a stable, prosperous and sustainable future. These consist of 17 objectives that aim to mitigate critical global challenges, including, for example, health, education, climate change, social inequality, and environmental degradation. These objectives require a mutual effort from governments, promoting international collaboration to ensure a successful transition period and the achievement of these targets by 2030 [1].

Scientific solutions should always be developed with a sense of responsibility, acknowledging that any technological advancement may affect different social groups as well as the conditions for future generations. Therefore, it is essential to rigorously evaluate potential negative consequences. The SDGs serve as an orienting guide when developing technology, ensuring the long-term sustainability.

This dissertation fits into an era of rapid technological transition, characterized by the exponential growth of scientific knowledge and the integration of artificial intelligence (AI) into diverse solutions. This dissertation implements AI methodologies aimed at addressing challenges in medical imaging diagnosis, ensuring that the proposed solutions are more generalized and adaptable to diverse clinical scenarios. This approach has the potential to enhance the efficiency of healthcare systems while simultaneously increasing the accessibility, reliability and diagnostic accuracy of the clinical diagnostic process on a global scale.

Consequently, this project contributes to a more sustainable future by directly contributing to the improvement of global healthcare and reducing inequalities in the accessibility of health services. Additionally, this work also contributes to scientific innovation, implementing solutions that could also be applied in different sectors. This research is directly aligned with the following Sustainable Development Goals (SDGs):

SDG 3 Ensure healthy lives and promote well-being for all at all ages.

SDG 9 Build resilient infrastructure, promote inclusive and sustainable industrialization, and foster innovation.

SDG	Target	Contribution	Performance Indicators and Metrics
3	3.7	**Ensure universal access to sexual and reproductive health-care services.** This project enhances diagnostic accuracy in the detection of levator ani muscle avulsion, providing clinicians with a robust and accurate tool to support earlier diagnosis and improve the standard of care for pelvic floor injuries.	**Indicator:** 3.7.1 - Proportion of women of reproductive age with access to modern diagnostic and family planning support. **Metrics:** - Performance of the AI models in diagnosing this condition. - Reduction in clinical assessment time per case.
	3.8	**Achieve universal health coverage.** This project studies the impact of implementing domain adaptation methods to adapt the model to different clinical settings. A model trained with high-quality data can be deployed to diagnose the disease in other centers or regions without requiring significant additional resources or data annotation.	**Indicator:** 3.8.1 - Coverage of essential health services **Metrics:** - Number of clinical scenarios covered. - Reduction of the performance degradation when using the model in a different setting.
9	9.5	**Enhance scientific research, upgrade the technological capabilities.** This research fosters scientific innovation by investigating various domain adaptation methods and analyzing their implementation details within a medical context, thereby contributing to the advancement of AI in healthcare.	**Indicator:** 9.5.1 - Research and development expenditure as a proportion of GDP. **Metrics:** - Number of studies conducted to improve the baseline classification of the condition. - Number of successfully implemented domain adaptation techniques.

Acknowledgements

Primeiramente, queria agradecer à minha família que sempre me ajudou em tudo para que eu conseguisse chegar até aqui e me permitiu que tivesse todo o tempo que precisasse para as minhas tarefas académicas com a máxima compreensão. Queria agradecer aos meus amigos, que sempre me apoiaram e contribuiram para a minha felicidade, tornando todo este processo muito mais fácil. E, finalmente, à minha namorada, Maria, que sempre esteve comigo a prestar o seu apoio incondicional e a motivar-me.

Queria agradecer aos meus colegas do laboratório de UZ Leuven que me integraram no grupo e tornaram a minha estadia muito mais confortável. Queria agradecer à Keke Shi que sempre me ajudou em tudo. Queria também agradecer aos meus colegas de laboratório do INESC TEC, com os quais aprendi muito nas reuniões que tive e também sempre me ajudaram nesta caminhada.

Por último, este trabalho não seria possível sem os meus orientadores Dr. João Pedrosa e Dra. Helena Williams, aos quais agradeço por todo o acompanhamento e revisão durante toda esta jornada.

Fábio Oliveira Rodrigues

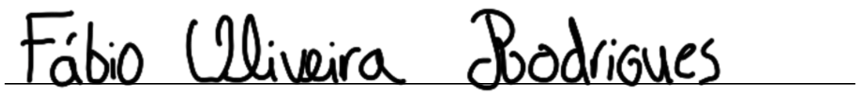
Declaration of honour

I declare that this dissertation is my own work and has not previously been submitted or assessed at this or any other institution.References to other authors (statements, ideas, thoughts), as well as the use of published works, strictly comply with the rules of attribution and are duly identified in the text and in the bibliographic references, in accordance with the applicable referencing standards.I am aware that the practice of plagiarism or self-plagiarism constitutes an academic offence, as established in the U.Porto Code of Ethics.

I further declare that I have included text(s) published in co-authorship, and that I have specified, with transparency and rigor, the contribution of each co-author, as well as stated whether there exists any other dissertation in which the same text(s) may also have been included. The authorization from the journal in which the work is published is also included, together with the express agreement signed by all co-authors authorizing the inclusion of the article.

I also declare that I have used Artificial Intelligence tools to complete part(s) of this dissertation, and that this use and its results are duly noted and have been carefully checked for errors, inaccuracies, unfounded attributions or other forms of bias in content and arguments, as well as determining authorship.

Fábio Oliveira Rodrigues



Declaration of Artificial Intelligence Use

During the writing of this dissertation, AI tools (Grammarly, Gemini, Claude) were used to assist with english grammar, argument structuring and ortography check. The author remains solely responsible for all the content generated, preparation of the data used, as well as the revision of all the manuscript.

"A river cuts through rock, not because of its power, but because of its persistence."

Jim Watkins

Contents

List of Figures

List of Tables

List of Acronyms

AdaBN	Adaptive Batch Normalization
AI	Artificial Intelligence
BMI	Body Mass Index
CDAN	Conditional Adversarial Domain Adaptation
CL	Contrastive Learning
CNN	Convolutional Neural Network
CMD	Central Moment Discrepancy
CORAL	Correlation Alignment for Unsupervised Domain Adaptation
CT	Computed Tomography
CTTA	Continual Test-Time Adaptation
CyCADA	Cycle-Consistent Adversarial Domain Adaptation
DA	Domain Adaptation
DANN	Domain-Adversarial Neural Network
DIF	Domain Invariant Feature
DRL	Disentangled Representation Learning
DSF	Domain Specific Feature
EDM	Explicit Discrepancy Minimization
EMA	Exponential Moving Average
EVUS	Endovaginal Ultrasound
FDA	Fourier Domain Adaptation
HM	Histogram Matching
GAN	Generative Adversarial Network
GRN	Global Response Normalization
IDM	Implicit Discrepancy Minimization
LUG	Levator Urethra Gap
ML	Machine Learning
MMD	Maximum Mean Discrepancy
MRI	Magnetic Resonance Imaging
PL	Pseudo Labeling
RKHS	Reproducing Kernel Hilbert Space
SDA	Supervised Domain Adaptation
SSDA	Semi-Supervised Domain Adaptation
SSL	Semi-Supervised Learning
TPUS	Transperineal Ultrasound
TTA	Test-Time Adaptation
TUI	Tomographic Ultrasound Imaging
UDA	Unsupervised Domain Adaptation
US	Ultrasound
VAE	Variational Autoencoder
ViT	Vision Transformer

Chapter 1

Introduction

1.1 Background and motivation

Throughout the last years, artificial intelligence (AI) has undergone exponential evolution, revolutionizing daily life across diverse fields. In healthcare, AI is widely applied in, for example, treatment selection, interpretation of patient genomes, patient monitoring, drug discovery, aided surgery, and disease diagnosis [2]. This integration of AI into healthcare streamlines clinical workflows by automating routine tasks, such as generating medical reports and organizing documentation, thereby freeing up clinicians' time for more complex patient care [3]. Moreover, AI significantly improves disease diagnosis by increasing both efficiency and diagnostic accuracy [2].

This high diagnostic accuracy has been achieved through continuous scientific progress, which has combined improvements in imaging modalities (e.g., acquisition of images with better resolution and reduced noise) with advances in AI [4]. Advancements in this field are characterized by the iterative refinement of architectures, aiming to overcome existing constraints and set new benchmark performances. These advancements have improved the ability of automated systems to identify subtle patterns in data, which has been successfully translated to the medical field to enhance diagnostic performance [5].

However, despite these advances, the application of AI in healthcare involves several challenges that must be carefully considered. These include, for example, label imbalance, often resulting from the low prevalence of diseases, data privacy concerns, and substantial data heterogeneity [6]. This data heterogeneity arises from multiple factors: different institutions employ varying medical protocols and use different imaging equipment; there are differences in patient demographics; and intra- and inter-operator variations occur in both image acquisition and diagnosis. All these factors contribute to significant variations across medical datasets [7, 8].

These challenges are particularly pronounced in medical ultrasound (US) imaging, a modality where high operator dependency and hardware-specific discrepancies [9] amplify such variability [9]. This is evident in the assessment of levator ani muscle (LAM) avulsion using US imaging, a pelvic floor injury affecting 10–30% of women following a normal vaginal delivery, which significantly impacts their quality of life [10]. The diagnostic process is inherently complex, often

hindered by subjective interpretation [11], which frequently leads to variability in diagnosis.

This inherent variability across diverse clinical environments poses a significant challenge for the deployment of AI models. They often exhibit substantial performance degradation when evaluated on data originating from distributions that differ from their training sets. This distribution shift implies that a model optimized for a specific source domain may fail to maintain its predictive performance upon deployment in a target institution, thereby limiting its clinical viability [7, 8].

For these reasons, it is crucial to ensure that AI models maintain consistent performance across new clinical datasets and populations, with minimal time required for adaptation and without the burden of new clinical labeling. Domain Adaptation (DA) methods [8] can be applied to the models to reduce the discrepancies between the two domains, hence decreasing this performance gap when applied to different settings. In this way, it is possible to use the information present in an already annotated dataset to train a model that leverages that information, while adapting it to achieve robust performance when deployed on a different dataset.

1.2 Objectives

Taking into account the previous challenges identified, this dissertation aims to investigate and optimize DA techniques tailored to the specific context of LAM assessment across multiple centers. The research is mainly structured around two primary objectives: to develop a model for the automated classification of LAM avulsion and analyse its performance decay when deployed in a different clinical setting; and then to implement and evaluate DA methods that mitigate this performance gap. The specific goals are outlined as follows:

- **Develop a classification model for the automated detection of LAM avulsion that achieves high diagnostic performance.** Study the performance across different model architectures, data augmentation strategies, and segmentation mask integration. Quantify the performance decay observed when the model is deployed in a clinical setting with different patient demographics and image acquisition equipment.

- **Implementation of different DA methods to reduce the performance degradation.** Evaluate the impact of applying DA methods across different methodological families, and analyse the sensitivity of model performance to variations in their respective hyperparameters. Identify and evaluate the most effective methods in reducing the domain discrepancies for this specific clinical scenario.

1.3 Dissertation Structure

Following this introductory chapter, the document is organized into the following chapters:

- **Chapter 2 - Clinical background:** Explains the anatomical and physiological implications of LAM avulsion, the diagnostic imaging modalities involved, and the inherent variability in its clinical assessment.
- **Chapter 3 - State of the Art:** Reviews the temporal evolution of image classification methods, transitioning into a detailed analysis of DA, exploring its taxonomy and the methodologies from the different families typically employed.
- **Chapter 4 - Methodology:** Describes the methodology implemented in this work, including the experiments conducted to obtain a model able to classify LAM avulsion and the implementation details of several DA methods used to mitigate the performance decay.
- **Chapter 5 - Results and Discussion:** Presents the results obtained for the baseline training, showing the results for each of the experiments conducted, as well as the results of the implemented DA methods, demonstrating the impact of each method's parameters.
- **Chapter 6 - Conclusion:** Summarizes the main findings of this dissertation, discusses the limitations of the current work, and proposes potential directions for future research.

Chapter 2

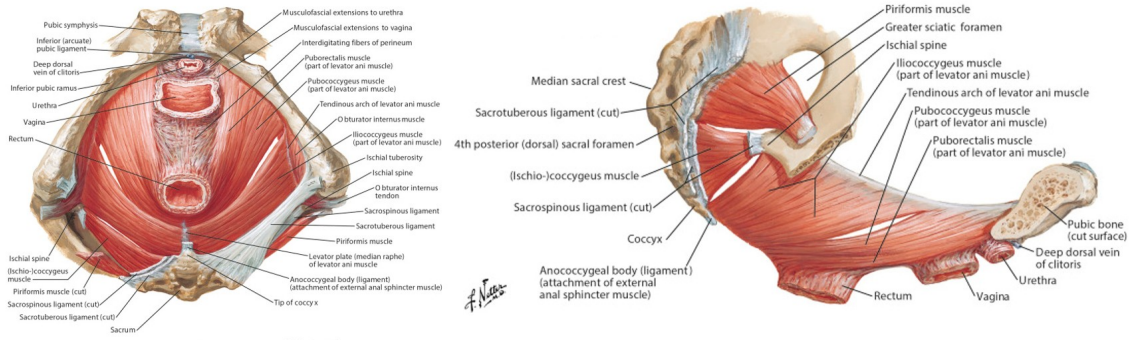
Clinical Background

Pelvic Floor Disorders affect millions of women worldwide [12] and include pelvic organ prolapse (drop from their position), urinary and fecal incontinence, and sexual dysfunction. LAM avulsion is a pelvic floor injury and is strongly associated with vaginal delivery, with a considerable impact on women's quality of life [13]. There are different imaging modalities that can be used to visually diagnose this condition. This chapter will describe pelvic floor anatomy and the pathophysiology of LAM avulsion, as well as the common imaging modalities that can be used for the diagnosis of this injury. It also discusses the inter-operator variability of pelvic floor injury assessment, which is strongly dependent on the clinician, image quality, and clinical practices [14].

2.1 Pelvic Floor Anatomy and Pathophysiology of Levator Ani Muscle Avulsion

The pelvic floor is composed of various anatomical components, including ligaments, muscles, fascia (connective tissue), and the pelvic organs (bladder, uterus, vagina, and rectum). These muscles are essential for supporting the pelvic organs and maintaining normal pelvic function, particularly through their role in bladder and bowel control [15].

Following DeLancey's anatomical and functional model [16], pelvic floor support is organized into a hierarchical system of suspension and fixation. Level I involves the upper third of the vagina and the cervix, where suspensory ligaments anchor these structures to the pelvic walls. Level II covers the mid-vagina, which is attached laterally to fascial structures. Level III comprises the lower vagina, fused with surrounding tissues such as the perineal body and, critically, the LAM, making it a level often directly implicated in childbirth-related injury. Figure 2.1 provides a more detailed illustration of this anatomy [15].



a) b)

Figure 2.1: Anatomy of the pelvic floor muscles: a) inferior view of the pelvic diaphragm; b) lateral view. Replicated from [15].

The LAM, composed of the pubococcygeus, puborectalis, and iliococcygeus muscles, forms a muscular sling that maintains pelvic organ support, continence, and structural integrity of the pelvic floor. During voiding and defecation, relaxation of the LAM allows the passage of urine and stool [17]. During vaginal delivery, the LAM is subjected to extreme biomechanical stress (mainly in the posterior-medial portion), achieving high circumferential stretch ratio, as the fetal head descends through the birth canal [18]. This can lead to detachment of the LAM from its insertion on the pubic bone, a condition known as LAM avulsion [19].

LAM avulsion can compromise a woman's health and overall quality of life, often leading to complex long-term physical and psychological problems. [20]. This condition is associated with pelvic organ prolapse [21], frequently manifesting as hiatal ballooning which is the enlargement of the levator hiatus area that facilitates the downward displacement of pelvic organs through the vaginal canal. Additionally, LAM compromises the structural integrity of the pelvic floor, affecting functional pelvic support that is related with an increased risk of urinary and fecal incontinence [22]. These complications can substantially affect patients' daily lives, with negative consequences for their social, physical, and sexual well-being [23, 24].

2.2 Medical Imaging Modalities for Assessment

To diagnose LAM avulsion and to assess pelvic floor integrity and function, clinicians rely on physical examination and medical imaging. One of the simplest methods involves vaginal palpation, which has moderate validity and reproducibility but can be difficult to perform, requiring extensive clinical training and being uncomfortable for the patient. With the advancement of medical imaging, the diagnosis of LAM avulsion became more accurate [13].

Magnetic resonance imaging (MRI) was the first imaging modality used for the assessment of LAM avulsion, due to its high spatial resolution and excellent anatomical detail, and is also widely used to diagnose other conditions. However, MRI is an expensive modality, associated with high equipment costs and longer acquisition times, extracting static images. For these reasons, US has

become the gold standard for diagnosing LAM avulsion, offering a cost-effective, accessible, and an easier modality that enables real-time dynamic imaging [13].

In US, LAM avulsion can be evaluated using either 3D transperineal ultrasound (TPUS) or 3D endovaginal ultrasound (EVUS). EVUS enables the extraction of detailed structural images with higher spatial resolution, since high frequency probes can be placed nearby the target tissue, allowing for a more precise evaluation even when the pelvic floor muscles are in a relaxed state. However, this method is more invasive than TPUS, since the probe needs to be placed inside the vaginal canal and is predominantly employed in static studies, as the probe can limit the natural muscle contraction. TPUS has become the gold standard for the assessment of LAM avulsion, due to its low cost, highly accessible, easy to use, allowing dynamic analysis (4D TPUS that is 3D volumes acquired over time) and for being non-invasive (probe placed externally on the perineum). Despite being of lower imaging resolution, there is a high diagnosis agreement of LAM avulsion using MRI, EVUS and TPUS volumes [25, 26]. Figure 2.2 illustrates the two US techniques.

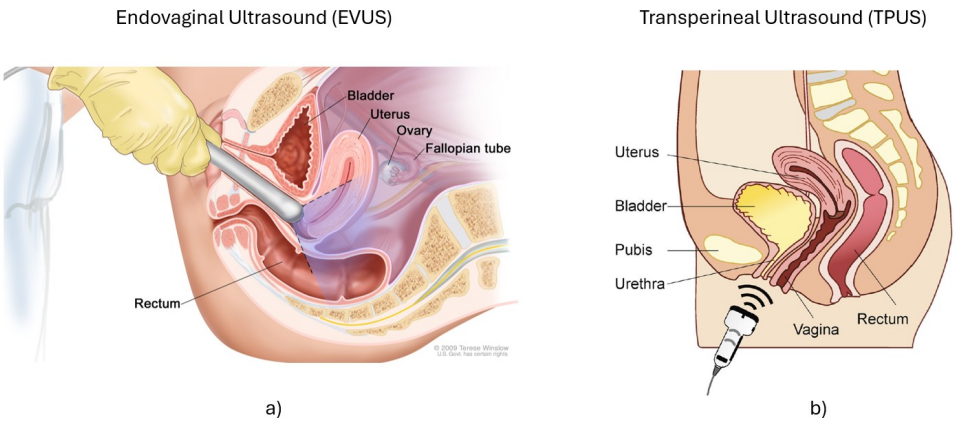
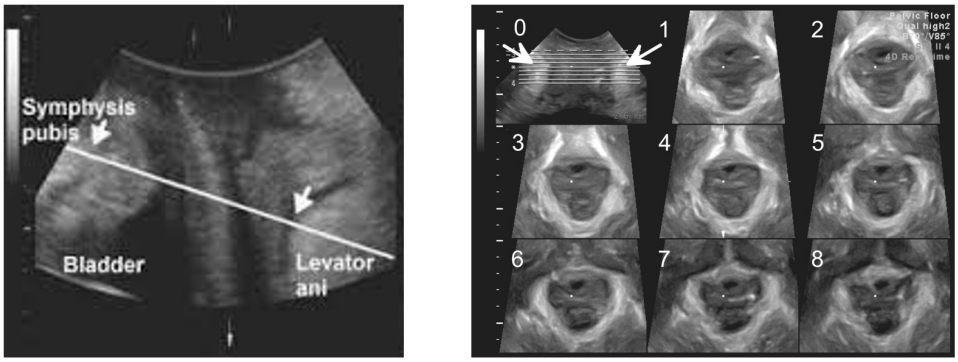


Figure 2.2: Schematic representation of pelvic floor US modalities: a) EVUS. Replicated from [27]); b) TPUS. Replicated from [28]

The development of these imaging modalities has provided a more robust and efficient workflow for LAM assessment. Clinicians frequently assess LAM avulsion by identifying asymmetry in the levator hiatus and discontinuities at the site of muscle insertion, which can lead to greater variability in the diagnosis. Therefore, a more precise and reproducible way to evaluate LAM avulsion is important to define [29].

Dietz et al. [19] introduced a key protocol for diagnosing LAM avulsion using TPUS imaging. After extracting the volumes, the plane of minimal hiatal dimensions (i.e., shortest distance between LAM and symphysis pubis) is located, as represented in Figure 2.3 a. From this plane, Tomographic Ultrasound Imaging (TUI) (Figure 2.3 b) extracts equidistant parallel slices with the plane in the centre (intervals of 2.5 mm, from 5 mm caudal to 12.5 mm cranial), from which the slices at the reference plane, 2.5 mm, and 5 mm above it are the most important for evaluation. In these slices, clinicians measure the levator urethra gap (LUG) that is defined as the minimum gap between the urethral center and the pelvic sidewall insertion of the most medial component of the LAM. A LUG above 2.5 cm means the presence of LAM avulsion.



a) b)

Figure 2.3: Image acquisition and processing for LAM avulsion assessment: a) Representation of the extraction of the minimal dimension hiatus plane. Adapted from [30]; b) Grid of US slices obtained via TUI. Replicated from [31].

LAM avulsion can be further divided taking into account the severity and the side of the deformation. If all the three central slices present a LUG above the threshold, the injury is classified as complete avulsion. Conversely, an increased LUG is detected in only one or two of these three central slices constitutes a partial avulsion. In terms of the side of the injury, LAM avulsion can be considered bilateral (both sides) or unilateral (left or right). The clinical relevance of partial avulsions remains uncertain, as research indicates that they show limited association with pelvic floor dysfunction and that their diagnosis may be susceptible to image acquisition artifacts [32]. Figure 2.4 illustrates LUG measurements in a normal (a) and a complete avulsion case (b).

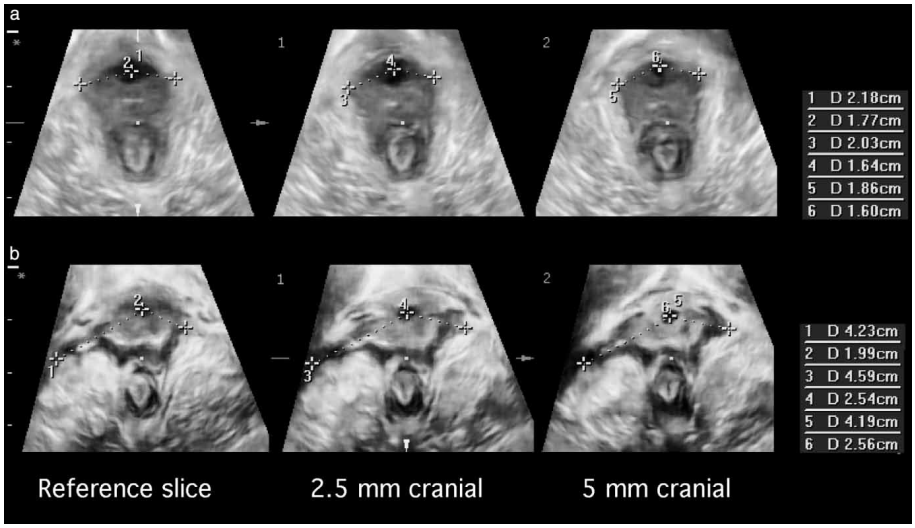


Figure 2.4: Measurement of the LUG on TUI. The images show the three central TUI slices used for determination of the LUG in a normal patient (a) and in a patient with unilateral avulsion (b). The patient's right side is shown on the left in all images, as if the pelvic floor was seen from below. Two of the left-hand measurements (25.4 mm and 25.6 mm for center and right images, respectively) are borderline abnormal. Replicated from [32].

2.3 Variability Across Clinical Settings

Despite the existence of a standardized protocol for the assessment of LAM avulsion, it remains subject to considerable variability across clinical settings. This is largely attributed to operator-dependent factors, particularly the difficulty in identifying the pubovisceral muscle insertion to measure the LUG, which often leads to limited inter and intra observer reliability. Additionally, many clinicians also assess LAM avulsion by analysing the asymmetry of the levator, depending on the measurement of the LUG only in dubious cases, leading to variability. In borderline cases, partial avulsion may be misclassified as an intact case due to differences in anatomical interpretation [11]. Image acquisition also plays a crucial role, as variations in the type of probe and US scanner used, in the pressure applied to the skin, and the contraction state of the muscle during the acquisition process can lead to relevant differences in the appearance of the levator hiatus, as well as impacting the overall quality of the image (e.g., contrast, field of view, resolution, noise) [9].

Consequently, the combination of low-quality images, inter-hospital variability, and the technical challenge of accurately localizing the muscle insertion leads to a complex diagnosis process, increasing the risk of misdiagnosis. Additionally, this manual assessment is a time-consuming process that places a heavy burden on clinical workflows. Therefore, given the recent advancements in AI, automating the imaging process is a logical step, allowing for an optimization of the diagnostic workflow, as well as facilitating a process that is severely affected by this inherent variability.

However, this variability remains a major challenge for the development of robust AI solutions, as models often suffer from performance degradation when faced with data from different settings. Simply automating the assessment of LAM avulsion is insufficient if models remain fragile to variations in acquisition protocols and hardware. To ensure reliable LAM avulsion classification across clinical environments, leveraging DA techniques that are able to guide the training process is essential, enabling the model to maintain robust performance when deployed in different clinical sites. For these reasons, the following chapter presents a bibliographic review of DA, with particular focus on its application to medical image analysis that can be used to improve the performance of AI models when exposed to diverse clinical settings.

Chapter 3

Bibliographic Review

This chapter presents the current state of the art in DA for medical imaging. Firstly, the temporal evolution of AI in classification tasks is described, specifically within medical imaging, introducing the different networks developed along with their advantages and limitations. Subsequently, the need and growth of DA methods are discussed, followed by a taxonomy of the problem. An analysis of different method types and their underlying principles is included, with a detailed focus on unsupervised methods, which are divided into several subcategories. Finally, a summary of the chapter is provided, reflecting on future prospects and important current areas of research.

3.1 Evolution of Classification Methods in Image Analysis

The need for automated medical image analysis has long been recognized in the biomedical field, with various strategies developed to improve the accuracy and precision of diagnostic methods. Within medical image analysis, image classification serves as the fundamental task of assigning a specific clinical label (e.g., pathological/healthy) to an entire image based on its visual content. Traditional methods, for instance, relied on morphological operations, binarization, watershed, and region growing to isolate objects of interest and, subsequently, through heuristic rules and manual feature design, the final classification was performed. However, these techniques were highly dependent on human intervention and lacked the capacity to capture complex patterns [33].

The introduction of machine learning (ML) had a major impact on the area of image analysis by enabling algorithms to learn from data rather than relying on explicitly defined rules, allowing a better adaptation to a specific problem. The models use manually extracted features as inputs (e.g., texture, shape, intensity, etc.), which allow discrimination between different classes [34]. Supervised learning methods [35], such as Support Vector Machines, Random Forests, Naive Bayes classifiers, and k-Nearest Neighbors, aimed to find optimal decision boundaries between classes using labeled data. On the other hand, unsupervised learning techniques [36], including k-means clustering, Gaussian Mixture Models, and hierarchical clustering, were used to group similar images in classes without requiring annotated data, which came with a trade-off of diagnosis quality [37]. Despite the advances and the vast application in different areas of ML, these techniques still

had problems, such as the dependence of the models in the process of extracting good and representative features and the impossibility to capture very complex patterns and detect subtle data variations inherent to medical imaging [38].

With the rise of ML, a subfield known as Deep Learning emerged, which led to the ability for automated performance on complex tasks. In the area of image analysis, convolutional neural networks (CNNs) [39] became the most used architectures, as they allowed for high-level feature extraction of images. This network was based on convolutional operations that use adaptable filters as shown in Figure 3.1. These filters capture important local features, such as edges and textures, and using convolutional operations it is possible to aggregate them into more complex features that are relevant for the classification of the images [40]. However, to perform complex tasks like medical image analysis, the size of the network increases substantially, which significantly increases the risk of overfitting and often leads to decreased performance on new data [41].

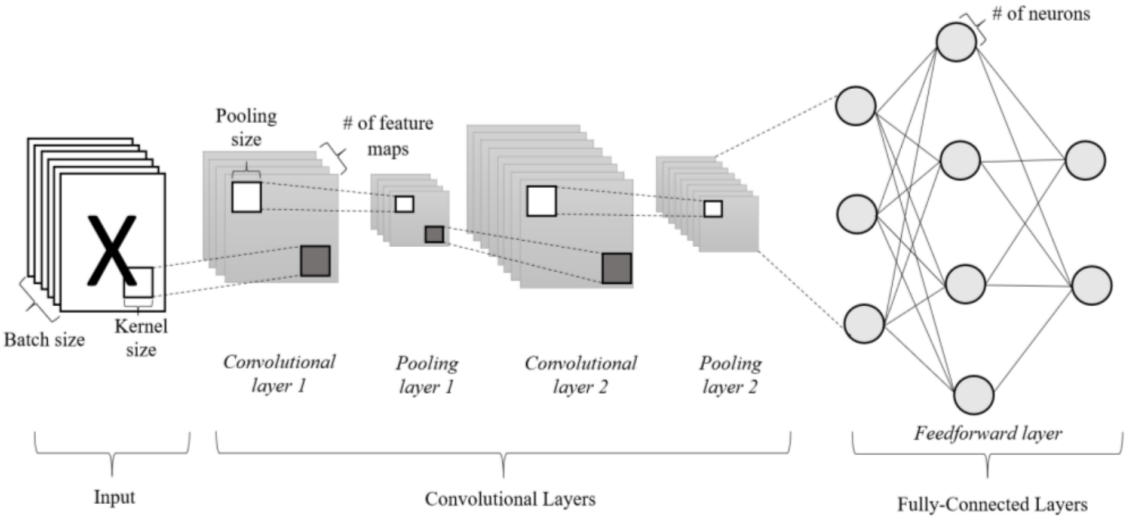


Figure 3.1: Typical architecture of a CNN. The input passes through convolutional and pooling layers to extract relevant features, producing a final label after the fully connected layer. Replicated from [42].

ResNet (2015) [43] was a major breakthrough in image analysis, surpassing almost every state-of-the-art result at the time. This network aims to solve the problem that exists in just scaling up the depth of the network, causing a degradation problem in which accuracy just gets saturated and starts decreasing and also the famous vanishing/exploding gradient problem, where gradients become extremely small or excessively large during backpropagation. It employs residual connections that help learn the necessary modifications to the input to produce accurate predictions. These connections allow an easier flow of the gradient through the network during backpropagation, enabling the training of much deeper networks and improving overall performance.

Despite the advances introduced by ResNets, there were more efficient means of learning than relying on shortcut connections to mitigate gradient degradation. DenseNet [44] addressed this limitation by implementing a densely connected architecture, in which each layer directly receives the feature maps extracted from all previous layers using feature concatenation as it is

possible to verify in Figure 3.2. This strategy facilitates an enhanced flow of information and gradient propagation throughout the network, and promotes feature reuse, substantially reducing the amount of redundant information, and simultaneously decreasing the number of parameters used.

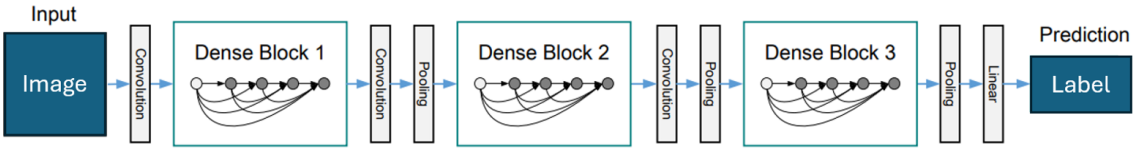


Figure 3.2: Overview of a DenseNet with three dense blocks, illustrating the feature reuse mechanism across layers within each block. The layers between two adjacent blocks are referred to as transition layers and change feature-map sizes via convolution and pooling. Adapted from [44].

EfficientNet was proposed by Tan & Le (2019) [45] as a family of CNNs that achieve state-of-the-art performance while being smaller and faster in inference. Before EfficientNet, scaling up CNNs to improve accuracy, was typically done by trial and error, increasing network depth, width, or input resolution, often leading to inefficient models with a large number of parameters, which comes at a high computational cost. EfficientNet addresses this limitation through its compound scaling method, which systematically balances these three dimensions - depth, width, and resolution - using carefully determined coefficients derived from a grid search on a baseline network (EfficientNet-B0).

Vision Transformers (ViTs) [46] were presented in the literature in 2020 after the massive success of Transformers in Natural Language Processing [47]. These architectures are based on attention mechanisms, which provide the model with a better long-term/global understanding of the image features. Unlike traditional CNNs, ViTs do not rely on local receptive fields but learn long-range dependencies across the entire image, making them particularly effective for capturing global context as presented in Figure 3.3. However, they generally require large amounts of training data or pretraining to achieve competitive performance compared to CNNs.

In subsequent years, Liu et al. (2022) [48] introduced the ConvNeXt, which is inspired by the ViT architecture. This network replaces Batch Normalization with Layer Normalization to mitigate potential performance degradation [49], incorporates a smoother activation function Gaussian Error Linear Unit instead of Rectified Linear Unit and bigger kernel sizes. In the next year, ConvNeXtV2 [50] was developed to even achieve better results, as shown in Figure 3.4 incorporating masked autoencoders [51] and Global Response Normalization (GRN) that enhances the network's ability to capture global context. In this way, it was possible to develop an architecture that combines the simplicity of the traditional CNNs with the ability to have a more global understanding of ViTs.

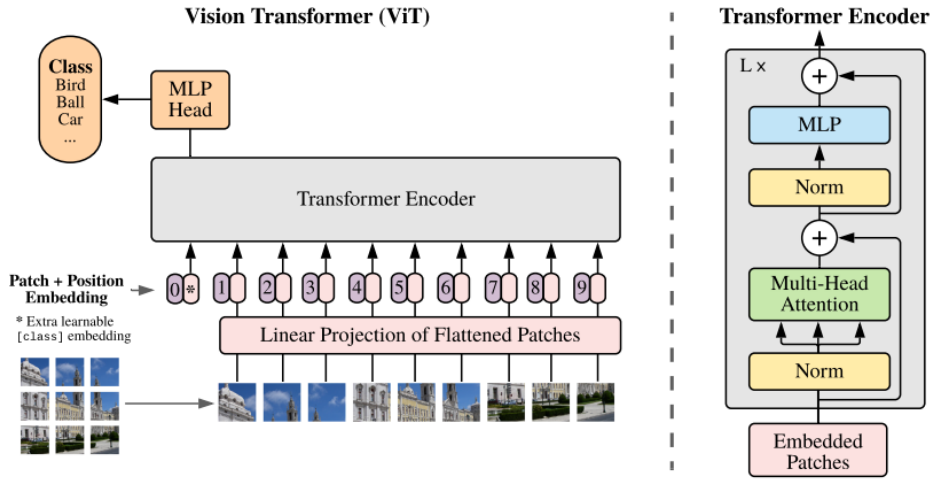


Figure 3.3: Schematic representation of the ViT. The images are divided into small patches associated with a position embedding so the spatial information is not lost and then are passed through a transformer encoder that allows the extraction of relevant features taking into account the relations between patches which are then fed through the decoder to solve the target task. Replicated from [46].

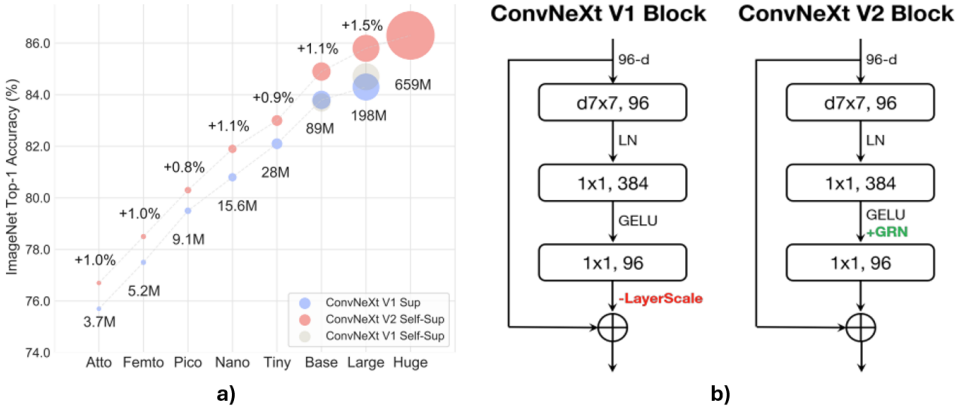


Figure 3.4: Comparison between ConvNeXt V1 and V2 a) ImageNet top-1 accuracy of ConvNeXt and ConvNeXtv2. The ConvNeXtV2 model performs significantly better than the previous version across a wide range of model sizes. b) Differences in the building block of ConvNeXt V1 and ConvNeXt V2. It is possible to see the inclusion of GRN leading to the removal of LayerScale as it became redundant. Adapted from [50].

Despite these advancements, several challenges remain to be addressed in the field of medical image analysis. These include, for instance, the typical class imbalance caused by the low prevalence of diseases [52], data privacy concerns [53] and data heterogeneity [7] across settings.

Data heterogeneity occurs because different hospitals employ varying acquisition protocols, utilize distinct equipment, different image modalities and patient populations have unique demographic distributions that impact disease prevalence and anatomical details. Consequently, due to this data heterogeneity, models trained on a specific dataset often show a significant drop in performance when applied to different settings [7]. For instance, a model trained to classify a disease

in CT (computed tomography) data will fail to effectively extract relevant features when applied to MRI data. This decrease in performance would require the training of new models for each clinical setting, which would be unfeasible, not only due to the time-consuming data collection and annotation requiring specialized clinical expertise but also because of the extensive regulatory demands. Therefore, there is a substantial advantage in developing methods that allow a model to be easily adapted to a new dataset, maintaining good performance [54].

3.2 Domain Adaptation

As previously discussed, medical image analysis is inherently characterized by high heterogeneity, arising from factors such as variations in acquisition protocols, imaging equipment, and patient demographics. This variability leads to what is commonly referred to as domain shift, a phenomenon where the characteristics of the training environment (source domain) differ from those of the deployment environment (target domain). These differences lead to difficulties in extracting relevant features for the intended task, resulting in a lack of robustness and a decrease in model performance [7].

To address this issue, many DA methods have been developed in recent years, aiming to improve model performance in a target domain by leveraging knowledge from a labeled source domain. This differs from Domain Generalization, where the primary objective is to learn domain-agnostic representations that ensure robustness across unseen domains, without requiring access to target domain data during the training phase [7]. Figure 3.5 illustrates an exponential growth in the number of articles addressing DA in medical image analysis.

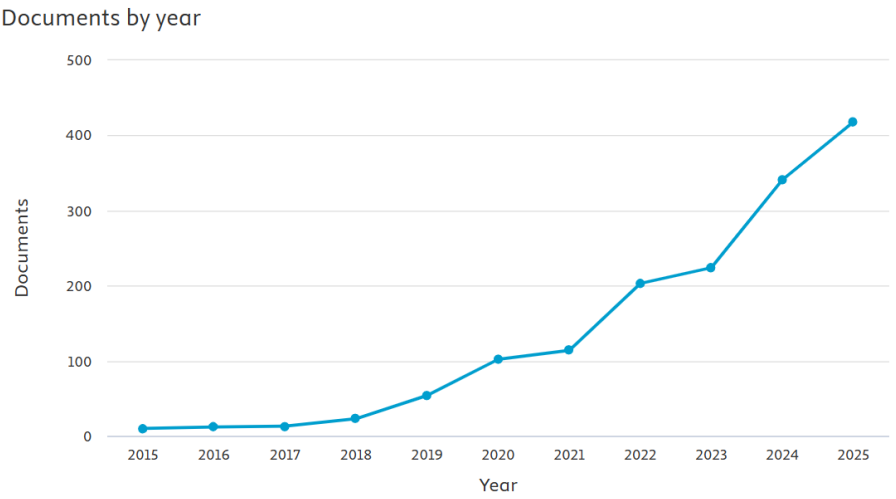


Figure 3.5: Number of articles published on domain adaptation in medical imaging according to Scopus [55] between 2015 and 2025 using as key words domain adaptation and medical image, illustrating the growth of research in this area over time.

3.2.1 Taxonomy of Domain Adaptation Problem

As discussed in the previous section, ML models frequently encounter a significant degradation in performance when applied to data that differs statistically from the training set. To mitigate this domain shift, DA techniques can be employed to adjust the models, maintaining the performance across different settings and bridging the gap between distinct data domains.

Formally, the domain shift problem can be described as the scenario where there is a source domain $S = \{(x_i^S, y_i^S)\}_{i=1}^{n_s}$, where x_i^S represents source samples and y_i^S their corresponding labels, and a target domain $T = \{(x_j^T, y_j^T)\}_{j=1}^{n_t}$. Both domains are defined over the same input and output spaces but follow different data distributions, such that $P_S(X,Y) \neq P_T(X,Y)$. The objective of DA methods is to reduce the discrepancy in model performance when tested on the target domain T, by adapting the model using a subset of n samples from the target domain during training, where $D'_T \subseteq D_T = \{(x_j^T, y_j^T)\}_{j=1}^{n_t}$ [56].

The literature presents several ways to categorize the domain shift problem, often leading to different taxonomies. However, well-established surveys focusing on medical image analysis [7, 8, 56] typically structure the problem around a main set of categories. Figure 3.6 presents this taxonomy, classifying the problem according to the availability of target domain data, the settings of the source and target domains, the nature of the domain shift, and the specific adaptation method used.

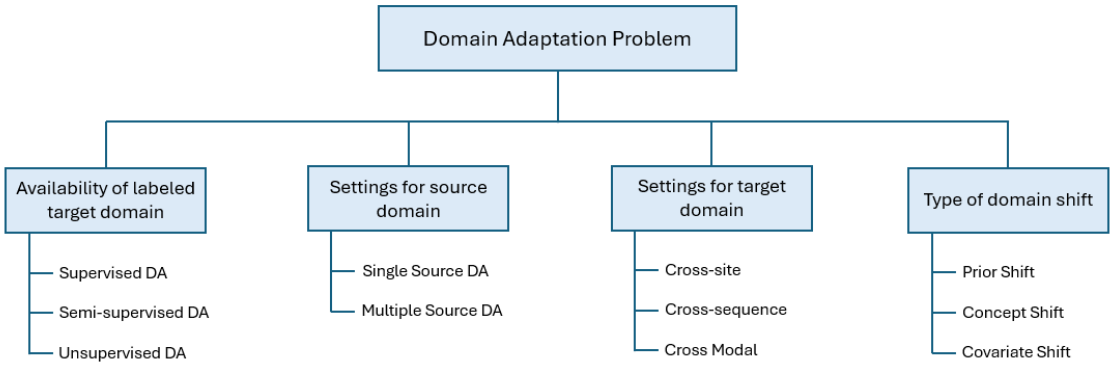


Figure 3.6: Taxonomy of the DA problem, categorized by the availability of labeled data in the target domain, the settings for the source and target domains, and the type of domain shift.

According to the availability of target domain labels, the techniques are referred to as supervised domain adaptation (SDA) when the labels of the full target domain, $D'_T = \{(x_j^T, y_j^T)\}_{j=1}^{n_t}$, are used together with the target data to correct the distributional shift, therefore making the adaptation task easier. If the target labels have low quality, are redundant or only a subset of the target domain is used, the methods are classified as semi-supervised domain adaptation (SSDA). In unsupervised domain adaptation (UDA), the models do not have access to the labels of the target domain images, $D'_T = \{(x_j^T)\}_{j=1}^{n'_t}$, which represents the most challenging scenario for DA [8].

In terms of the settings for the source domain, ML algorithms can have access to a single source domain ($M = 1$) during training, which makes the learning of invariant features that can generalize to a different target domain more difficult due to lower data diversity [57], or have

access to multiple source domains ($M > 1$), allowing a better understanding of which features remain constant across domains, thus facilitating the task of DA [58].

Regarding the target domain, the literature identifies three different settings. Cross-site DA refers to the scenario where datasets are collected from different institutions, which may have, for example, variations in the equipment used [59]. Cross-sequence DA can be further divided into cross-temporal [60], when data is obtained at different phases of treatment, and cross-protocol [61], where differences arise due to variations in image acquisition protocols (e.g., different parameters or T1 vs. T2 MRI images). Finally, when DA focuses on creating models that can easily adapt across different image modalities, such as when the source domain consists of CT images and the target domain is MRI images, it is referred to as cross-modal DA [62].

Irrespective of the type of domain shifts, three different situations can be defined. When a certain disease has a substantially different incidence rate in the source domain ($P_S(Y)$) compared to the target domain ($P_T(Y)$), this represents a case of prior shift [63], also known as a class imbalance problem, where $P_S(Y) \neq P_T(Y)$, but the conditional probabilities remain constant, $P_S(X \mid Y) = P_T(X \mid Y)$. If the relationship between the data and the output differs across domains, there is a concept shift [64], where $P_S(Y \mid X) \neq P_T(Y \mid X)$, which normally occurs in cross-modality settings, for example, when features in CT images have a different meaning than in MRI images. Finally, when the meaning of the features remains constant across domains, there is a covariate shift [65], where $P_S(X) \neq P_T(X)$, normally caused by variations in the machines and environment, but $P_S(Y \mid X) = P_T(Y \mid X)$. The different types of shifts are represented graphically in Figure 3.7.

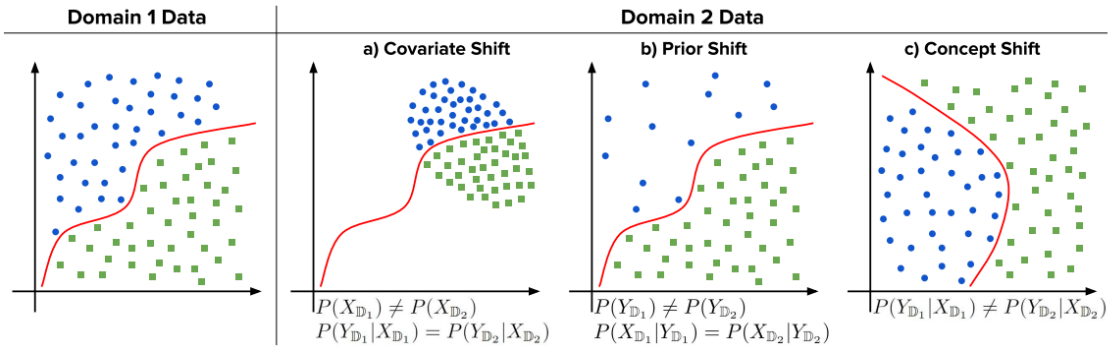


Figure 3.7: Illustration of the three kinds of domain shifts in which Domain 1 represents the source domain and Domain 2 represents the target domain. While Covariate Shift (a) and Prior Shift (b) maintain the original decision boundary despite changes in data distribution or class balance, with Concept Shift (c) there is a change in the classification rule. Replicated from [66].

While the classical taxonomy in Figure 3.6 covers the main aspects of domain shift, other research directions have emerged to address distinct deployment constraints, such as differences in the label space of the source and target domains. Partial DA corresponds to the scenario where the target label space is a subset of the source label space, existing source labels that should not be considered since they do not exist in the target domain. On the other hand, Open Set DA deals with the existence of unknown classes in the target domain. Ultimately, Universal DA has emerged

to deal with all these differences in labels without any prior knowledge of the category overlap, requiring the model to classify known classes correctly while rejecting samples from unknown categories, which represents one of the most useful scenarios, identifying new pathologies in the clinical setting [67].

Another important area of research is Continual or Lifelong DA, which consists of the ability of the models to accumulate knowledge from various domains across time without degrading past performance, ensuring a dynamic learning environment, preserving past information. This is particularly important in clinical practice because data is always changing due to demographic shifts, development of more advanced equipment, variations in disease prevalence, and other factors. The primary obstacle in this setting is catastrophic forgetting, where the model, during the adaptation process, starts losing all the previous knowledge, highlighting the need for more effective strategies than those currently available [68].

Furthermore, focusing on the strict privacy requirements inherent in clinical settings, Source-Free DA has emerged as a critical research area, operating without access to source data, relying solely on pre-trained models, and adapting them using unlabeled target data [69].

3.2.2 State of the Art of Methods

This section provides a comprehensive review of several well-known methods of DA in the literature that can be applied across various domains, including medical image analysis. To provide a structured and cohesive overview, the methods are divided based on target domain availability. This allows a clear division of the methods, specifying in which scenarios they can be applied, according to the label availability. This taxonomy is also motivated by the fact that other classification criteria, for example those based on specific target domain settings, often result in a significant overlap of techniques. Consequently, the following discussion is divided into SDA and SSDA, where a subset of target labeled data is accessible, and UDA, which addresses the scenario in which there is no access to target domain labeled data, representing a more challenging problem.

3.2.2.1 Supervised and Semi-Supervised Methods

In certain scenarios, it is possible that the target domain contains a set of annotated data, which facilitates the adaptation of a model initially trained on the larger source domain. SDA leverages these labeled target samples to improve the model's performance on the target domain, typically by fine-tuning the previous model with the new samples or jointly optimizing on both source and target labeled data, and reducing the discrepancy between the domains.

Fine-tuning is the most typical approach in SDA and consists of using the pre-trained weights of a model trained on a similar dataset and adjusting them using the samples from the new target domain. Depending on the differences between the source and target domains, different fine-tuning strategies can be implemented, taking into account that the initial layers of a CNN generally capture generic low-level features, while deeper layers learn more task-specific representations

[70]. In this way, depending on the dataset similarity, strategies can involve freezing all layers and substituting the original fully connected layers (classification head, for instance) with a new trainable layer, fine-tune only the top layers or even gradually unfreezing all the layers for a more stable adaptation [71].

Supervised Discrepancy Minimization is a technique of SDA that aims to reduce the discrepancy between the domains, taking into account the labels of the data, whose objective is not only to align the representations of the features across domains, but also to align them in terms of the label. This means that samples with the same label from different domains should be similar in the feature space, allowing a good performance in the main task. This can be achieved by minimizing discrepancy metrics between samples with the same label, and adding that factor in the global loss function, as in the work by Chacon et al. (2018) [72] where a two-stream U-Net architecture was employed for DA. In this work, one stream processes source samples while the other handles target samples, utilizing correlation alignment and Maximum Mean Discrepancy (MMD) as regularization terms to align the two domains. Other works also explore the use of a triplet loss function to further enforce that samples with the same label across domains are closer in the feature space, while samples with different labels remain separated [73].

In terms of SSDA methods, they can be divided into three main categories. Cases where there is only access to a portion of the labels of the target domain are referred to as incomplete supervision. This is the most common type, and the techniques previously described in supervised learning can be used here, although this approach must account for the fact that the amount of data of the target domain is often significantly lower in comparison to the source domain. Other techniques such as pseudo-labeling (PL) [74] or consistency regularization [75] can also be used. Inexact supervision happens when there is only access to coarse-grained labels, meaning the labels are more generic and unspecific (e.g., 'animal' instead of 'dog'), and therefore, the intended task is more difficult to perform. Methods commonly rely on Multiple Instance Learning [76] to infer the precise details from the coarse labels. The last type is inaccurate supervision, which represents one of the most difficult scenarios since the labels can be erroneous, leading to noise propagation and hindering model's convergence. The methods use robust loss functions [77] to minimize the influence of erroneous labels and apply label correction mechanisms to fix the noisy annotations [56].

3.2.2.2 Unsupervised Methods

The acquisition of high-quality labeled data can be extremely expensive and time-consuming, often requiring the involvement of skilled professionals. This challenge is particularly pronounced in medical image analysis, where obtaining high-quality labels depends on well-organized hospital or clinical systems, and the participation of health experts can be difficult to coordinate. To address this issue, many techniques have been developed in recent years to improve DA without relying on target domain labels. The following subsections describe the main families of UDA methods and their underlying principles. In Figure 3.8, an overview of the different UDA methods is presented.

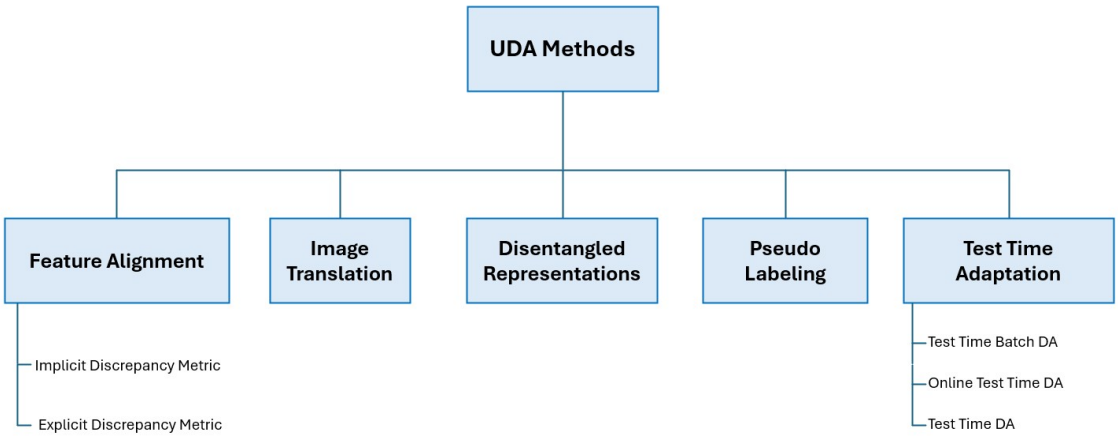


Figure 3.8: Taxonomy of UDA methods, categorized by their underlying alignment and adaptation strategies.

Feature Alignment

In UDA, feature alignment methods aim to reduce the domain shift between the source and target domain by aligning their representation in the feature domain, encouraging the model to extract domain-invariant features, learning consistent and transferable representations across domains. These techniques can be divided into two subcategories: explicit discrepancy minimization (EDM) and implicit discrepancy minimization (IDM).

EDM methods define a metric that quantifies the discrepancy between features extracted from the source and target domains, aiming to minimize this metric, encouraging the model to learn domain-invariant features, which improve transferability and ensure that representations remain constant across domains. Such discrepancy metrics can be based on statistical measures, including distances between distributions, feature correlations, or higher-order moments, which are explicitly optimized during training.

One of the fundamental works in this area was a kernel based approach that aimed to compare the distribution in a non-parametric way. In the work of Long et al. (2015) [78], the Deep Adaptation Networks were introduced, exploring how the minimization of the MMD [79] can be used to improve the DA of the models. MMD enables the comparison of different distributions, making it possible to determine whether two samples were drawn from different distributions by mapping samples into a reproducing kernel Hilbert space (RKHS) and computing the difference between their mean embeddings in that space and using a statistical test.

While MMD aligns the distributions in a high-dimensional RKHS, other methods simplify the alignment by defining it through explicit descriptive statistics. Correlation Alignment for UDA (CORAL) [80] is a very simple method that minimizes the domain shift by aligning the second-order statistics (covariance) of source and target distributions, ensuring that the linear relationships between the features become similar, being a post-adjustment after the extraction of the features. Deep CORAL enables the simultaneous extraction of features that lead to high performance on

the task and invariant to the domain, by jointly optimizing the training towards minimizing the CORAL and task loss.

However, aligning the distributions based solely on the covariance may be insufficient to capture their main characteristics, making it necessary to take into account higher-order statistics. Central Moment Discrepancy (CMD) [81], in contrast to MMD and CORAL, enables alignment that captures higher-order, non-linear differences between feature distributions by comparing the central moments of the features (mean, variance, skewness, kurtosis), which are important characteristics that allow a better alignment between source and target distributions.

In addition to the EDM methods that align based on statistical moments, a recent approach takes into consideration the global geometric structure and the cost of aligning the distributions. Optimal Transport [82] is a method that measures the minimum cost of transporting one distribution into another, based on a predefined distance between samples, and aims to learn a mapping between the source and target feature distributions which can be non-linear, allowing the alignment of complex distributions while maintaining geometric and semantic structure of the data.

IDM methods do not explicitly define a discrepancy metric but rely mostly on an adversarial training, in which the invariance is learnt indirectly by using an additional neural network component, such as a domain discriminator that forces the features to be indistinguishable across domains.

Domain-Adversarial Neural Network (DANN) [83] uses a feature extractor that is shared by two components: a label predictor and a domain classifier. The goal of the label predictor is to optimize the extraction of features that are relevant to the intended task, updating the feature extractor using standard backpropagation. The goal of the domain classifier is to determine the corresponding domain of the extracted features, updating the feature extractor through a gradient reversal layer, so that the optimization is performed in the direction of confusing the domain classifier, making the features more invariant to the domain. In this way, the features are relevant to the target task and invariant to the domain. The proposed architecture is represented in Figure 3.9.

Adversarial Discriminative DA [84] is similar to DANN, but instead of using a shared feature extractor, it employs two separate ones, which can be better if the domains are very heterogeneous, allowing a more domain specific approach. First, the source feature extractor is trained using only the source samples. In the next step, the weights of the source extractor are frozen, and a target feature extractor is trained with the goal of confusing the domain discriminator, thereby enabling the alignment of target features with the source. Conditional Adversarial DA (CDAN) [85] extends DANN by conditioning the domain discriminator on both the feature representations and the classifier predictions, allowing the network to align the feature distributions based on the domain and in the predicted class prioritizing samples based on their prediction entropy, which helps preserve class separability while achieving domain-invariant representations.

Contrastive learning (CL) can also be considered an EDM method that separates feature representations of pairs with opposite labels while bringing closer those with the same label in the

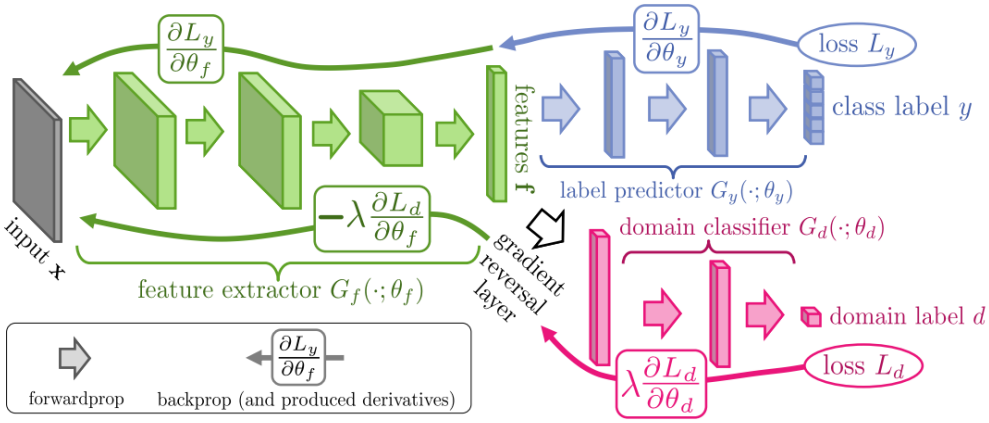


Figure 3.9: The architecture of DANN includes a feature extractor (green) and a label predictor (blue), forming a standard feed-forward network. UDA is achieved by adding a domain classifier (red) connected via a Gradient Reversal Layer (GRL), which multiplies the gradient by a negative constant during backpropagation. Training minimizes both label prediction (source) and domain classification (all samples) losses. This ensures feature distributions become indistinguishable, resulting in domain-invariant features. Replicated from [83]

feature space. In this way, it is possible to extract better image features that do not focus on background details and are therefore more invariant to variations in background and acquisition protocols, allowing for improved DA [86]. Chen et al. (2020) [87] introduced SimCLR, an approach that showed the significant potential of CL in the classification of images in an unsupervised way using augmentations. Li et al. (2025) [88] used CL techniques in mammographic image analysis, using CycleGAN and style blending to create images with different styles, using CL to align representations across both vendor styles and mammographic views.

EDM methods provide more control over the adaptation process, since they allow the use of an explicitly defined metric to guide the alignment in the feature space. In contrast, IDM methods are more flexible and dataset-specific, as they implicitly learn a way to align features, making them particularly useful when the differences between source and target distributions are complex and non-linear.

Image Translation

The image translation methods objective is to align the source domain with the target domain at the image level by transforming the image to the style of the ones in the target domain. Certain approaches rely on simple and direct transformations. For instance, Histogram Matching (HM) [89] finds a mapping function that transforms the intensity distribution of the source domain to match that of the target domain. Similarly, Fourier Domain Adaptation (FDA) [90] achieves style alignment by decomposing images into the frequency domain. It replaces the low-frequency amplitude spectrum of the source domain with the ones of the target domain, as these components are

associated to global statistics such as brightness and contrast, while preserving the phase spectrum and the high-frequency amplitude spectrum, which define the image semantics.

After the massive success of the Generative Adversarial Network (GAN) [91], many works started applying this to DA. One of the most well-known techniques in this category is CycleGAN [92], which consists of two mapping functions that transform the source domain into the target domain and vice versa, and it uses a cycle-consistency loss to ensure that the image transformed through the cycle coincides with the original image, as represented in Figure 3.10. DualGAN [93] is a technique very similar to CycleGAN, as it also employs a reconstruction error serving the same purpose as the cycle consistency loss, alongside an adversarial loss. However, the core principle of DualGAN is based on dual learning: two opposite tasks (X→Y and Y→X) are trained simultaneously, allowing them to mutually reinforce each other during training.

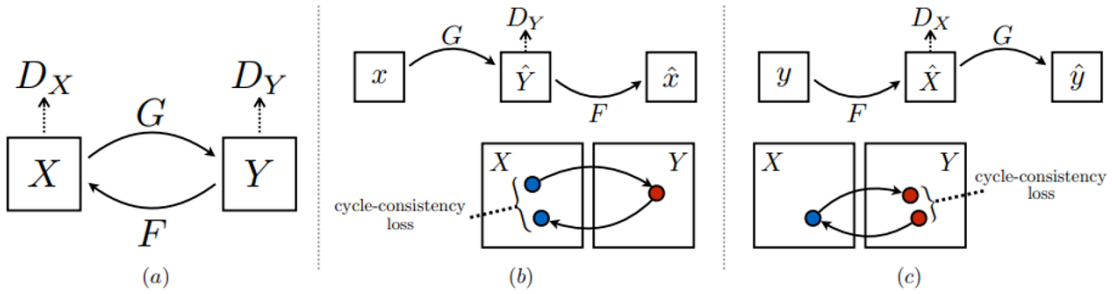


Figure 3.10: Overview of the CycleGAN method (a) The model contains two mapping functions $G : X \rightarrow Y$ and $F : Y \rightarrow X$, and associated adversarial discriminators D_Y and D_X. D_Y encourages G to translate X into outputs indistinguishable from domain Y, and vice versa for D_X and F. To further regularize the mappings, there are two cycle consistency losses that capture the intuition that if the images are translated from one domain to the other and back again it should be obtained the starting point: (b) forward cycle-consistency loss: $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$, and (c) backward cycle-consistency loss: $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$. Replicated from [92]

The applicability of these image translation methods is particularly relevant in medical image analysis, where they facilitate the simulation of data across different domains that can be used for DA. Wolterink et al. (2017) [94] used CycleGANs to do the conversion of brain 2D MRI scans to brain 2D CT scans, showing that cross-modality translation without requiring paired datasets is possible, which can enhance the DA for subsequent tasks such as segmentation or diagnosis. Tomar et al. (2021) [95] also employ a GAN framework, but enhance it with a self-attentive spatial adaptive normalization applied to the generator's activations, aiming to better preserve the anatomical geometry during image translation. Diniz et al. (2025) [96] employed a CycleGAN to translate 3T MRI images into 7T MRI images, aiming to exploit the higher resolution and contrast of 7T scans without the need for costly physical scanners.

In order to merge the advantages of the feature alignment methods and the image translation methods, new joint-learned approaches arose that consisted of transforming the source domain images to the target domain, and then do feature adaptation to close the remaining gap that might still exist.

Cycle-Consistent Adversarial DA (CyCADA) [97] adapts the representation both in the pixel and feature-level, improving the alignment between domains, using a pixel, feature, semantic, and cycle consistency loss. The source image is transformed into the style of the target domain, and by using the labels from the source domain, both the original and translated images are constrained to produce features that maintain performance on the task, ensuring semantic consistency. The translated image should be able to confuse the target domain discriminator, so that the images share the same style (pixel level). There is also a discriminator in the feature level to make sure that the features fail in the same distribution. Finally, the source image should be the same as the final reconstructed image. A schematic overview of the CyCADA framework is shown in Figure 3.11.

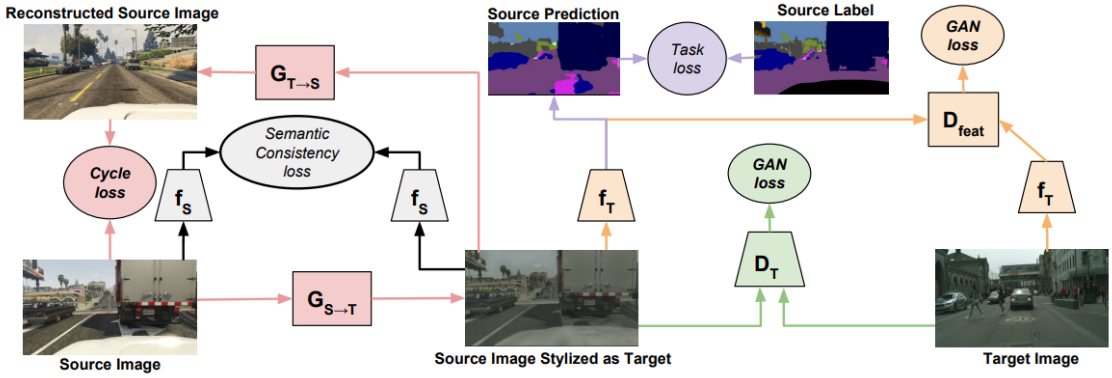


Figure 3.11: Cycle-consistent adversarial adaptation of pixel-space inputs. By directly remapping source training data into the target domain, the low-level differences between the domains are removed, ensuring that the task model is well-conditioned on target data. The image-level GAN loss (green), the feature level GAN loss (orange), the source and target semantic consistency losses (black), the source cycle loss (red), and the source task loss (purple) are represented. The target cycle is omitted. Replicated from [97]

Disentangled Representations

Several methods that are commonly used in DA have as objective to ensure separability between the features that are relevant to the content of the image (anatomy, for instance) - domain invariant features (DIFs) - and the ones that are related to the style (such as texture, scale, rotation) of the images and are more variable across domains - domain-specific features (DSFs). The core concept behind Disentangled Representation Learning (DRL) is to force the model to learn the joint distribution $P(X,Y)$ in a way that the image X can be factorized into two independent latent representations: $z_{content}$ (DIFs) and z_{style} (DSFs), allowing the classifier to use the $z_{content}$ for the final prediction, allowing a good DA [98].

Pioneering methods focused on the theoretical mechanism of factorization using Variational Autoencoders (VAEs) as a mathematical framework to decompose the latent representation of an

image. The central objective was to force the separation of the image's factors of variation into independent latent codes, such as content and style, primarily by modifying the VAE's loss function (the ELBO). Higgins et al. (2017) introduced the β-VAE [99] that added a hyperparameter $\beta > 1$ to penalize the Kullback-Leibler Divergence term, forcing the factors to be statistically independent. Similarly, methods like InfoVAE [100] utilized regularization based on Mutual Information Minimization to refine the independence objective in the latent space.

While VAEs established the fundamental theory, ensuring that the learned $z_{content}$ is truly invariant to domain shifts, DA requires stronger regularization and, therefore, the DRL methodology evolved to incorporate adversarial techniques.

In the context of DA, disentanglement is often achieved using adversarial training in which a domain discriminator network is employed that attempts to predict the source domain using only the content-relevant features. This adversarial competition forces the encoder to learn representations that are domain-agnostic. Yang et al. (2019) [101] implemented this by utilizing a VAE-GAN hybrid, where reconstruction loss minimized the difference between the encoders and generators (promoting them to be inverses to each other), and a GAN component encouraged disentanglement by generating images with swapped style. This architecture is represented in Figure 3.12.

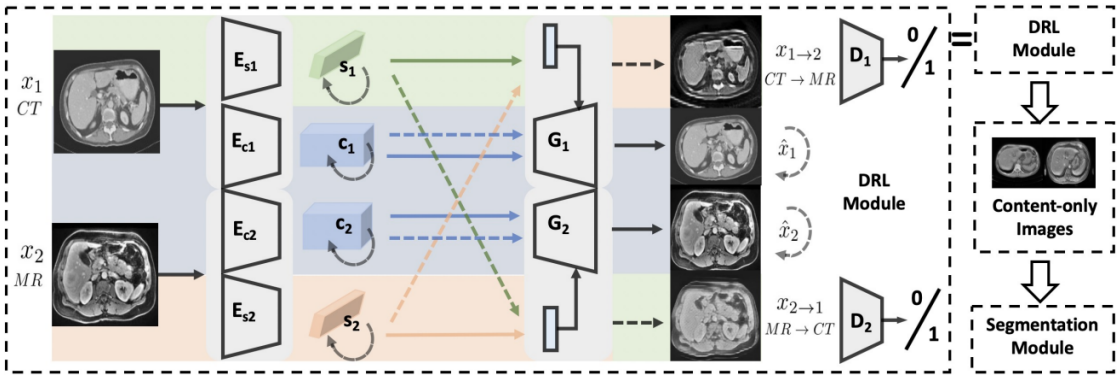


Figure 3.12: Overview of the proposed DRL framework. Left: Structure of the DRL module. The solid lines indicate in-domain reconstruction, while dotted lines represent cross-domain translation. Right: The complete pipeline for DA via Disentangled Representations. Replicated from [101]

Chartsias et al. (2019) [102] provided a structural approach to this problem, particularly in cardiac image analysis. They proposed the Spatial Decomposition Network that explicitly factorizes medical images into a spatial anatomical factor (s)-a multi-channel map representing cardiac structures - and a non-spatial modality factor (z) that encodes the specific image characteristics such as the contrast, brightness, and noise. The success of this method relied on enforcing the anatomical factor (s) to be discrete and categorical, which inherently prevented it from encoding modality information, thereby achieving domain invariance suitable for semi-supervised and multimodal learning.

The methodology continued to evolve through the integration of CL in this area. Unlike methods relying solely on reconstruction and adversarial losses, CL explicitly enforces disentanglement by maximizing the similarity of content features ($z_{content}$) across images that share the same underlying anatomy but possess different styles or domain characteristics. This approach ensures the content encoder learns representations that are inherently robust and invariant to domain shifts. For example, Gu et al. (2023) [103] proposed Contrastive Domain Disentanglement and Style Augmentation, which leverages contrastive loss to generate anatomical representations resilient to domain shift in medical image segmentation, representing a key direction in modern DRL for DA.

Pseudo labeling

PL gained popularity in the area of Deep Learning in 2013 with the work by Lee et al. [104], which introduced this method as a semi-supervised learning (SSL) approach, where the model's own predictions on unlabeled data are used as pseudo-labels to improve training. In this way, in a scenario where labeled data is scarce, these techniques can take advantage of that data and be used to improve the adaptation of the models, ensuring the alignment of features.

In the area of DA, PL methods are particularly relevant due to their ability to leverage the information that exists in the unlabeled target data. The incorporation of these samples using pseudo-labels enhances the model's discriminative ability, forcing it to learn more robust decision boundaries that also work within the target distribution, thereby reducing the impact of the domain shift. While most of these methods emerged within the SSL field, they have been adapted for DA, providing an effective strategy to improve the robustness of models across different distributions [105].

One of the initial relevant works in this area aimed to solve the high instability inherent to PL. To address this, research focused on consistency regularization. The π-Model [106] applies small perturbations to an unlabeled input and forces the model to produce consistent predictions using a soft consistency loss. Temporal Ensembling, introduced in the same article, also generates pseudo-labels but uses a consistency loss between the current prediction of the input and the exponential moving average (EMA) of its predictions from previous epochs, achieving better stability than the π-Model. The development of these techniques within the SSL framework was a crucial step, providing the necessary tools for the subsequent application of PL in DA methods.

After the huge success of EMA for stability, the Mean Teacher strategy [107] simplifies the EMA implementation by using the Teacher Model (whose weights are updated via EMA of the Student) to generate targets, while the Student Model is updated via backpropagation, enabling more stability. This strategy involves augmenting each unlabeled sample with two different augmentation sets, generating two versions of the same image, and passing them through the teacher and student model and computing their consistency loss (mean squared error between predictions) and the real classification loss in the case of labeled samples. This method is represented in Figure 3.13. In other works, the unsupervised task loss using pseudo labels is also integrated [108].

Perone et al. [109] successfully adapted the Mean-Teacher strategy for medical imaging segmentation, maintaining high model performance when applied to images acquired by different MRI scanners without using labeled data.

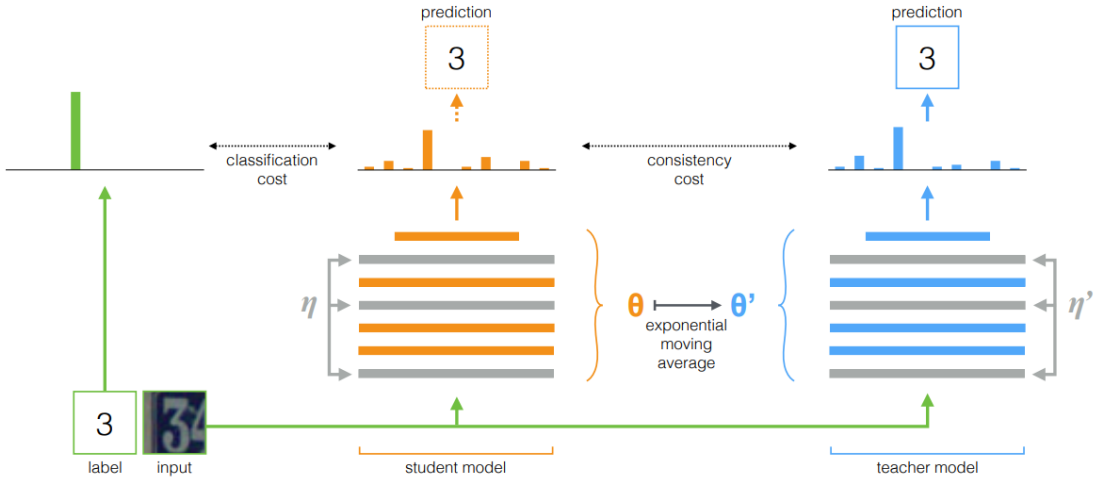


Figure 3.13: Representation of the Mean-Teacher method. A training batch is shown with a single labeled example. Both student and teacher models evaluate the input with noise. The softmax output of the student model is compared with the one-hot label using classification cost and with the teacher output using consistency cost. After the student gradient updates, teacher weights are updated as an EMA of the student's. While both models predict, the teacher is typically more accurate by the end of training. A training step with an unlabeled example would be similar, except no classification cost would be applied. Replicated from [107]

The Teacher-Student technique was further improved through specific mechanisms to increase the quality and precision of the pseudo-labels. A notable extension is AdaMatch [110], which adapts the FixMatch [111] framework specifically for DA scenarios. The core principle of forcing consistency between weakly and strongly augmented versions of an unlabeled image is maintained. However, to be applied in DA, AdaMatch employs a distribution alignment to ensure the target pseudo-label distribution matches the source label distribution, it uses a dynamic confidence threshold for the target domain based on the model's average confidence in the source domain and it uses a mechanism to stabilize the initial transition between the source and target distributions. Many works in medical image analysis employ this concept of filtering pseudo labels based on the confidence of the prediction [112, 113].

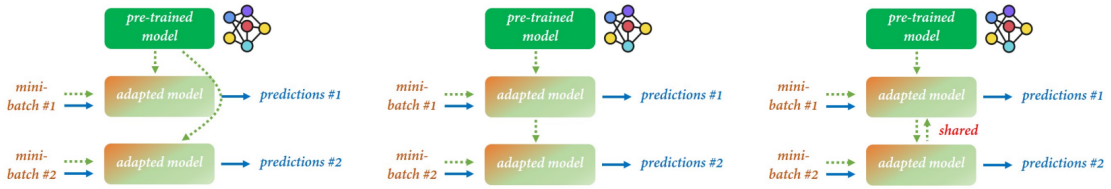
Despite all advances in the techniques, depending on the specific scenario and the dataset, PL can still have many problems. In cases of UDA, if the target domain is very different from the source domain, the pseudo labels originated by the model can always have a larger error associated which can cause the propagation of the error during training and problems in the convergence and adaptation of the model to the target domain [114]. Additionally, in highly imbalanced problems, the model tends to generate a larger number of pseudo-labels corresponding to the majority class, since its predictions for those samples are usually more confident, which can increase the imbalance problem, leading to even a bigger degradation in the performance in classifying minority

classes, highlighting the importance of applying techniques to address this issue [115].

Test Time Adaptation

In contrast to the previous DA methods described, Test-Time Adaptation (TTA) aims to solve the domain shift problem by enabling a more dynamic and personalized adjustment to the target domain. While other approaches typically utilize part of the target data during the training phase to adjust the model, TTA only uses the target data during the inference phase, updating the model's parameters and leveraging the target data in an unsupervised way that can be achieved, for instance, by using a proxy task and minimizing an auxiliary objective function. TTA methods have several advantages because they not only reduce the domain shift across domains and allow good performance of the models, but also solve the problems associated with privacy concerns, since it is possible to have a good adaptable model that can be easily tailored to the new domain data without sharing any data between institutions [116].

The methods can be divided into test-time DA, in which the target domain is divided by batches and the model is adjusted based on all those batches before making the prediction; test-time batch adaptation, in which each batch is independent of another, so the weights are just adjusted temporarily to make the predictions of each batch and online test-time adaptation in which there is a continuous adjustment of the weights of the model, so previous information about batches is kept [116]. In Figure 3.14 the different types are presented.



(a) test-time batch adaptation (b) online test-time adaptation (c) test-time domain adaptation

Figure 3.14: The TTA paradigm aims to adapt the pre-trained model to various types of unlabeled test data, including single mini-batch in (a), streaming data in (b), or an entire dataset in (c), before making predictions. During the adaptation process, either the model or the input data can be adjusted to reduce the distribution shifts. The dotted green arrow indicates the test-time training phase before inference, while the blue arrow corresponds to pure inference. Replicated from [116].

One of the fundamental adaptation strategies that paved the way for many variations TTA methods was presented by Li et al. (2016) [117], in which they introduced Adaptive Batch Normalization (AdaBN). This technique is based on adapting the batch normalization layer statistics (μ, σ^2), substituting the accumulated source statistics (which can vary severely if the domain shift is large) while keeping the affine parameters (γ and β) fixed to ensure that the semantic knowledge learned from the source domain is not lost, which leads to better DA. TENT [118] was based on the work of AdaBN, but instead of changing only the layer statistics, verified that it would be better to adjust the affine parameters by introducing a small cycle of "training" with the test data,

in which the gradient of the entropy loss in relation to these parameters, optimizing them towards a lower entropy.

Other methods are based on defining an appropriate self-supervised proxy task, as the one presented by Shuhau et al. (2020) [119] in which they developed a method that applies TTA in an unsupervised manner by leveraging a downstream task: solving a jigsaw puzzle. The main network is adjusted based on the performance of each new sample from the target domain on this auxiliary task. Other works in the medical field try to combine different approaches. Dong et al. (2024) [120] proposed a novel method, InTEnt, that allows the adaptation to a new domain using a single unlabeled test image, creating an ensemble of adapted models with different batch normalization layer statistics based on the target sample. Zhu et al. (2025) [121] implemented a Continual Test-Time Adaptation (CTTA) framework that uses entropy and variance to guide the adaptation and employs a student-teacher model to provide stable supervision while updating the network on unlabeled target data.

The methods that are based on adapting the batch normalization layers have as an advantage their simplicity, the slight adjustment of the model's parameters guarantees that the information learnt in the training is not lost and does not rely on choosing the right proxy task, which can be complicated. However, if the shift is big enough, adapting only these layers may not be enough. The ones that rely on using proxy tasks, normally change more parameters in the network (the last layers of feature extractor, for instance), which can be more efficient in correcting the domain shift. On the other hand, these methods face the challenge of selecting an appropriate proxy task that correlates well with the primary classification task, and the overall training and adjustment process requires more computational power and time.

One problem that is related to most of the TTA methods, if it is not taken into account, is the catastrophic forgetting [122] that is normally more prevalent in CTTA, leading to the loss of the initial information that was used to train the model. Ideally, the model should be constantly adaptable to be able to learn with the new data, but not lose the information that was previously learnt. To mitigate this problem between adaptability and stability, CTTA methods often employ techniques based on parameter restriction, such as using Fisher Information [122] to maintain the most essential weights, or EMA [123] to regularize the adaptation process.

3.3 Summary and Future Prospects

In recent years, the field of DA has evolved rapidly, driven by the huge need and importance to increase the model robustness and reduce the reliance on using a large quantity of new data when the task is similar but the domains change. In the biomedical area, this evolution is very important due to the scarcity of labeled data and the high variability between imaging protocols and modalities, resulting in a common gap that limits the performance of models already well trained in new settings. Consequently, the research trajectory has shifted from basic statistical corrections to more complex approaches that allow for more adaptable and autonomous approaches that correct the domain shift.

When labels for the target domain were available, the task of DA was significantly easier, relying mostly on fine-tuning or discrepancy metric minimization. However, in real-world scenarios, where the UDA setting is more frequent, many other techniques have been developed. Initially, techniques focused on utilizing target data during the training phase to promote an alignment of feature distributions, addressing the primary cause of domain shift, which are known as feature alignment methods. Subsequently, more advanced approaches started using the big advances in Generative AI to create synthetic data, leading to image translation-based methods, which shifted the focus from the feature level to the pixel level. Subsequent work brought a joint alignment at both the feature and image levels, providing a more robust solution to the domain shift problem.

Other techniques have focused on learning disentangled representations, aiming for a clear separation between DIF (that are relevant for the domain shift problem) and DSF. Meanwhile, PL was based on a different principle that consists of using the model to generate temporary labels for unlabeled data, thereby increasing the utility of unlabeled data. Finally, TTA is a very promising technique that involves adjusting the model in real-time to the new data, allowing for a more dynamic adaptation process, addressing critical challenges such as data privacy.

In conclusion, in recent years, there has been a massive evolution in making models more adaptable, leading to the development of a large number of diverse DA methods that target different problems and settings. However, due to the several challenges associated with the medical field, there are still limitations in applying these methods across certain domains. In the specific case of TPUS imaging, it often involves complex 3D volume pre-processing and is characterized by high noise levels and limited dataset sizes, which are common challenges in the medical field. Additionally, as described in Chapter 2, the diagnosis of conditions such as LAM avulsion is typically associated with high data heterogeneity, making it crucial to test and tailor DA methods specifically for these clinical conditions.

Chapter 4

Methodology

4.1 Dataset

4.1.1 Description

TPUS data was collected from two distinct centers: UZ Leuven and Sydney Urodynamic Centres. The Leuven cohort consisted of 169 patients evaluated 1 year after delivery, including 39 diagnosed with major avulsions and 130 intact patients. The Sydney dataset encompassed 84 symptomatic patients, comprising 24 major avulsion cases and 60 intact cases. Following the standardized protocol described in Section 2.2, the three central axial slices were extracted from each volume at 2.5 mm intervals, resulting in the total number of slices for each condition illustrated in Figure 4.1. As expected, since the Sydney dataset contains only symptomatic patients, a higher ratio of avulsion cases exists when compared to the Leuven dataset.

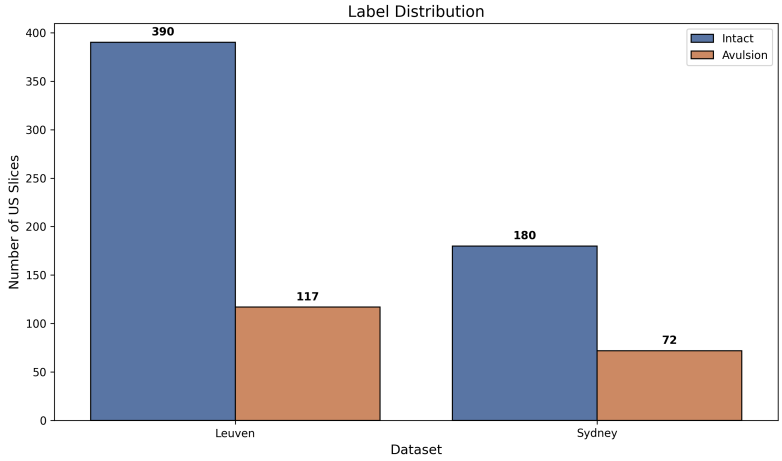


Figure 4.1: Dataset label distribution, comprising the three central axial slices from each volume.

Both datasets included demographic information and equipment specifications. Table 4.1 presents the statistical analysis of these data, performed using JASP [124].

In terms of the demographic data, significant differences were observed between datasets regarding patient age, body mass index (BMI), and parity (Student's t-test showed $p < 0.001$). These main differences can be attributed to the fact that both datasets were obtained from different clinical settings/care pathways (symptomatic and one year after delivery). Sydney dataset patients were significantly older than those in the Leuven dataset (age is a risk factor for LAM avulsion) and presented higher BMI and parity.

Regarding the US equipment used, Sydney scans were primarily obtained using the Voluson E22, whereas Leuven scans were acquired using different hardware versions, with the Voluson E10 being the most predominant. The Voluson E22 corresponds to a newer generation of US systems (released in 2022), focusing not only on acquisition speed but also on significant advances in spatial and contrast resolution, which results in superior image quality compared to the Voluson E10 (launched in 2014).

Therefore, there is both a covariate shift (due to differences in demographics and imaging equipment) and a prior shift (due to different label distributions), as explained previously in Section 3.2.1. In Figure 4.2, the differences between the two datasets are visible, as they were acquired with different US hardware, resulting in variations in brightness and contrast.

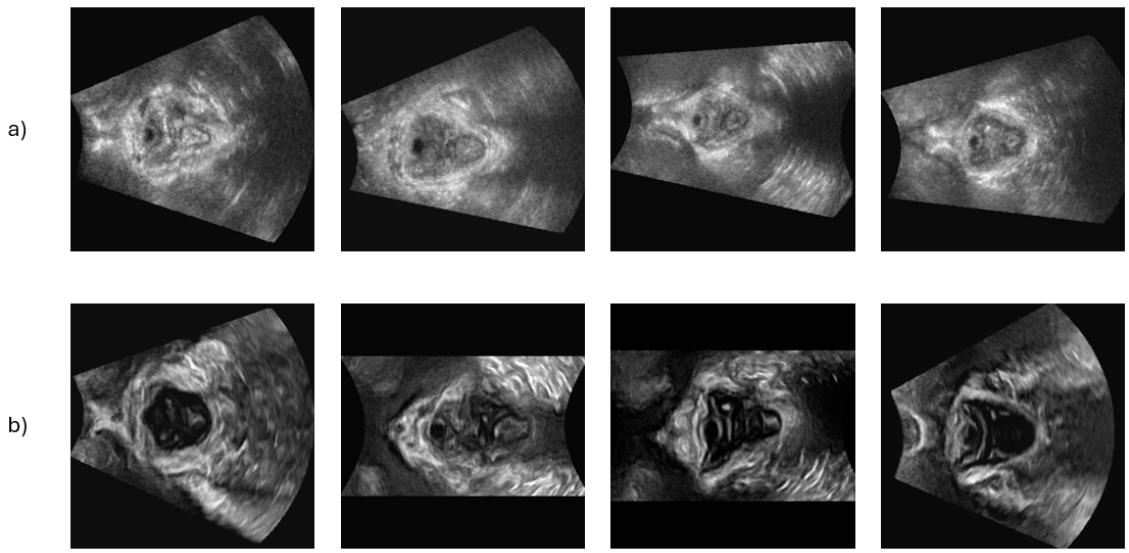


Figure 4.2: US images from the two datasets: (a) Leuven dataset acquired with Voluson E10; (b) Sydney dataset acquired with Voluson E22.

Table 4.1: Demographic characteristics of the study population and US equipment used stratified by dataset origin. This table summarizes the age, body mass index (BMI), parity and the diversity of equipment used for each dataset. N_missing represents the count of patients with unavailable data.

Variables	Dataset Leuven (N=169)	Dataset Sydney (N=84)	p-value
Age			$< .001$
N_missing	23	0	
Minimum	20	23	
Maximum	45	85	
Median (IQR)	32 (29-35)	52 (39-67)	
BMI			$< .001$
N_missing	24	0	
Mean $\pm$ SD	24.7 ± 5.8	28.1 ± 5.4	
Minimum	16.4	18.2	
Maximum	61.0	43.0	
Median (IQR)	23.1 (21.1 - 26.8)	27.2 (24.3 - 31.9)	
Parity			$< .001$
N_missing	25	0	
Minimum	1	0	
Maximum	4	8	
Median (IQR)	2 (1-2)	2 (1-3)	
US Equipment (Number of US)			
N_missing	0	2	
Voluson E10	145	0	
Voluson SWIFT	20	0	
Voluson S10 Expert	3	0	
Voluson E22	1	82	

4.1.2 Preprocessing

To maintain uniformity, all images underwent the same preprocessing sequence before any experiments. The first step was voxel size normalization to the dataset's mean voxel dimensions, ensuring that the interpolation did not result in significant loss of information. In this way, all images were standardized to represent the same biological scale. Subsequently, all images were cropped to the minimum bounding box including the fan-shaped region, ensuring that all relevant US data was preserved while removing redundant background information and noise. To preserve the original aspect ratio, images were padded with zeros to a square format. This ensured that subsequent resizing to a 448×448 resolution maintained the correct spatial proportions of the anatomical structures. Finally, intensity values were scaled using min-max normalization per image, mapping all voxel intensities to the $[0, 1]$ interval, ensuring that the models received inputs in a fixed range. Figure 4.3 illustrates the preprocessing pipeline applied to a US image.

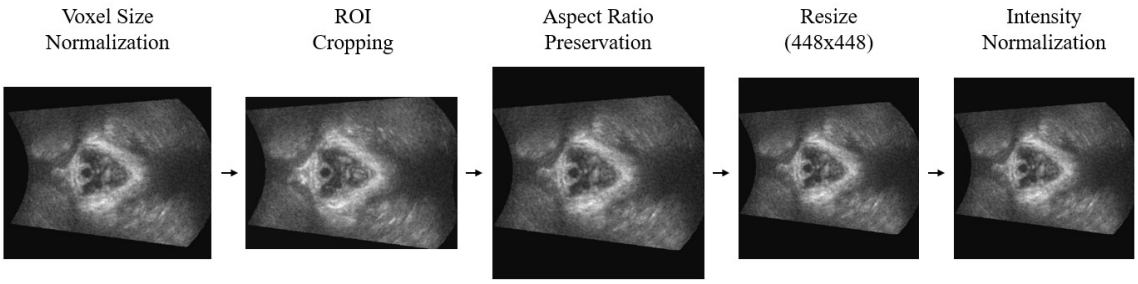


Figure 4.3: Overview of the image preprocessing pipeline. It includes voxel size normalization, region of interest (ROI) cropping, aspect ratio preservation, resizing to 448x448 resolution and intensity normalization.

4.2 Training Settings

The evaluation of DA methods requires the establishment of upper and lower performance bounds. The upper bound represents the ideal scenario, where the training, validation, and test sets were extracted from the same dataset, sharing the same distribution. Conversely, the lower bound is obtained when the model is evaluated on a dataset different (target domain) than the one on which it was trained and evaluated (source domain), reflecting the existing domain shift between the two datasets.

Comparability of the upper and lower bounds was ensured by doing a random subsampling stratified by label in the Leuven dataset, which contained more images than the Sydney dataset, matching the dataset sizes while maintaining the original label distribution. This approach removes any potential performance impact resulting from differences in the size of the training and validation sets, making it possible to directly compare the effect of different distributions in the training/validation and testing sets.

Due to the limited size of the datasets, the models were evaluated using a 5-fold cross validation, in which the test set iterates across the entire dataset, ensuring that performance was assessed on the entire dataset, while the training and validation set were randomly split in each fold. The data was split into 60/20/20% for training, validation, and testing, respectively. To avoid data leakage, this split was performed at the patient level, ensuring that all three slices from each patient were assigned to the same split. Label distribution was preserved across the different splits.

Before applying DA methods, it is important to have a base model with a good upper bound performance, capable of achieving high accuracy when tested within the same data distribution. It is also crucial to verify the existence of a domain shift (lower bound considerably smaller than the upper bound), so that the DA methods could reduce this gap in performance. Therefore, Section 4.3 describes the experiments conducted to obtain a good baseline model that is able to classify LAM avulsion. After obtaining the model, several UDA methods from the categories detailed in Chapter 3 were employed. Disentangled Representations based methods were not tested, since they normally rely on complex networks requiring large datasets.

The UDA methods leveraged unlabeled data from the target domain to improve classification performance on the test set. Figure 4.4 illustrates the generic application of a UDA method with the objective of increasing Leuven lower bound performance. Section 4.4 describes different experiments that implemented those methods, using identical training hyperparameters, data splits, and network architecture as the lower and upper bound setting to ensure direct comparability.

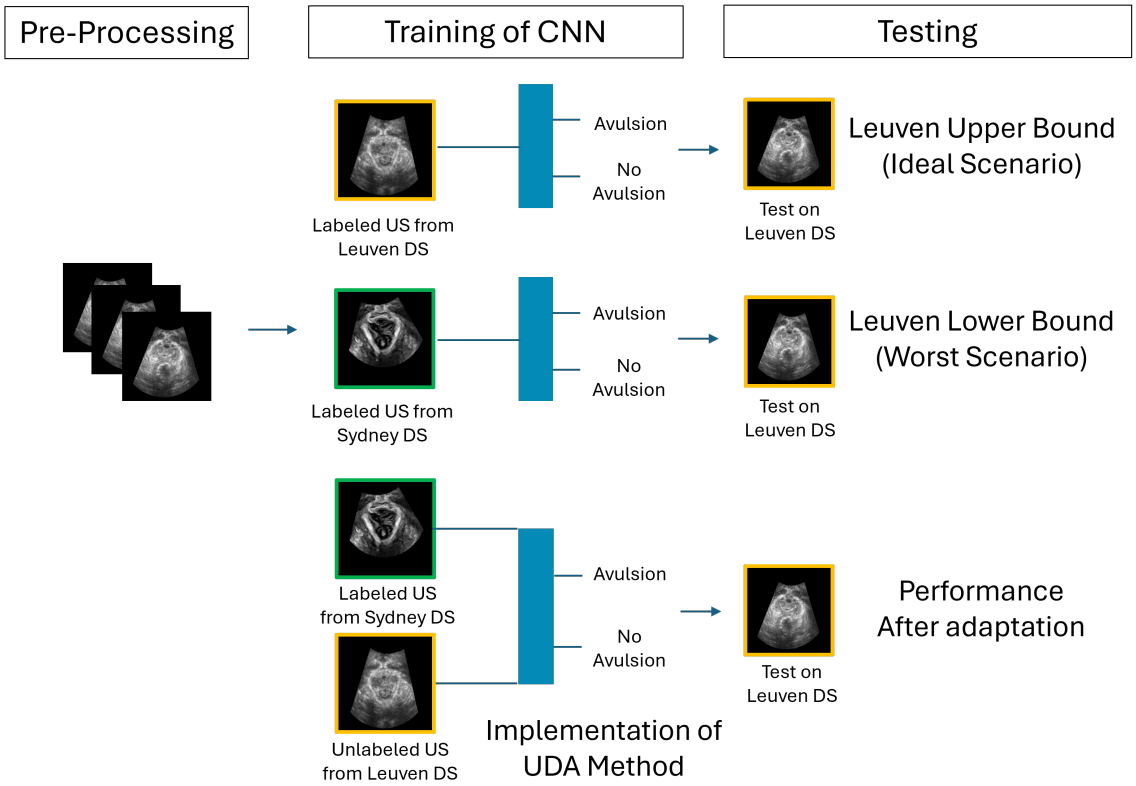


Figure 4.4: Proposed generic domain adaptation pipeline. It includes the preprocessing followed by the training of the baseline models. To bridge the domain gap, the UDA method leverages unlabeled samples in the training set from the target domain (Leuven) to improve the Leuven lower bound performance.

4.3 Model architecture and training details

To evaluate the impact of using different architectures for the classification of LAM avulsion, experiments were conducted to compare four distinct backbones. Due to the limited size of the dataset, architectures without a large number of parameters but with distinct structural designs were selected to evaluate which mechanisms lead to better performance in this scenario. A ResNet-18 [43] was employed, a widely used baseline in image classification; EfficientNet-B0 [45], a highly efficient convolutional architecture; ConvNeXtV2 [50], representing one of the most recent advancements in convolutional design as described in Section 3.1; and a ViT-Tiny patch 16 [46]

to evaluate transformer-based architectures. All models were implemented using PyTorch Image Models (timm) library [125].

To address the visual similarity and high correlation between the three central slices, the effect of applying data augmentations was studied, in order to add variability to the training set, facilitating the model's convergence and reducing the risk of overfitting.

Using MONAI library [126], random horizontal and vertical flips were applied, alongside random zooming into the central region of the image (scaling factor: 1.0 to 1.2) to reduce the influence of the US border on the model's decision. To simulate variability in US probe orientation, random rotations were performed within the range of $-15°$ to $15°$. Furthermore, to emulate acquisitions with different US equipments and to add data variability without compromising the symmetry of the levator region, random intensity scaling and contrast adjustment in an interval of [0.8,1.2] were applied. Finally, Random Gaussian Noise (with std of 0.02) was incorporated to simulate variations in tissue texture and image acquisition artifacts. All augmentations were applied with a probability of 0.4. These parameters were chosen to avoid excessive deformation of the image content. Figure 4.5 illustrates the effect of the different augmentation types on a US image.

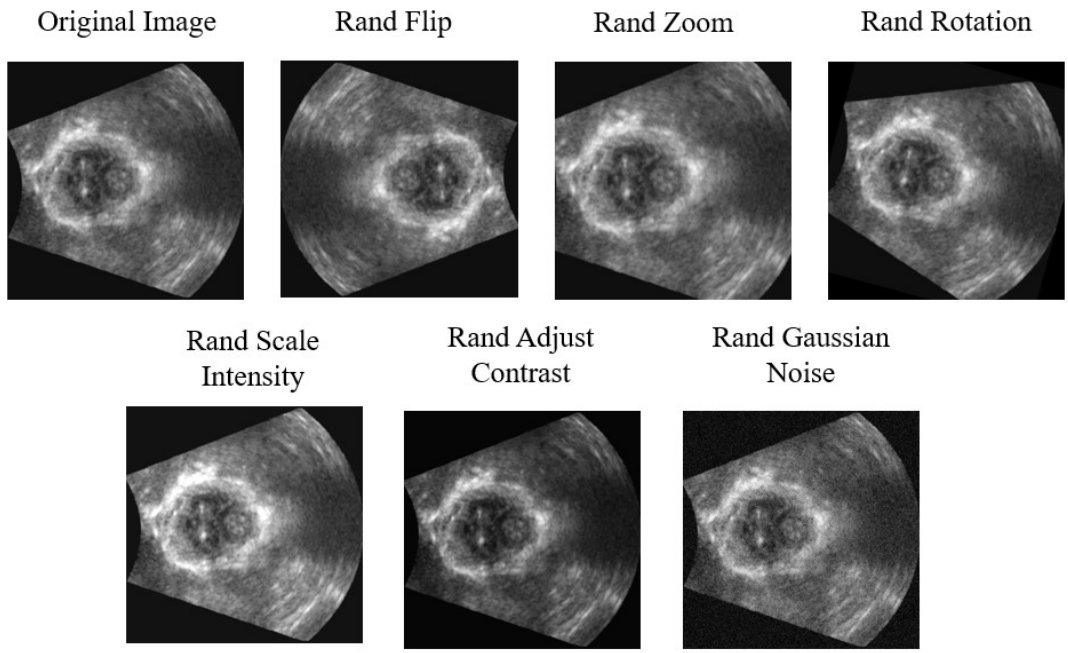


Figure 4.5: Different types of augmentations used during training to increase data diversity.

Additionally, an experiment was conducted to study the effect of incorporating a segmentation mask as part of the preprocessing pipeline. The masks were obtained using a pre-existing model developed by the Department of Development and Regeneration at UZ Leuven, which was trained on a separate dataset (different from the datasets used in this study), comprising 899 US images and subsequently tested on a cohort of 161 US images. The implementation utilized the nnU-Net framework based on MONAI, with a learning rate of 0.001 and an early stopping criterion set at 10 epochs. It achieved a Dice Similarity Coefficient (DSC) of 0.9487 ± 0.0313, demonstrating high

accuracy in segmenting the levator ani region. To ensure the inclusion of all anatomical borders of the levator, a morphological dilation was applied to the masks. Furthermore, the combined effect of using the segmentation mask and data augmentation was also analysed. Figure 4.6 presents the US images with the applied levator mask.

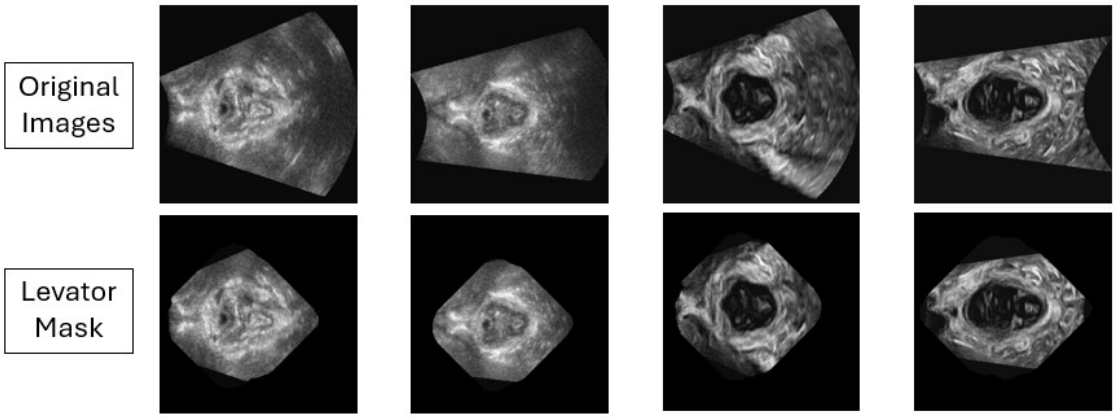


Figure 4.6: Segmentation of the levator region. Original US images (top) and the resulting ROI after applying the dilated segmentation mask (bottom).

The models were trained for 100 epochs with a batch size of 32, using the `AdamW` optimizer with an initial learning rate of 10^{-5} and a `ReduceLROnPlateau` scheduler. Early stopping was applied based on the validation AUC with a patience of 15 epochs. For the training process, a Binary Cross Entropy [127] with logits loss function was used, which combines the sigmoid activation and binary cross-entropy into a single and more numerically stable formulation. A weighting factor was incorporated to mitigate class imbalance within the dataset, which is calculated per fold. This loss function is represented in Equation 4.1, where N corresponds to the size of the training set, x_n denotes the logit output from the network, $y_n \in \{0,1\}$ represents the sample's label, w_n is the inverse frequency of the sample's class in the dataset, and $\sigma(x_n) = (1+e^{-x_n})^{-1}$ corresponds to the sigmoid activation function.

$$L_{BCE} = -\sum_{n=1}^{N} w_n \left[y_n \cdot \log(\sigma(x_n)) + (1-y_n) \cdot \log(1-\sigma(x_n)) \right] \tag{4.1}$$

Model performance was evaluated at the patient level by calculating the mean of the predicted probabilities across the three central slices.

4.4 DA Methods

After selecting a baseline model based on the previously described methods, which establishes the upper and lower bounds, several DA techniques were implemented to adapt the model to the target domain, aiming to reduce domain shift. Table 4.2 summarizes the different DA methods implemented and their corresponding family.

Table 4.2: Overview of the implemented domain adaptation methods, categorized by their respective families.

Family	**Method**
Feature Alignment (EDM)	MMD CORAL CMD
Feature Alignment (IDM)	DANN CDAN
Image Translation	HM FDA
Pseudo-Labeling	FixMatch
Test-Time Adaptation	TENT

In these experiments, a fully unsupervised DA setting was assumed, where no labeled target data was available for validation. This posed a significant challenge for model selection, as there was not a reliable proxy metric to evaluate target domain performance throughout the training process. This difficulty was further increased by the substantial domain shift, which restricted the use of other unsupervised proxy metrics reported in the literature [128].

Consequently, the source domain validation AUC was used as the model selection metric for all methods except the IDM methods. For the IDM methods, this source validation AUC was combined with the absolute difference between the domain discriminator accuracy and 0.5, ensuring the selection of a model that extracts relevant, domain-invariant features.

After selecting the optimal hyperparameters for each method, an additional training phase was conducted using the test AUC as the model selection criterion. This allows for an evaluation of the potential of each method (Oracle AUC), as it isolates their capacity to mitigate domain shift, independent of the challenges associated with unsupervised model selection.

4.4.1 Feature Alignment

Feature Alignment methods aim to minimize the domain shift by aligning the source feature distribution P (with samples $X = \{x_1, \ldots, x_n\}$) and the target feature distribution Q (with samples $Y = \{y_1, \ldots, y_m\}$) within a shared latent space. In terms of the EDM methods, in which an explicit discrepancy metric between the features is defined, MMD, CORAL and CMD were tested. For the IDM methods, DANN and CDAN were employed using adversarial learning to force the model to learn invariant features.

EDM Methods

The MMD [79] is a non-parametric metric that measures the distance between two probability distributions P and Q by comparing their mean embeddings in an RKHS. A Radial Basis Function (RBF) kernel was used that allows the mapping to the RKHS, as it is shown in Equation 4.2,

where $\|x_i - x_j\|^2$ denotes the squared Euclidean distance between two feature vectors, and σ^2 is the kernel bandwidth parameter that controls the sensitivity of the similarity measure. A multi-scale kernel approach was adopted by using a set of 5 bandwidths, allowing the MMD to capture alignments at different scales. The central bandwidth was determined using the median heuristic, a standard robust approach in the literature that sets the kernel width based on the median of pairwise distances between samples [129].

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{4.2}$$

Using this multi-kernel approach with L kernels, the MMD is calculated according to Equation 4.3. Term 1 and Term 2 compute the intra-domain similarity within the source and target domains, respectively (where a smaller $\|x_i - x_j\|^2$ leads to a higher RBF kernel output k_{σ_l}), while Term 3 quantifies the cross-domain similarity between features. By employing this metric as a regularization factor with a coefficient α in the loss function (Equation 4.4), the minimization process encourages an increase in the inter-domain feature similarity. Term 1 and Term 2 ensure that the model preserves the features structure of each domain, preventing the model from converging to a trivial solution, such as a feature collapse. Two different values of α (0.05 and 0.1) were tested, as well as an alpha that is linearly increased (factor of 0.01 to the max value of 0.1) across epochs, to prioritize discriminative features before enforcing domain alignment.

$$\mathrm{MMD}^2(X,Y) = \sum_{l=1}^{L} \left[\underbrace{\frac{1}{n^2}\sum_{i,j=1}^{n} k_{\sigma_l}(x_i, x_j)}_{\text{Term 1}} + \underbrace{\frac{1}{m^2}\sum_{i,j=1}^{m} k_{\sigma_l}(y_i, y_j)}_{\text{Term 2}} - \underbrace{\frac{2}{nm}\sum_{i,j=1}^{n,m} k_{\sigma_l}(x_i, y_j)}_{\text{Term 3}} \right] \tag{4.3}$$

$$\mathscr{L}_{total} = \mathscr{L}_{task}(f(x_s), y_s) + \alpha \cdot \mathrm{MMD}^2(f(X_s), f(X_t)) \tag{4.4}$$

CORAL [80] is a very simple method that focuses directly on aligning the second-order statistics (covariance) of the source and target features statistical distribution. The covariance matrices for the source domain ($C_S \in \mathbb{R}^{d\times d}$) and target domain ($C_T \in \mathbb{R}^{d\times d}$) are computed using extracted feature matrices X_S and X_T by centering the features around zero (subtracting the feature mean), as represented in Equation 4.5.

$$C_S = \frac{1}{n-1}(X_S - \bar{X}_S)^T(X_S - \bar{X}_S), \quad C_T = \frac{1}{m-1}(X_T - \bar{X}_T)^T(X_T - \bar{X}_T) \tag{4.5}$$

The CORAL loss defined in Equation 4.6 can be computed by calculating the squared Frobenius norm of the difference between the source covariance matrix C_S and the target covariance matrix C_T where d corresponds to the number of features, serving as a normalization factor for the dimensionality size.

$$\mathscr{L}_{\text{CORAL}} = \frac{1}{4d^2}\|C_S - C_T\|_F^2 = \frac{1}{4d^2}\sum_{i=1}^{d}\sum_{j=1}^{d}(C_{S,ij} - C_{T,ij})^2 \tag{4.6}$$

The CORAL loss was also incorporated as a regularization term, as shown in Equation 4.7. Given the small magnitude of the CORAL loss, constant values of α (5,10) were tested, as well as a linear ramp-up schedule that increases α from 0.5 to a maximum value of 10 throughout the training epochs.

$$\mathscr{L}_{total} = \mathscr{L}_{\text{task}}(f(x_s), y_s) + \alpha \cdot \mathscr{L}_{\text{CORAL}}(f(X_s), f(X_t)) \tag{4.7}$$

CMD [81] allows control of the number of statistical moments used to align the feature distributions of the source and target domains, a tanh normalization was applied to limit the range of the features to $[-1,1]$. The CMD loss can be computed using Equation 4.8, where $\mu(X)$ denotes the mean of the feature distribution, $\phi_k(X) = \mathbb{E}[(X - \mathbb{E}[X])^k]$ represents the k-th order central moment with $\mathbb{E}[\cdot]$ denoting the mathematical expectation operator, and the term $\frac{1}{2^k}$ serves as a normalization factor to prevent higher-order moments from dominating the loss.

$$\mathscr{L}_{\text{CMD}}(X_S, X_T) = \|\mu(X_S) - \mu(X_T)\|_2 + \sum_{k=2}^{K}\frac{1}{2^k}\|\phi_k(X_S) - \phi_k(X_T)\|_2 \tag{4.8}$$

By adjusting the hyperparameter K, the model progressively aligns increasingly complex statistical features, ranging from the mean (first-order) to higher-order shape properties like skewness and kurtosis. The value of K was set to 5, and different values of α were tested (0.1, 0.5) and a linear increase starting from 0.05 up to a maximum of 0.5.

$$\mathscr{L}_{total} = \mathscr{L}_{\text{task}}(f(x_s), y_s) + \alpha \cdot \mathscr{L}_{\text{CMD}}(f(X_s), f(X_t)) \tag{4.9}$$

IDM Methods

DANN [83] implements an unconditional adversarial alignment strategy, designed to learn domain invariant features. A domain classifier branch was introduced in the network, that uses the same extracted features but aims to classify them as source or target domain features as previously represented in Figure 3.9. The domain classifier consisted of a small network comprising a fully connected layer followed by a ReLU activation, a dropout layer and a final linear classification layer.

During backpropagation, in the domain discriminator branch, the gradients are calculated and used to update the weights of the domain discriminator, thereby improving its ability to classify the domain. A GRL was employed that intercepts the gradients calculated by the domain discriminator during backpropagation and scales them by a negative factor, forcing the feature extractor to update its weights in a direction that increases the domain discriminator's loss. This promotes the learning of domain-invariant features that are also useful for the classification of avulsion, as the task related gradients flow normally through the corresponding branch.

The gradient used to update the feature extractor weights can be calculated using Equation 4.10, where $\nabla_{\theta_f}\mathscr{L}_{total}$ is the total gradient computed for the feature extractor's parameters (θ_f), $\nabla_{\theta_f}\mathscr{L}_{task}$ is the gradient of the main task loss with respect to the feature extractor's parameters, λ is the adaptation factor that controls the impact of this reverse gradient in updating the feature extractor, and $\frac{\partial \mathscr{L}_d}{\partial \theta_f}$ represents the gradient of the domain classification loss ($\mathscr{L}_d$) with respect to the feature extractor's parameters.

$$\nabla_{\theta_f}\mathscr{L}total = \nabla\theta_f\mathscr{L}_{task} - \lambda \cdot \frac{\partial \mathscr{L}_d}{\partial \theta_f} \tag{4.10}$$

CDAN [85] is very similar to the DANN method, but it uses label information into the adversarial training process. For target domain samples, where ground truth labels are not accessible, CDAN uses the class predictions ($\hat{y}$) generated by the task classifier to condition the domain alignment. This is achieved by employing a multilinear mapping that consists in doing an outer product between the features extracted and the task predictions, and then using this transformed features in the domain classifier.

In this setup, the domain discriminator classifies the domain using features conditioned by the task label predictions. This forces a class-conditional alignment, ensuring that features from both domains are matched according to their predicted class labels. At the start of the training, the predicted label for the target domain is very prone to errors, accentuated by the existence of the domain shift. Therefore, it is important to use a small λ in the initial epochs.

Both of these methods were implemented testing different λ configurations. A λ scheduler that follows Equation 4.11 was used, in which γ controls the slope, p is defined as the ratio of the current epoch to the total number of epochs and λ_{max} is the maximum value λ can reach. Different values of λ_{max} (0.1, 0.5 and 1) were studied. The γ parameter was fixed at 10.

$$\lambda = \lambda_{max} \cdot \left(\frac{2}{1+\exp(-\gamma \cdot p)} - 1 \right) \tag{4.11}$$

4.4.2 Image Translation

The objective of image translation methods is to adapt source domain images to the style of the target domain. Given the visible differences in the appearance (brightness and contrast) of the US images between the two datasets, these techniques could reduce the domain shift. For this study, methods that are based on generative networks were not evaluated, since they typically require large datasets for stable training. Therefore, HM and FDA were implemented.

Histogram Matching

HM [89] is a DA approach that enforces alignment between the Cumulative Distribution Functions (CDFs) of two distinct domains. A Global Histogram Matching (GHM) and an Instance Histogram Matching (IHM) were implemented.

The intensity histogram of the target domain was computed while excluding background pixels, thereby mitigating the effect of differing background sizes across domains. This filtered target histogram is used to compute the Probability Mass Function (PMF), $P_d(i_t)$, which represents the relative frequency of each intensity level. The CDF can be obtained by integrating the PMF, as represented in Equation 4.12, where $F_t(z)$ represents the CDF of intensity z.

$$F_t(z) = \sum_{i=0}^{z} P_d(i_t) \tag{4.12}$$

Finally, the intensity mapping transformation is obtained by applying the inverse of the target domain's CDF to the source domain's CDF values ($F_s(i_s)$). This process maps the original pixel intensities to new, matched values ($i_{matched}$) that effectively align the source intensity distribution with that of the target domain, as defined by Equation 4.13.

$$i_{matched} = F_t^{-1}(F_s(i_s)) \tag{4.13}$$

In the case of the GHM, the global intensity histogram of the target domain (mean of all histograms) is used, while the IHM only selects a random image from the test set and computes the intensity of that histogram. Beyond transforming the training set to match the style of the target domain, the inverse approach for GHM was also evaluated. In this configuration, using the baseline model that was trained on source domain data, the performance was evaluated when employed on test set (target domain) that was transformed to mimic the style of the images in which the model was previously trained. These configurations are represented in Figure 4.7.

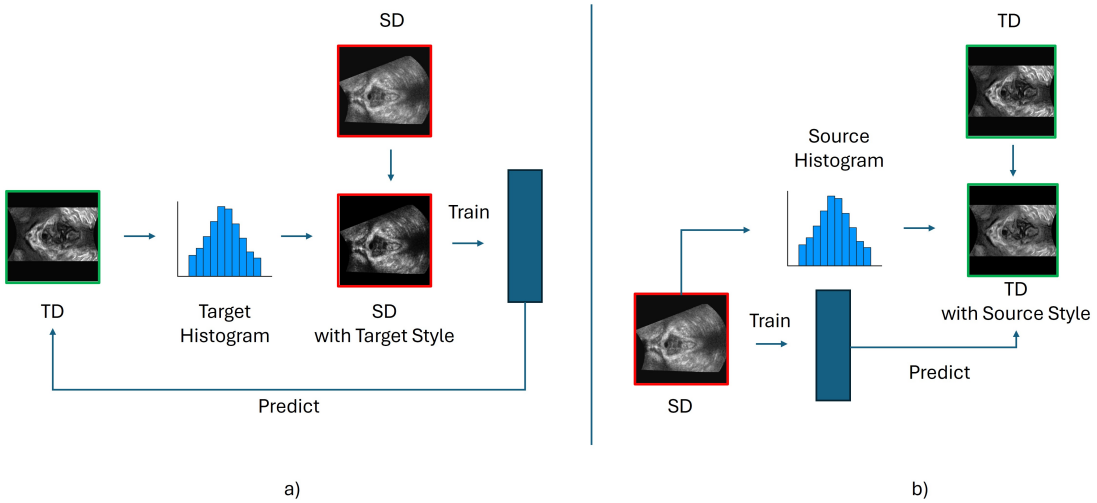


Figure 4.7: Schematic overview of the proposed HM frameworks. Red and green boundaries denote source (SD) and target domain (TD) images, respectively. (a) Adaptation of the training set to the target style, where the target histogram is derived from either a single random target image (IHM) or the mean histogram of the target domain (GHM). (b) Adaptation of the test set to the source style using a global source histogram.

Fourier Domain Adaptation

FDA [90] aims to align the styles of two domains by decomposing images into phase and amplitude components via the Fourier Transform. It is established that the low-frequency components of the amplitude spectrum capture the global style and appearance of the image, while the phase component contains its structural information.

FDA exploits this by replacing the low-frequency component of the source image's amplitude spectrum, $|\mathscr{F}(x_s)|$, with that of the target image, $|\mathscr{F}(x_t)|$, within a region defined by the mask M_β, as it is represented in Figure 4.8. The phase component of the source image is maintained. This representation is then mapped back to the spatial domain using the inverse Fourier transform ($\mathscr{F}^{-1}$), resulting in a transformed source image (x_{st}) as represented in Equation 4.14:

$$x_{st} = \mathscr{F}^{-1}\left(M_\beta \odot \mathscr{F}(x_t) + (1 - M_\beta) \odot \mathscr{F}(x_s)\right) \tag{4.14}$$

This process ensures that x_{st} maintains the structural content of the source image x_s, while adopting the style of the target domain D_t. The β controls how much of the source image's amplitude spectrum is replaced. A large β value can introduce artifacts, as higher frequencies within the amplitude spectrum capture semantic features of the images.

Similarly to the HM experiments (Figure 4.7), a global representation of the target domain amplitude spectrum (mean of the spectra) was also calculated, representing the global style of the image and was used to transform the source images (excluding background). Additionally, this was done in two directions: training a model that uses the training data transformed with the style of the target domain, as well as using the previously baseline trained model and testing it on transformed target domain images. Finally, FDA was also implemented in the standard way, where each source domain image is transformed by randomly selecting an image from the target domain. While this approach introduces greater transformation variability, it also carries the risk of incorporating outliers that do not accurately represent the target domain's style. All these experiments were conducted with $\beta \in \{0.01, 0.05, 0.09\}$ to study the effect of this parameter.

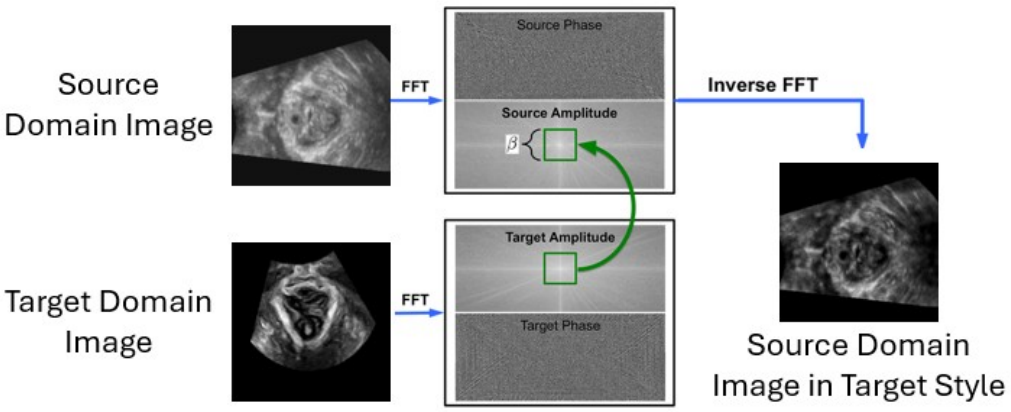


Figure 4.8: Representation of the FDA method, illustrating the effect of the β parameter on the size of the amplitude spectrum being replaced. Adapted from [90].

4.4.3 Pseudo Labeling

For the PL methods, since the baseline model showed a very poor performance in the target dataset, the process of generating pseudo-labels would be very erroneous. Therefore, due to the nature of the problem, a simple version of FixMatch [111] was tested but applied to the DA problem.

For each unlabeled image from the target domain, two versions are generated: one weakly augmented (using the same augmentations applied to the source domain) and one strongly augmented (using the previous transformations but with higher ranges and a probability of 0.6). The weakly augmented images are passed through the model to obtain a prediction. The maximum predicted probability (confidence), calculated as $\max(p, 1-p)$, is evaluated, and if this confidence exceeds the threshold $\tau = 0.70$, the resulting pseudo-label is used to train the model with the strongly augmented version of the image. While standard FixMatch implementations utilize $\tau = 0.95$, a reduced threshold of $\tau = 0.70$ was used, leading to a higher quantity of pseudo-labeled samples, since the model is not confident in the target domain due to the existence of a large domain shift. Figure 4.9 represents the PL approach applied.

As represented in Equation 4.15, due to the fact that the unsupervised loss ($\mathscr{L}_{\text{pl}}$) calculated using pseudo-labels can contain significant noise when the model is still learning, λ is designed to gradually increase the influence of the target domain, ensuring that the supervised loss ($\mathscr{L}_{\text{sup}}$) guides the models training. The same scheduler as the one used in DANN and CDAN (Equation 4.11) was used, testing λ_{max} of 0.1, 0.5 and 1.

$$\mathscr{L}_{\text{total}} = \mathscr{L}_{\text{sup}} + \lambda \cdot \mathscr{L}_{\text{pl}} \tag{4.15}$$

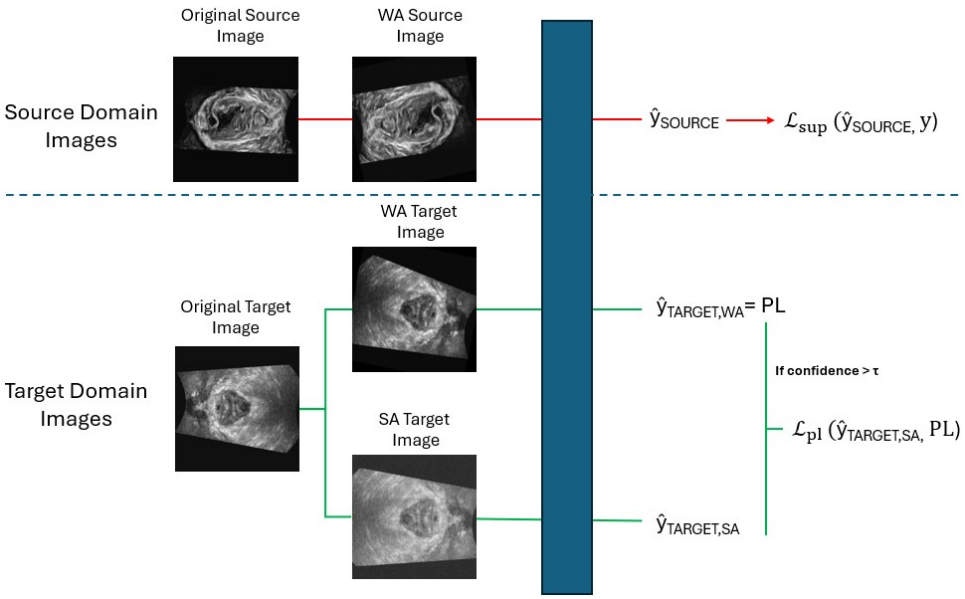


Figure 4.9: Overview of the PL method. The framework utilizes Weak Augmentation to generate high-confidence pseudo-labels for the target domain, which subsequently supervise the model's training on the Strong Augmentation branch.

4.4.4 Test Time Adaptation

Unlike the previous methods, TTA methods address the domain shift by adapting model parameters during the inference phase. TENT [118] was implemented, which performs this optimization by minimizing the predictive entropy of the model's output on unlabeled target data.

This method considers that a good model should produce highly confident, low-entropy predictions on target domain samples. Consequently, a brief training loop is performed on the target domain test set, during which the affine parameters of the normalization layers are updated to minimize prediction entropy. Given the binary nature of the classification task, the entropy loss can be defined in Equation 4.16, where f_θ denotes the model with parameters θ, and $p = \sigma(f_\theta(x))$ the predicted probability for an input x.

$$\mathscr{L}_{\text{ent}} = -\left[p\log(p) + (1-p)\log(1-p)\right] \tag{4.16}$$

By minimizing this loss, the model refines its decision boundaries, effectively becoming more confident in the target domain. It performs forward-backward passes for a number of steps, updating the normalization parameters using the gradient of $\mathscr{L}_{\text{ent}}$, allowing the alignment of its internal feature statistics with the target distribution. The rest of the architecture is frozen.

This method was tested in the pre-trained baseline model using a learning rate of 10^{-5} , testing the performance when using 1 and 10 adaptation steps, as suggested in the literature. This configuration ensures that the model does not converge to a trivial solution, that completely minimizes entropy at the expense of the discriminative ability of the model, which may occur with higher learning rates or an increased number of adaptation steps.

4.5 Evaluation Metrics

The performance of the classification models after applying those methods was evaluated using the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. The models produce a probabilistic output that can be used to generate a prediction (if the output is above a certain threshold it corresponds to a positive labeled sample in case of a binary problem). The ROC curve can then be computed by plotting the True Positive Rate (TPR) on the y-axis and the False Positive Rate (FPR) on the x-axis across different threshold values. The AUC value is obtained by integrating this curve, representing the probability that a randomly chosen positive sample will be ranked higher than a randomly chosen negative one, and is commonly used to evaluate the discriminative ability of the model. An AUC of 0.5 corresponds to a random classifier, while an AUC of 1.0 represents a model capable of perfectly separating the classes [130].

In addition to the AUC, t-Distributed Stochastic Neighbor Embedding (t-SNE) [131] plots were computed to visually assess the transformations that occur in the feature space when applying the different feature alignment methods. t-SNE is a dimensionality reduction technique that can be used to show high-dimensional data in a 2D space, preserving local structures and relationships.

Source and target domain features were used to analyse whether two distinct clusters were formed: one comprising positive avulsion cases and another comprising intact cases from both domains.

To statistically evaluate the performance improvements achieved by the DA methods compared to the baseline, the Wilcoxon signed-rank test [132] was employed. This non-parametric test is suitable for paired samples, comparing the mean AUC scores obtained by each method against the lower bound across the five cross-validation folds. For a set of differences d_i between paired observations, the test statistic W can be calculated using Equation 4.17, where R_i denotes the rank of the absolute difference $|d_i|$ and $\text{sgn}(d_i)$ is the sign function. A one-sided Wilcoxon signed-rank test was conducted to test the hypothesis that the DA method achieves a superior AUC compared to the baseline. Results are considered statistically significant for p-values lower than 0.05.

$$W = \sum_{i=1}^{N} [\text{sgn}(d_i) \cdot R_i] \tag{4.17}$$

Additionally, given the limited sample size (5 folds), Cohen's d [133] is reported to quantify the magnitude of the observed improvement, accounting for the small scale of the dataset. This metric is defined as the ratio of the mean difference to the standard deviation of those differences (s), as represented in Equation 4.18. Values of $d \approx 0.2$, 0.5, and 0.8 are conventionally interpreted as small, medium, and large effect sizes, respectively [134].

$$d = \frac{\bar{x}_{\text{method}} - \bar{x}_{\text{baseline}}}{s} \tag{4.18}$$

Chapter 5

Results and Discussion

5.1 Baseline Training

In this section, the results of an initial study focused on training a baseline model for the classification of LAM avulsion are reported. Initially, the performance of using different architectures was evaluated, and based on these results, the impact of implementing data augmentations and incorporating segmentation masks was further investigated.

Table 5.1 presents the Mean AUC values for the classification of LAM avulsion across various architectures, highlighting the trade-off regarding the number of parameters. As discussed in Chapter 3, the convnextv2_atto is a convolutional network that incorporates mechanisms from vision transformers. It combines the advantage of requiring less training data to achieve high performance with a superior ability to capture global image features. Consequently, this network achieved the best performance on both datasets (0.8083 and 0.7154 for the Sydney and Leuven datasets, respectively) while maintaining the lowest parameter count. The vit_tiny_patch16_448 served as the second-best model, likely due to its transformer based architecture, followed by the efficientnet_b0, which, as a pure convolutional network without an attention mechanism, yielded lower results. Finally, resnet18 failed to perform adequately in this task, likely because its large parameter count relative to the dataset size led to overfitting, lacking the advanced feature extraction capabilities of modern architectures.

It is also important to mention that all results showed a high standard deviation in the AUC across the five folds, primarily caused by the limited size of the dataset. In addition to being prone to overfitting, selecting the optimal model using the validation set poses extra challenges. Due to the reduced dataset size, the validation set may not accurately reflect the performance on the test set, leading to the selection of a sub-optimal model that does not perform well in some of the folds.

Additionally, it is interesting to note that across all architectures, the Sydney dataset consistently yielded higher performance than the Leuven dataset, despite the normalization of sample sizes. This discrepancy can be attributed to the superior image quality of the Sydney dataset, acquired using more recent US equipment. Furthermore, in contrast to the Leuven dataset, whose

patients were randomly selected one year postpartum, the Sydney cohort comprises symptomatic patients; this results in a higher prevalence of LAM avulsion cases, which may facilitate the network's learning process.

Table 5.1: Mean AUC ($\pm std$) performance comparison across Sydney and Leuven datasets with standard deviation across the 5 folds. Bold values indicate the highest Mean AUC achieved per dataset.

Architecture	**Parameters (M)**	**Sydney AUC**	**Leuven AUC**
convnextv2_atto	3.39	**0.8083 ± 0.1477**	**0.7154 ± 0.0875**
efficientnet_b0	4.01	0.6283 ± 0.1027	0.5795 ± 0.0844
vit_tiny_patch16_448	5.54	0.7242 ± 0.0802	0.6449 ± 0.0500
resnet18	11.17	0.5258 ± 0.1895	0.4308 ± 0.1838

Given its superior performance, the convnextv2_atto architecture was selected for the subsequent experiments. The impact of implementing data augmentations and incorporating a levator hiatus mask on performance was investigated, assessing both the upper bound (testing on images from the same dataset) and the lower bound (evaluating model robustness when testing on images from a different clinical institution).

As shown in Table 5.2, applying data augmentation led to a slight improvement in the upper bound performance for both datasets. This increase is attributed to the added variability during training, which is crucial for mitigating overfitting, particularly in this scenario, in which the model is using three similar slices from the same patient. Furthermore, augmentations yielded an improvement in the lower bound performance for the Leuven dataset, as the increased diversity helped reduce the negative impact of domain shift.

The integration of the segmentation mask resulted in a significant performance increase across both datasets, most notably in the Sydney cohort. This can be explained by the mask's ability to preserve only the relevant diagnostic signal while removing noise that is irrelevant for the diagnosis. Consequently, this also led to a substantial improvement in lower bound performance; by filtering out image noise, the model training process is facilitated, and it learns more invariant features, becoming significantly more robust to differences between the datasets. Adding augmentations on top of the masks increased the upper bound performance, achieving good results in the classification of LAM avulsion, AUC of 0.9233 and 0.8051 in Sydney and Leuven dataset, respectively.

Table 5.2: Impact of data augmentation and segmentation masks on convnextv2_atto performance across domains. The Lower Bound AUC represents performance when testing on a dataset different from the training set, while the Upper Bound AUC corresponds to testing on the same dataset used for training. Bold values indicate the highest Mean AUC achieved per dataset.

	Test Sydney		**Test Leuven**	
Configuration	**Upper Bound AUC**	**Lower Bound AUC**	**Upper Bound AUC**	**Lower Bound AUC**
Base	0.8083 ± 0.1477	0.6033 ± 0.2326	0.7154 ± 0.0875	0.5192 ± 0.1850
Aug	0.8167 ± 0.1333	0.5892 ± 0.2276	0.7513 ± 0.1600	0.5654 ± 0.1566
Base + Mask	0.9067 ± 0.0962	$\mathbf{0.8033 \pm 0.1108}$	0.7718 ± 0.1103	0.7128 ± 0.1803
Aug + Mask	$\mathbf{0.9233 \pm 0.0751}$	0.7892 ± 0.1423	$\mathbf{0.8051 \pm 0.1476}$	$\mathbf{0.7308 \pm 0.2296}$

Figure 5.1 illustrates the t-SNE projections of the feature space generated by the baseline model (using data augmentation). These plots highlight the significant performance gap between the upper and lower bounds: AUC values drop from 0.8167 to 0.5892 when trained on the Sydney dataset, and from 0.7513 to 0.5654 when trained on the Leuven dataset. This observed divergence in the latent space provides a visual representation of this domain shift. A clear decision boundary is visible between the avulsion and intact features within the source domain, consistent with the model's high discriminative performance. In contrast, features from the target domain are not clearly separable by class and are situated in a distinct region of the latent space.

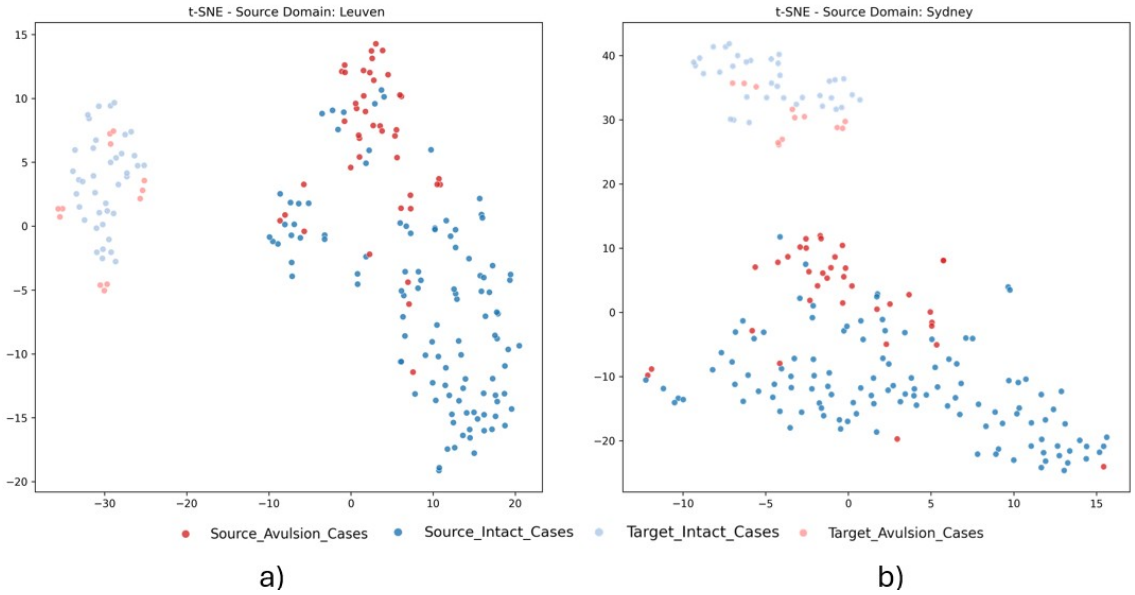


Figure 5.1: Representation of the feature space of the baseline model from a representative fold, illustrated by t-SNE plots. a) Model trained on Leuven dataset b) Model trained on Sydney dataset.

Despite the higher performance achieved by using the segmentation mask in the classification of LAM avulsion, this configuration was not used in the subsequent DA experiments. Although effective on these datasets, the robustness of the segmentation model has not been thoroughly validated across diverse clinical sites. This limits its reliability as a pre-processing step, as it relies

on the assumption that the segmentation masks are consistently accurate. Since the objective of this work is to evaluate DA techniques, it is not ideal to introduce a pre-processing step that could potentially fail when applied to data from different hospitals. Furthermore, the augmentation only configuration already exhibits a significant domain shift while maintaining strong upper-bound performance on both datasets. Therefore, it provides a suitable baseline to assess the efficiency of DA techniques without adding unnecessary complexity to the pre-processing phase.

5.2 DA Methods

In the following subsections, the results for each method are presented, organized by their respective methodology family. Each method was evaluated under various hyperparameter configurations, as previously detailed in Section 4.4, to examine their specific impact on the DA process. Following the selection of the optimal hyperparameters for each method, a benchmarking section is provided. It includes the computation of the Oracle AUC for each method, representing their potential in mitigating domain shift, as well as a detailed analysis of performance gains and and an evaluation of their statistical significance.

5.2.1 Feature Alignment

EDM Methods

As can be seen in Table 5.3, DA methods such as MMD, CORAL, and CMD are strongly dependent on hyperparameter configurations. Through exploring different values of α to align the source and target feature distributions, the results indicate that these methods struggle to consistently bridge the performance gap between the lower and upper bounds.

For the Sydney dataset, CORAL was the EDM method that facilitated a better adaptation process, achieving an AUC of 0.6550 ± 0.1455 ($\alpha = 10$), considerably higher than the lower bound and with a smaller standard deviation across folds. This suggests that, for this alignment direction, matching the covariance of the feature distribution is the most effective strategy. For the Leuven dataset, implementing MMD as a regularization factor with an α of 0.05 led to a substantial increase in the mean AUC compared to the lower bound (0.6756 ± 0.2189). However, there was also an increase in the standard deviation, meaning that this alignment completely failed in some folds.

It is also possible to see that the best results obtained using CMD did not outperform those achieved with MMD and CORAL, which can be attributed to the fact that CMD also depends on the number of moments being aligned (K value), making parameter tuning more sensitive to the specific setting. Additionally, if the K value is not properly tuned, it may lead to the alignment of noise in the distributions, since aligning higher-order moments can capture non-discriminative details when K is too high, or under-align the distributions when K is too low, which leaves important domain discrepancies unresolved.

Table 5.3: Mean AUC (± std) for convnextv2_atto after the implementation of EDM methods, such MMD, CORAL, and CMD, with different alpha values. Bold values indicate the highest Mean AUC achieved by each method per dataset.

		Mean AUC	
Setting	**Alpha**	**Test Sydney**	**Test Leuven**
Upper Bound	—	$\mathbf{0.8167 \pm 0.1333}$	$\mathbf{0.7513 \pm 0.1600}$
MMD	0.05	0.5475 ± 0.1493	$\mathbf{0.6115 \pm 0.2143}$
	0.1	$\mathbf{0.6258 \pm 0.1636}$	0.6115 ± 0.2407
	Linear Increase	0.4867 ± 0.1672	0.6090 ± 0.2227
CORAL	5	0.5783 ± 0.1736	0.5577 ± 0.1420
	10	$\mathbf{0.6550 \pm 0.1455}$	$\mathbf{0.5872 \pm 0.1893}$
	Linear Increase	0.6417 ± 0.2348	0.5603 ± 0.1430
CMD	0.1	0.5667 ± 0.2622	$\mathbf{0.5974 \pm 0.1902}$
	0.5	0.4992 ± 0.1004	0.5808 ± 0.1695
	Linear Increase	$\mathbf{0.6033 \pm 0.1898}$	0.5641 ± 0.1347
Lower Bound	—	$\mathbf{0.5892 \pm 0.2276}$	$\mathbf{0.5654 \pm 0.1566}$

These methods aim to align the global distribution of features, which can lead to the misaligned mapping of classes. For instance, avulsion cases from the target domain may be incorrectly mapped to the feature representations of intact source cases during training. While this achieves a successful global alignment of feature distributions, it sacrifices the discriminative ability of the model in the target domain. This is accentuated by the fact that the lower-bound performance is already close to random (AUC of 0.5), which indicates a substantial domain shift. As a result, the target-domain features lie far from the source-domain feature region, which hinders correct alignment. Figure 5.2 illustrates this behaviour, showing the evolution of the latent feature space across training epochs when applying the CMD method with an α of 0.5. In epoch 1, there is a clear formation of two clusters of points, each representing one domain. Across the epochs, the global feature distributions become aligned, with the points moving closer together in feature space; however, there is no clear boundary separating the intact and avulsion cases.

Figure 5.2: Representation of the feature space from a representative fold across different epochs (1,10,20) using CMD method with an α of 0.5.

IDM Methods

Table 5.4 displays the results obtained for the IDM methods that use adversarial training to promote the alignment of the domains. For both of the datasets, the best results were achieved by implementing CDAN with a λ_{max} of 0.5, resulting in an AUC of 0.5950 $\pm$ 0.0935 and 0.7231 $\pm$ 0.1143 for the Sydney and Leuven dataset, respectively. These values are both superior to the mean AUC of the lower bound and with a smaller std, resulting in an improvement in the performance of the model, successfully reducing the domain shift.

Table 5.4: Mean AUC ($\pm$ std) for convnextv2_atto after the implementation of IDM methods, such DANN and CDAN, with different λ_{max} values. Bold values indicate the highest Mean AUC achieved by each method per dataset.

		Mean AUC	
Setting	λ_{max}	**Test Sydney**	**Test Leuven**
Upper Bound	—	$\mathbf{0.8167 \pm 0.1333}$	$\mathbf{0.7513 \pm 0.1600}$
DANN	0.1	0.5525 ± 0.1802	$\mathbf{0.6423 \pm 0.1443}$
	0.5	$\mathbf{0.5650 \pm 0.1669}$	0.6295 ± 0.0828
	1	0.4817 ± 0.1720	0.5590 ± 0.1446
CDAN	0.1	0.5050 ± 0.1800	0.5744 ± 0.1581
	0.5	$\mathbf{0.5950 \pm 0.0935}$	$\mathbf{0.7231 \pm 0.1143}$
	1	0.5325 ± 0.0497	0.6654 ± 0.1542
Lower Bound	—	$\mathbf{0.5892 \pm 0.2276}$	$\mathbf{0.5654 \pm 0.1566}$

As expected, the best results achieved by CDAN were able to outperform those obtained with DANN (0.5650 $\pm$ 0.1669 and 0.6423 $\pm$ 0.1443 for the Sydney and Leuven datasets, respectively). This can be explained by the fact that CDAN aligns the domains while taking into consideration the labels of the samples, as opposed to DANN. This results in the alignment of intact features from the source domain with intact features from the target domain, and avulsion features from the

source domain with avulsion features from the target domain. Figure 5.3 shows the reorganization of the feature space across epochs, where by Epoch 20, source and target features corresponding to the same label are visibly closer in the embedding space.

Figure 5.3: Representation of the feature space across different epochs (1,10,20) using CDAN method with an λ_{max} of 0.5.

The results also show that the performance of these methods is dependent on the regularization factor of the domain discriminator loss (λ). In these methods, maintaining a balance between the domain discriminator and the feature extractor is important, so that neither component dominates the training process. While low values of λ lead to a training in which the features of the two domains are not properly aligned, leading to high domain discriminator accuracy, high values of λ force the discriminator into a state where it consistently misclassifies domains, driving its accuracy toward 0% as it essentially learns to invert the labels. To achieve good domain alignment using these methods, during the initial epochs of training, the model should focus on distinguishing intact from avulsion cases and then consistently force the features extracted to be invariant, resulting in a discriminator accuracy close to 50%. Therefore, the best results were mainly achieved using a λ_{max} of 0.5, allowing a better balance. Figure 5.4 shows the impact of the λ parameter in the domain discriminator accuracy using λ_{max} of 0.1, 0.5 and 1.

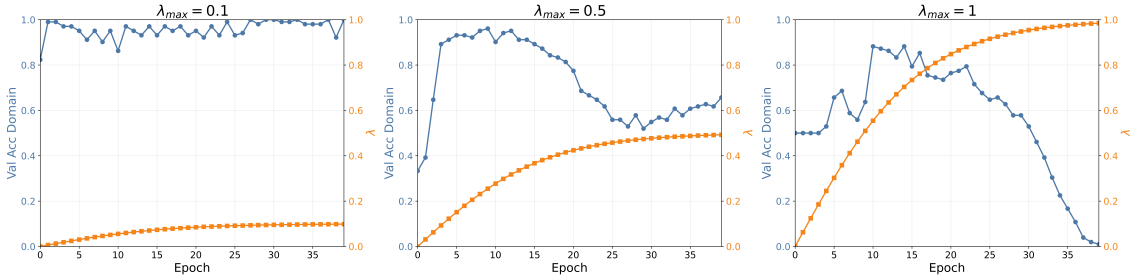


Figure 5.4: Impact of the regularization factor λ on the domain discriminator accuracy (blue line) throughout training, using λ_{max} of 0.1,0.5 and 1. The orange line represents the scheduled increase of λ over epochs.

5.2.2 Image Translation

Histogram Matching

The HM method was successfully implemented, aligning the intensity distributions of both domains, as shown in Figure 5.5. The CDF of the original image was properly aligned to the reference CDF, resulting in a transformation of the image style. As can be seen from the global reference histograms of both domains, the Leuven dataset typically has a higher brightness than the Sydney dataset and by applying HM, Leuven images become darker, while Sydney images exhibit increased intensity levels.

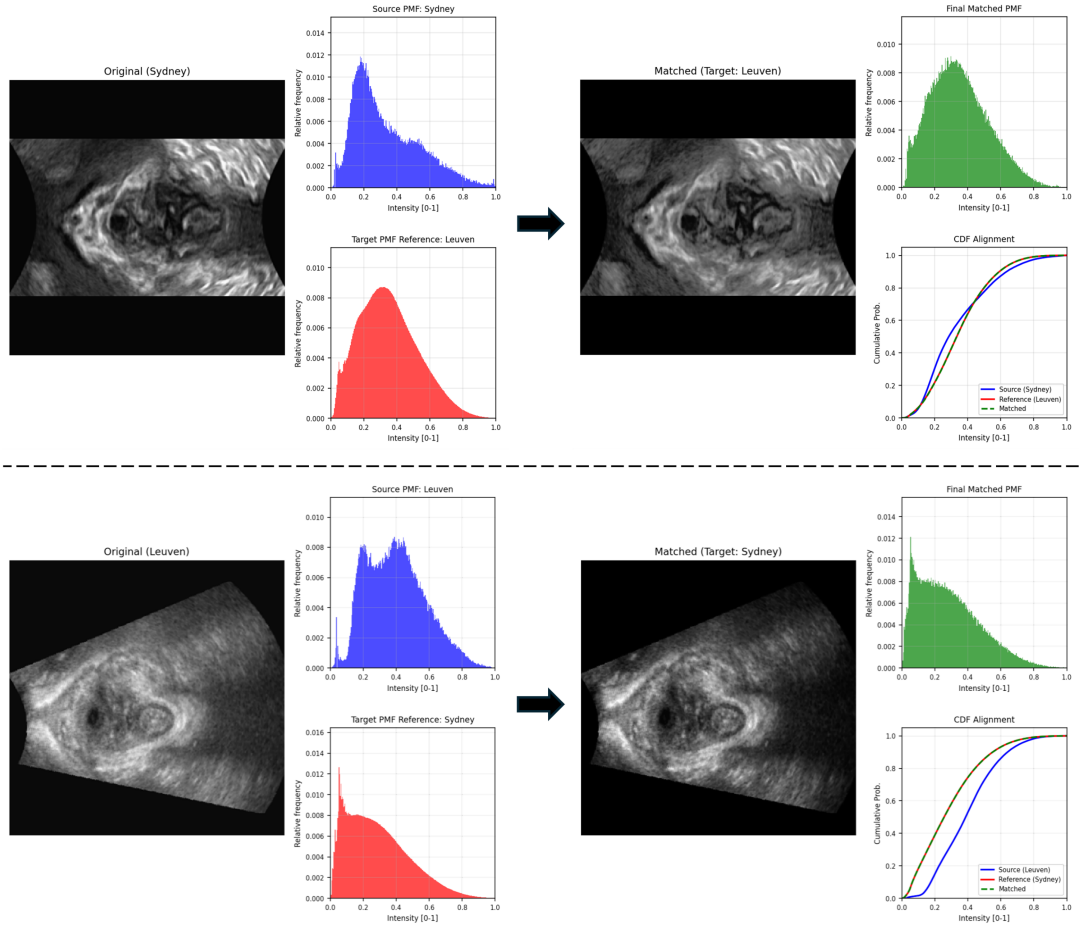


Figure 5.5: Visual representation of the HM process, demonstrating the alignment of image intensity distributions between domains, displaying the PMF and CDF to illustrate the convergence of the source image statistics towards the target domain reference.

Figure 5.6 shows the results of applying HM as a DA method. This method was able to reduce the domain shift, as the models achieved better performance when compared to the lower bound. Using the same augmentations as the baseline model, for the Sydney test set, GHM transforming the training images to the target domain style was able to achieve a mean AUC of 0.7008

(±0.1597). Conversely, for the Leuven test set, the best performance was obtained with the Instance HM method, achieving a mean AUC of 0.6321 (±0.1751).

Since the intensity distributions of the domains were explicitly aligned, the application of intensity and contrast augmentations could nullify these effects, effectively reversing the benefits of the HM. Therefore, despite not being completely comparable to the lower and upper bounds due to the differences in the augmentation set, the effect of removing these augmentations was studied. For the Sydney dataset, there was a very small increase in performance in GHM configuration, however, due to the size of the standard deviation (± 0.2026), this increase can be considered negligible. In the Leuven dataset, this led to a considerable decrease in performance. Despite possibly changing the intensity distribution slightly, the augmentations introduced variability in the training, which led to a more stable training and the mitigation of the domain shift.

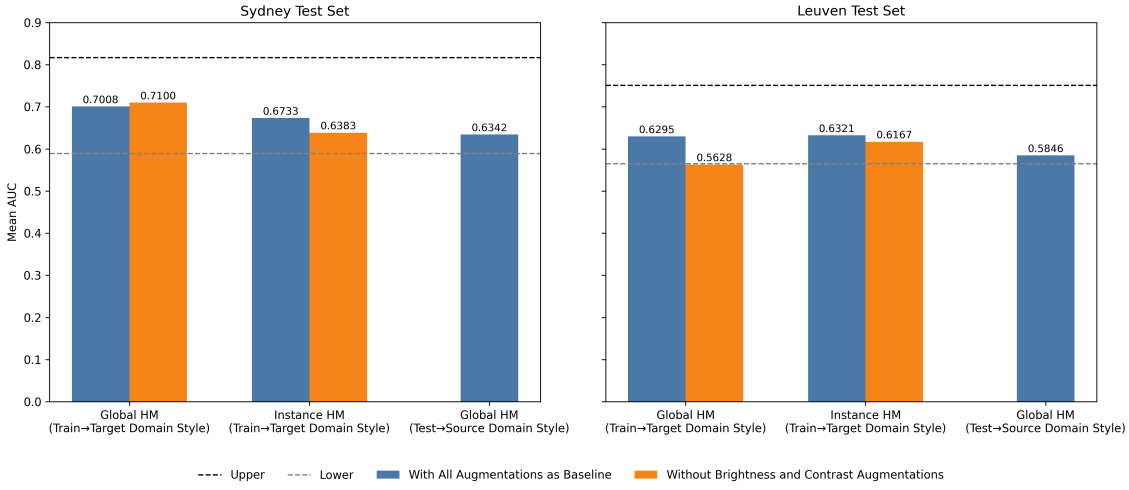


Figure 5.6: Results of applying the IHM and GHM methods to reduce domain shift. Blue bars represent the results obtained using the same augmentations as the baseline models. Orange bars represent the results when removing the random brightness and contrast augmentations.

Fourier Domain Adaptation

FDA was implemented, testing the different configurations (Global FDA and Instance FDA), using various β values, as shown in Table 5.5. For the Sydney dataset, Instance FDA led to a considerable increase in AUC, with a lower standard deviation than the lower bound, achieving 0.6617 ± 0.1963 ($\beta = 0.01$). In this case, the variability in style introduced by randomly selecting a target-domain representation was advantageous when adapting from Leuven to Sydney images. On the other hand, for the Leuven dataset, the best performance was achieved using the Global FDA configuration, where the training set was transformed to the target-domain style, reaching a mean AUC of 0.6564 ± 0.0759 ($\beta = 0.05$). This can be explained by the fact that, in this dataset, the stability introduced by using a global representation of the target dataset allowed better adaptation of the model, reducing the potential noise from outliers.

The results of implementing Global FDA in the opposite direction (i.e., adjusting test images to the source domain style) did not outperform the ones obtained in the other configurations. This indicates that it is more effective to adapt the training process to the target-domain style, rather than attempting to align test images with the source domain after, as the FDA transformation can introduce spectral artifacts that the model was never exposed to during training, leading to the degradation of the performance. Introducing this process during training allows the model to classify avulsion accurately even in the presence of some cases with artifacts, increasing its overall robustness to such distortions while adjusting the style.

Table 5.5: Mean AUC ($\pm$ std) for convnextv2_atto after the implementation of different FDA configurations (Global FDA and IHM FDA), with different β values. Bold values indicate the highest Mean AUC achieved by each configuration per dataset.

		Mean AUC	
Setting	**Beta**	**Test Sydney**	**Test Leuven**
Upper Bound	—	$\mathbf{0.8167 \pm 0.1333}$	$\mathbf{0.7513 \pm 0.1600}$
Global FDA (Train→Target Domain Style)	0.01	$\mathbf{0.6050 \pm 0.2351}$	0.5859 ± 0.1432
	0.05	0.5475 ± 0.1828	$\mathbf{0.6564 \pm 0.0759}$
	0.09	0.5142 ± 0.1731	0.5718 ± 0.1619
Global FDA (Test→Source Domain Style)	0.01	0.5858 ± 0.1871	$\mathbf{0.5577 \pm 0.1380}$
	0.05	0.5783 ± 0.1567	0.4692 ± 0.1174
	0.09	$\mathbf{0.6050 \pm 0.1457}$	0.4231 ± 0.1311
Instance FDA (Train→Target Domain Style)	0.01	$\mathbf{0.6617 \pm 0.1963}$	$\mathbf{0.5756 \pm 0.1194}$
	0.05	0.5475 ± 0.1773	0.5436 ± 0.1326
	0.09	0.5033 ± 0.1963	0.5487 ± 0.1109
Lower Bound	—	$\mathbf{0.5892 \pm 0.2276}$	$\mathbf{0.5654 \pm 0.1566}$

Additionally, the best results for the Sydney and Leuven datasets were obtained when using a β value of 0.01 and 0.05, respectively. As it is possible to verify in Figure 5.7, higher values of β lead to the appearance of artifacts in the transformed images, in accordance with the theoretical presumption that higher frequencies of the amplitude spectrum contain structural and semantic details. This is more visible in the Sydney dataset, where the loss of structural details and increased blurriness explain why the optimal performance was obtained with a smaller β value of 0.01.

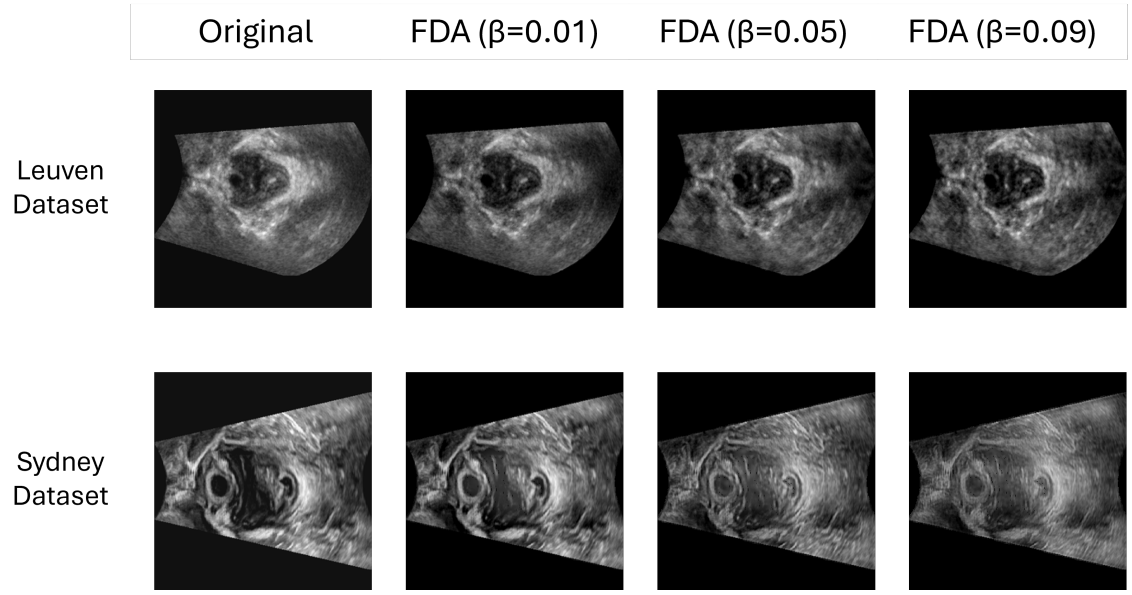


Figure 5.7: Visual representation of the effect of different β values (0.01, 0.05, and 0.09) on the image appearance for both datasets.

5.2.3 Pseudo Labeling

Table 5.6 presents the results of implementing FixMatch method using a fixed threshold ($\tau = 0.70$) while varying the λ_{max} parameter to control the weight of the pseudo-label loss during optimization. Implementing this method with $\lambda_{max} = 0.1$ yielded the best performance on both datasets, achieving 0.6425 ± 0.1964 and 0.5885 ± 0.1371 for the Sydney and Leuven datasets, respectively. Since pseudo-labels can be erroneous, mainly during the initial epochs, a smaller λ_{max} ensures that they serve only as auxiliary supervision, while the labeled source domain data remains the dominant factor in the loss optimization process.

Table 5.6: Mean AUC ($\pm$ std) for convnextv2_atto after the implementation of a PL method (FixMatch) with different λ_{max} values. Bold values indicate the highest Mean AUC per dataset.

		Mean AUC	
Setting	λ_{max}	**Test Sydney**	**Test Leuven**
Upper Bound	—	$\mathbf{0.8167 \pm 0.1333}$	$\mathbf{0.7513 \pm 0.1600}$
FixMatch	0.1	$\mathbf{0.6425 \pm 0.1964}$	$\mathbf{0.5885 \pm 0.1371}$
	0.5	0.6200 ± 0.1839	0.5449 ± 0.1569
	1	0.5900 ± 0.1774	0.5679 ± 0.2131
Lower Bound	—	$\mathbf{0.5892 \pm 0.2276}$	$\mathbf{0.5654 \pm 0.1566}$

The effect of changing the confidence threshold was also studied, as it directly impacts both the reliability and the number of pseudo-labels used. Figure 5.8a displays the number of pseudo-labels generated across epochs. As expected, using a smaller τ value results in a considerable

increase in the number of pseudo-labels, particularly in the initial epochs. However, these labels are more likely to be erroneous, potentially misguiding the model's learning process. Conversely, employing higher τ values (e.g., 0.80) leads to fewer pseudo-labels, for which the model is more confident and which are more likely to be correct. However, this significantly reduces the number of used target domain samples, complicating the DA process. Consequently, the best results were achieved using an intermediate τ value (0.70), which effectively balances the quantity of pseudo-labels with their reliability, as shown in Figure 5.8b.

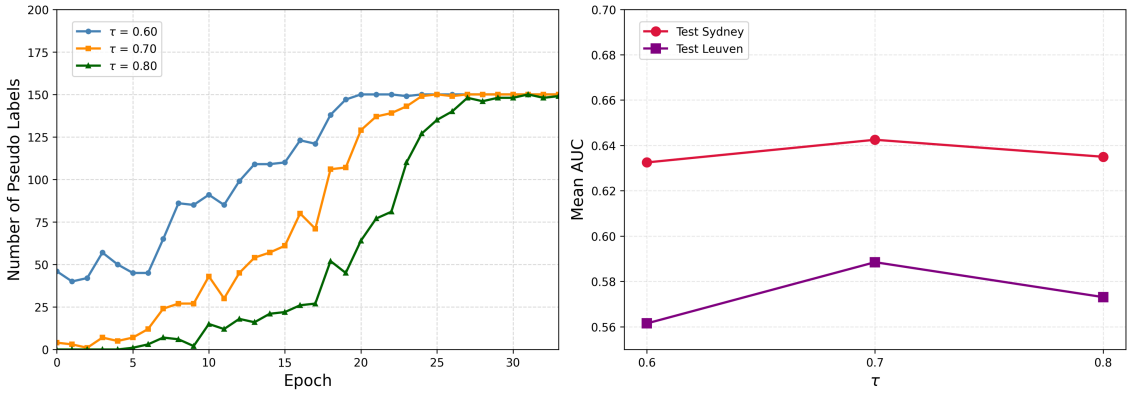


Figure 5.8: Impact of the confidence threshold (τ) on the PL method. (a) Evolution of the number of pseudo-labels generated per epoch. (b) Influence of the threshold τ on the final Mean AUC performance for both Sydney and Leuven test sets.

5.2.4 Test Time Adaptation

TTA methods aim to reduce domain shift by adjusting model weights during inference. Table 5.7 shows the results when applying Tent to the baseline models. This method improved performance when adjusting the weights of the model trained on Leuven data for testing on Sydney data, achieving a performance of 0.6583 ± 0.2202 in one step. Conversely, on the Leuven dataset, this method resulted in performance degradation.

Although increasing the number of steps drives the adaptation process toward lower entropy, this optimization may lead the model in the opposite direction of what is intended, particularly since the baseline performance is very close to being random (AUC of 0.5). Consequently, the model fails to improve because it lacks sufficient discriminative ability in the target domain to effectively guide the adaptation process, which was verified when using 10 steps in the optimization cycle.

The architecture used, convnextv2_atto, is not ideal for this TTA method because it relies mainly on LayerNorm2d and GRN rather than Batch Normalization, the setting for which Tent was originally developed. Given that Layer Normalization layers do not rely on global running statistics (such as mean and variance), they operate locally by computing these values for each input sample. Since Tent primarily optimizes affine parameters, it is restricted to a single sample optimization that lacks the domain-level anchor provided by Batch Normalization to stabilize the

adaptation process. This may lead to overfitting on individual samples, thereby failing to provide the stability required to mitigate domain shifts.

Table 5.7: Mean AUC (± std) for convnextv2_atto after the implementation of a TTA method (Tent) with different adaptation steps. Bold values indicate the highest Mean AUC per dataset.

		Mean AUC	
Setting	**Steps**	**Test Sydney**	**Test Leuven**
Upper Bound	—	$\mathbf{0.8167 \pm 0.1333}$	$\mathbf{0.7513 \pm 0.1600}$
Tent	1	$\mathbf{0.6583 \pm 0.2202}$	$\mathbf{0.5026 \pm 0.0911}$
	10	0.5808 ± 0.2106	0.4346 ± 0.2024
Lower Bound	—	$\mathbf{0.5892 \pm 0.2276}$	$\mathbf{0.5654 \pm 0.1566}$

5.3 Performance Benchmarking

Following the previous experiments that aimed to tune the hyperparameters for each method and assess their impact on the DA process, the best configurations were fixed for each dataset and used to conduct a more detailed analysis, as shown in Table 5.8. For the HM methods, despite not always achieving the best performance, the configuration applying all augmentations was the one tested in this experiment, since it is comparable with the lower and upper bounds.

Table 5.8: Summary of the best-performing methods per DA family. It shows the configuration that yielded the highest Mean AUC on the respective test set across all hyperparameter experiments.

Family	**Method**	**Test Sydney**	**Test Leuven**
Feature Alignment (EDM)	MMD	$\alpha = 0.1$	$\alpha = 0.1$
	CORAL	$\alpha = 10$	$\alpha = 10$
	CMD	$\alpha =$ LI	$\alpha = 0.1$
Feature Alignment (IDM)	DANN	$\lambda_{max} = 0.5$	$\lambda_{max} = 0.1$
	CDAN	$\lambda_{max} = 0.5$	$\lambda_{max} = 0.5$
Image Translation (HM)	Global HM (Train -> Target)	With Aug	With Aug
	Global HM (Test -> Source)	With Aug	With Aug
	Instance HM (Train -> Target)	With Aug	With Aug
Image Translation (FDA)	Global FDA (Train -> Target)	$\beta = 0.01$	$\beta = 0.05$
	Global FDA (Test -> Source)	$\beta = 0.09$	$\beta = 0.01$
	Instance FDA (Train -> Target)	$\beta = 0.01$	$\beta = 0.01$
Pseudo Labeling	FixMatch	$\lambda_{max} = 0.1$	$\lambda_{max} = 0.1$
Test-Time Adaptation	Tent	steps=1	steps=1

Additionally, in a fully UDA setting, where target samples are unavailable during training and validation, identifying the epoch at which the model achieves the best discriminative capacity in the target domain can be a difficult task. Therefore, an Oracle AUC was computed as explained in

Chapter 4.4. For the methods that implied using the already pre-trained baseline model, the Oracle AUC was not possible to obtain, since it did not imply a training process.

As visible in Tables 5.9 and 5.10, all DA methods exhibit an Oracle AUC higher than the Oracle AUC of the lower bound, successfully reducing the domain shift. Moreover, the Oracle AUC was significantly higher than the Mean AUC obtained using the validation AUC of the source domain as an early stopping criterion. This reinforces that the model selected during training did not often correspond to the best performing model in the target domain. For both datasets, CDAN was the method that showed the highest Oracle AUC, which can be attributed to the fact that the alignment takes into consideration the labels of the samples. When using a good model selection criterion, it is possible to achieve AUCs closer to the upper bound.

Although the Oracle AUC demonstrates the method's potential, it is not a reliable metric for real-world deployment. Therefore, using models selected based on source validation AUC, a more comprehensive analysis was performed. This evaluation includes the mean performance increase over the lower bound with the standard deviation, the number of folds in which the DA outperformed the lower bound, Cohen's d effect size, and the statistical significance evaluated using the Wilcoxon signed-rank test.

For the Sydney dataset (Table 5.9), GHM was able to achieve the best gain in Mean AUC (0.1117 ± 0.1386), when adapting the training set to the target domain style. This method was able to outperform the lower bound in 4 out of 5 folds, obtaining a high Cohen distance (>0.80), indicating that the DA yields a substantial performance improvement over the lower bound despite not achieving statistical difference (p-value > 0.05). All of the methods, except DANN, were able to outperform the lower bound, despite not achieving statistical relevance.

Regarding the Leuven dataset, CDAN was the method that obtained the most significant increase in performance (0.1577 ± 0.0672), almost achieving the upper bound, reducing the domain shift effect. This method outperformed the lower bound in all folds and, thereby, was able to achieve statistical significance and an high Cohen distance, reinforcing this improvement. All the other methods were also able to outperform the lower bound, except the Global FDA when adapting the test set to the source style and the Tent Method.

Although most methods did not achieve a statistically significant improvement, the Cohen's d value provides a measure of the effect size, indicating that several DA methods still led to substantial performance gains. Given the small sample size and the sensitivity of this statistical test to fold variance, p-values may underestimate the real advantages in implementing these methods. Consequently, interpreting other metrics such as effect size alongside the mean performance increase offers a more comprehensive understanding of the impact of each method in reducing these discrepancies between domains.

Table 5.9: Performance evaluation of DA methods on the Sydney test set. Results show Mean AUC, Cohen's d, and Win Rate (W/T) relative to the LB baseline. The values are reported as Mean $\pm$ Standard Deviation across 5-fold cross-validation. * indicates statistical significance with $p < 0.05$ using Wilcoxon signed-rank test. Best value for each metric in bold.

Method	Oracle AUC	Mean AUC	Diff (Mean $\pm$ STD)	W/T	Cohen D
Upper Bound	**0.8533 $\pm$ 0.1108**	**0.8167 $\pm$ 0.1333**	-	-	-
MMD	0.7775 $\pm$ 0.1695*	0.6258 $\pm$ 0.1636	0.0367 $\pm$ 0.1591	3/5	0.2305
CORAL	0.7733 $\pm$ 0.1367	0.6550 $\pm$ 0.1455	0.0659 $\pm$ 0.2672	3/5	0.2465
CMD	0.7008 $\pm$ 0.1143	0.6033 $\pm$ 0.1898	0.0142 $\pm$ 0.1136	3/5	0.1248
DANN	0.7384 $\pm$ 0.1424	0.5650 $\pm$ 0.1669	-0.0242 $\pm$ 0.2896	3/5	-0.0834
CDAN	**0.7842 $\pm$ 0.0258**	0.5950 $\pm$ 0.0935	0.0058 $\pm$ 0.2083	3/5	0.0280
GHM (Adapt Train)	0.7567 $\pm$ 0.1847	**0.7008 $\pm$ 0.1597**	**0.1117 $\pm$ 0.1386**	4/5	**0.8055**
GHM (Adapt Test)	-	0.6342 $\pm$ 0.1587	0.0450 $\pm$ 0.0912	4/5	0.4937
IHM (Adapt Train)	0.6983 $\pm$ 0.1158	0.6733 $\pm$ 0.1902	0.0667 $\pm$ 0.1509	4/5	0.4417
GFDA (Adapt Train)	0.6667 $\pm$ 0.2205	0.6050 $\pm$ 0.2351	0.0159 $\pm$ 0.0517	3/5	0.3067
GFDA (Adapt Test)	-	0.6050 $\pm$ 0.1457	0.0158 $\pm$ 0.1251	2/5	0.1266
IFDA (Adapt Train)	0.7125 $\pm$ 0.1831	0.6617 $\pm$ 0.1963	0.0725 $\pm$ 0.0933	3/5	0.7769
FixMatch	0.7142 $\pm$ 0.1694*	0.6425 $\pm$ 0.1964	0.0533 $\pm$ 0.0877	3/5	0.6083
Tent	-	0.6583 $\pm$ 0.2202	0.0692 $\pm$ 0.1284	2/5	0.5387
Lower Bound	**0.6492 $\pm$ 0.1894**	**0.5892 $\pm$ 0.2276**	-	-	-

Table 5.10: Performance evaluation of DA methods on the Leuven test set. Results show Mean AUC, Cohen's d, and Win Rate (W/T) relative to the LB baseline. The values are reported as Mean $\pm$ Standard Deviation across 5-fold cross-validation. * indicates statistical significance with $p < 0.05$ using Wilcoxon signed-rank test. Best value for each metric in bold.

Method	Oracle AUC	Mean AUC	Diff (Mean $\pm$ STD)	W/T	Cohen D
Upper Bound	**0.7833 $\pm$ 0.1387**	**0.7513 $\pm$ 0.1600**	-	-	-
MMD	0.7282 $\pm$ 0.1783	0.6115 $\pm$ 0.2407	0.0462 $\pm$ 0.1776	3/5	0.2599
CORAL	0.6897 $\pm$ 0.1493	0.5872 $\pm$ 0.1893	0.0218 $\pm$ 0.1220	3/5	0.1787
CMD	0.7333 $\pm$ 0.1788	0.5974 $\pm$ 0.1902	0.0321 $\pm$ 0.0821	2/5	0.3905
DANN	0.7231 $\pm$ 0.1294*	0.6423 $\pm$ 0.1443	0.0769 $\pm$ 0.1254	4/5	0.6136
CDAN	**0.7756 $\pm$ 0.1324***	**0.7231 $\pm$ 0.1142***	**0.1577 $\pm$ 0.0672**	**5/5**	**2.3464**
GHM (Adapt Train)	0.6744 $\pm$ 0.1239	0.6295 $\pm$ 0.1124	0.0641 $\pm$ 0.1852	3/5	0.3462
GHM (Adapt Test)	-	0.5846 $\pm$ 0.0928	0.0192 $\pm$ 0.0828	2/5	0.2322
IHM (Adapt Train)	0.7141 $\pm$ 0.1600	0.6321 $\pm$ 0.1751	0.0842 $\pm$ 0.1291	4/5	0.6520
GFDA (Adapt Train)	0.7077 $\pm$ 0.0712	0.6564 $\pm$ 0.0759	0.0910 $\pm$ 0.1563	3/5	0.5825
GFDA (Adapt Test)	-	0.5577 $\pm$ 0.1380	-0.0077 $\pm$ 0.0570	1/5	-0.1347
IFDA (Adapt Train)	0.6333 $\pm$ 0.1003	0.5756 $\pm$ 0.1194	0.0102 $\pm$ 0.1529	4/5	0.0670
FixMatch	0.6590 $\pm$ 0.1149	0.5885 $\pm$ 0.1371	0.0231 $\pm$ 0.1328	2/5	0.1738
Tent	-	0.5026 $\pm$ 0.0911	-0.0628 $\pm$ 0.0910	1/5	-0.6903
Lower Bound	**0.6474 $\pm$ 0.1060**	**0.5654 $\pm$ 0.1873**	-	-	-

Chapter 6

Conclusion

This dissertation demonstrates the inherent heterogeneity within the medical field, where variations in clinical protocols, patient demographics, and image acquisition equipment can cause a domain shift across different clinical settings and institutions. This leads to performance degradation when models are deployed on data that follows a different distribution, a problem that can be attenuated by incorporating DA techniques.

6.1 Contributions and Results

This project began by analysing two datasets extracted from distinct institutions: UZ Leuven and the Sydney Urodynamic Centres. The demographic data and the US equipment used to obtain the images were analysed and revealed significant differences in patient age, BMI, and parity, which influence image appearance. Furthermore, the Sydney dataset predominantly utilized more recent US hardware, resulting in higher image quality compared to the Leuven dataset. Additionally, the Sydney cohort, consisting primarily of symptomatic patients, presented a larger proportion of avulsion cases, which induced a prior shift between the two datasets.

After verifying the existence of a domain shift, experiments were conducted to analyse the feasibility of diagnosing LAM avulsion in these datasets. Models were trained and tested on each dataset using different architectures. The ConvNeXtV2-Atto achieved the best results and was, therefore, selected for the subsequent work.

An upper bound (model trained and tested on the same dataset) and a lower bound (model trained on a different dataset than the one on which it was tested) were obtained, testing the effect of using augmentations and the application of a segmentation mask. The combination of augmentations and the segmentation mask led to the best upper bound performance, achieving an AUC of 0.9233 ± 0.0751 and 0.8051 ± 0.1476 for the Sydney and Leuven datasets, respectively. However, due to the complexity of the segmentation model added in this configuration, which could fail on other datasets, the augmentation only configuration was used in the DA experiments, as it achieved a good upper bound (Sydney - 0.8167 ± 0.1333 and Leuven - 0.7513 ± 0.1600) and a significantly decreased lower bound (Sydney - 0.5892 ± 0.2276 and Leuven - 0.5654 ± 0.1566).

Several DA methods were implemented with the objective of increasing this lower bound performance, by using unlabeled samples from the target domain to reduce the domain shift. Various families of methods, each employing distinct mechanisms, were implemented and properly tuned to analyse their impact on performance.

Feature alignment methods based on using an explicit discrepancy metric as a regularization factor in the loss function were successfully implemented, leading to a better organization of both domains in the feature space. CORAL achieved a performance of 0.6550 ± 0.1455 on the Sydney dataset, and MMD was the most efficient method on the Leuven dataset, reaching a mean AUC of 0.6115 ± 0.2143. For the IDM methods, CDAN outperformed the DANN method as it aligns distributions based on the predicted labels, reducing the domain shift by using a moderate value of λ_{max} to ensure that the adversarial component is balanced with the label discriminator.

Image translation methods were able to align the visual appearance of both datasets. HM obtained a performance of 0.7008 ± 0.1597 (Sydney) and 0.6321 ± 0.1751 (Leuven), significantly reducing the performance gap. FDA also reduced the domain shift, though it proved highly dependent on the β factor, which led to image artifacts when not properly tuned.

Experiments implementing the Pseudo-Labeling method (FixMatch) showed that this approach performs best when using an intermediate confidence threshold (0.70) for target domain labels. It also benefits from a smaller weight in the loss function, as early pseudo-labels can introduce erroneous information into the model's training process.

Subsequently, the Test-Time Adaptation method (Tent), which aims to adjust a subset of the model's weights during inference, was investigated. This method only reduced the domain shift for the Sydney dataset and failed completely on the Leuven dataset, particularly when increasing the number of steps in the training loop. This occurred because this method is not effective when the lower-bound performance is already close to random. Additionally, using architectures that do not use batch normalization layers may result in a degradation in performance.

Finally, after identifying the best-performing method with parameters properly tuned for each family, a detailed comparison featuring statistical analysis was conducted, including the calculation of an Oracle AUC to evaluate the potential of each method in mitigating domain shift. For the Sydney dataset, GHM was the method that was able to achieve the highest performance. For the Leuven dataset, CDAN was the most effective method, yielding a statistical increase and reducing the domain gap by 0.1577 ± 0.0672. This method proved superior to the others, as it aligns the domains not only based on the global distribution but also by taking the labels into account. Oracle AUC also showed that CDAN was the method with the greatest potential in reducing these discrepancies, if the optimal stopping epoch can be determined.

6.2 Limitations and Future Work

Despite the fact that all methods were successfully implemented and capable of reducing domain shift, this work presented certain limitations and remains open to future improvements.

One of the main limitations of this work is the small size of the datasets. Since no public dataset was used, the acquisition and annotation of these private datasets were time-consuming tasks, requiring highly specialized clinical expertise and uniform label processing across institutions to ensure the datasets were compatible for the same application (both datasets were annotated with intact and complete avulsion cases). The small size of the datasets limited the type of DA methods implemented, since commonly used generative approaches were not tested, as they typically require large quantities of data.

Additionally, although all DA methods led to an increase in mean AUC, only one achieved statistical significance. This can also be attributed to the small size of the datasets, which caused high variance in performance across folds, ultimately resulting in non-significant improvements when applying statistical tests.

For future work, it would be interesting to apply these DA methods across more than two centers. Additionally, it would be valuable to investigate scenarios involving different types of domain shifts, such as adaptation from MRI to US (two modalities commonly used for LAM assessment). Moreover, this dissertation did not explore metrics for evaluating adaptation progress in the target domain during the training, which would have been useful for monitoring the model's performance on the validation set, facilitating model selection.

6.3 Final Remarks

This dissertation analysed the existence of domain shift across medical settings and its underlying implications when applying AI models in clinical environments. The two main objectives initially defined were successfully achieved.

The models exhibited strong performance in the classification of LAM avulsion, and the implemented DA methods demonstrated the potential to mitigate performance degradation caused by domain shift. Ultimately, this work provides a detailed analysis of the implementation of several methods, highlighting the advantages of obtaining models that perform effectively in a target domain without requiring new clinical annotations, which is typically a time-consuming process that requires experienced clinicians, hindering clinical workflows.

References

[1] UNESCO and International Centre for Engineering Education. *Engineering for sustainable development: delivering on the Sustainable Development Goals*. Paris: UNESCO, 2021. ISBN: 978-92-3-100437-7.

[2] Kun-Hsing Yu, Andrew L. Beam, and Isaac S. Kohane. "Artificial intelligence in healthcare". In: *Nature Biomedical Engineering* 2.10 (2018), pp. 719–731. DOI: 10.1038/s41551-018-0305-z.

[3] Kevin Pierre, Adam G. Haneberg, Sean Kwak, Keith R. Peters, Bruno Hochhegger, Thiparom Sananmuang, Padcha Tunlayadechanont, Patrick J. Tighe, Anthony Mancuso, and Reza Forghani. "Applications of Artificial Intelligence in the Radiology Roundtrip: Process Streamlining, Workflow Optimization, and Beyond". In: *Seminars in Roentgenology* 58.2 (2023). SI: AI in Radiology, pp. 158–169. ISSN: 0037-198X. DOI: https://doi.org/10.1053/j.ro.2023.02.003.

[4] B. Abhisheka, S. K. Biswas, B. Purkayastha, et al. "Recent trend in medical imaging modalities and their applications in disease diagnosis: a review". In: *Multimedia Tools and Applications* 83.15 (2024), pp. 43035–43070. DOI: 10.1007/s11042-023-17326-1.

[5] Xin Luo, Zhaoguo Zhang, Jinlong Chen, and Chang Liu. "From Convolution to Attention: A Historical Review and Comparative Analysis of Deep Learning Models in Image Classification". In: *2025 6th International Conference on Electronic Communication and Artificial Intelligence (ICECAI)*. 2025, pp. 300–308. DOI: 10.1109/ICECAI66283.2025.11171030.

[6] Luciano M. Prevedello, Safwan S. Halabi, George Shih, Carol C. Wu, Marc D. Kohli, Falgun H. Chokshi, Bradley J. Erickson, Jayashree Kalpathy-Cramer, Katherine P. Andriole, and Adam E. Flanders. "Challenges Related to Artificial Intelligence Research in Medical Imaging and the Importance of Image Analysis Competitions". In: *Radiology: Artificial Intelligence* 1.1 (2019), e180031. DOI: 10.1148/ryai.2019180031.

[7] Jee Seok Yoon, Kwanseok Oh, Maciej A. Mazurowski, Yooseung Shin, and Heung-Il Suk. "Domain Generalization for Medical Image Analysis: A Survey". In: (2023). DOI: 10.13140/RG.2.2.36040.49929.

[8] Hao Guan and Mingxia Liu. "Domain Adaptation for Medical Image Analysis: A Survey". In: *IEEE Transactions on Biomedical Engineering* 69 (Mar. 2022), pp. 1173–1185. DOI: 10.1109/TBME.2021.3117407.

[9] Hans Peter Dietz. *Pelvic floor ultrasound: a review*. Apr. 2010. DOI: 10.1016/j.ajog.2009.08.018.

[10] H. P. Dietz. "Clinical consequences of levator trauma". In: *Ultrasound in Obstetrics & Gynecology* 39.4 (2012), pp. 367–371. DOI: https://doi.org/10.1002/uog.11141.

[11] G. A. Van Veelen, K. J. Schweitzer, and C. H. Van Der Vaart. "Reliability of pelvic floor measurements on three- and four-dimensional ultrasound during and after first pregnancy: Implications for training". In: *Ultrasound in Obstetrics and Gynecology* 42 (5 Nov. 2013), pp. 590–595. ISSN: 09607692. DOI: 10.1002/uog.12523.

[12] Rocío Adriana Peinado-Molina, Antonio Hernández-Martínez, Sergio Martínez-Vázquez, Julián Rodríguez-Almagro, and Juan Miguel Martínez-Galiano. "Pelvic floor dysfunction: prevalence and associated factors". In: *BMC Public Health* 23 (1 Dec. 2023). ISSN: 14712458. DOI: 10.1186/s12889-023-16901-3.

[13] H. P. Dietz, F. Moegni, and K. L. Shek. "Diagnosis of levator avulsion injury: A comparison of three methods". In: *Ultrasound in Obstetrics and Gynecology* 40 (6 Dec. 2012), pp. 693–698. ISSN: 09607692. DOI: 10.1002/uog.11190.

[14] A. Samešová, B. Packet, L. Cattani, et al. "Image-quality assessment of the levator ani immediately after delivery". In: *International Urogynecology Journal* 37 (2026), pp. 1347–1356. DOI: 10.1007/s00192-025-06448-9.

[15] Sabina Tim and Agnieszka I. Mazur-Bialy. *The Most Common Functional Disorders and Factors Affecting Female Pelvic Floor*. Dec. 2021. DOI: 10.3390/life11121397.

[16] J.O.L. DeLancey. "Chapter Two - Pelvic Floor Anatomy and Pathology". In: *Biomechanics of the Female Pelvic Floor*. Ed. by Lennox Hoyte and Margot Damaser. Academic Press, 2016, pp. 13–51. ISBN: 978-0-12-803228-2. DOI: https://doi.org/10.1016/B978-0-12-803228-2.00002-7.

[17] Elizabeth A. Doxford-Hook, Elizabeth Slemeck, Candice L. Downey, and Fiona A. Marsh. *Management of levator ani avulsion: a systematic review and narrative synthesis*. Nov. 2023. DOI: 10.1007/s00404-023-06955-4.

[18] Nikhil Sindhwani, Christian Bamberg, Nele Famaey, Geertje Callewaert, Joachim W. Dudenhausen, Ulf Teichgräber, and Jan Deprest. "In vivo evidence of significant levator ani muscle stretch on MR images of a live childbirth". In: *American Journal of Obstetrics and Gynecology* 217.2 (2017), 194.e1–194.e8. DOI: 10.1016/j.ajog.2017.04.014.

[19] H. P. Dietz, A. Abbu, and K. L. Shek. "The levator-urethra gap measurement: A more objective means of determining levator avulsion?" In: *Ultrasound in Obstetrics and Gynecology* 32 (7 Dec. 2008), pp. 941–945. ISSN: 09607692. DOI: https://doi.org/10.1002/uog.6268.

[20] Karin Estendahl and Lotta Danielsson. "Women's experiences of life and healthcare after levator ani avulsion: a qualitative interview study". In: *BMC Women's Health* 25.1 (2025), p. 336. DOI: 10.1186/s12905-025-03892-z.

[21] HP Dietz and JM Simpson. "Levator trauma is associated with pelvic organ prolapse". In: *BJOG: An International Journal of Obstetrics & Gynaecology* 115.8 (2008), pp. 979–984. DOI: https://doi.org/10.1111/j.1471-0528.2008.01751.x.

[22] Hans Peter Dietz and Cheung Shek. "Levator avulsion and grading of pelvic floor muscle strength". In: *International Urogynecology Journal* 19.5 (2008), pp. 633–636. DOI: 10.1007/s00192-007-0491-9.

[23] Hans Peter Dietz. "Pelvic organ prolapse - a review". In: *Australian Family Physician* 44.7 (July 2015). PMID: 26590487, pp. 446–452.

[24] Zeelha Abdool, Ka Lai Shek, and Hans Peter Dietz. "The effect of levator avulsion on hiatal dimension and function". In: *American Journal of Obstetrics and Gynecology* 201 (1 2009), 89.e1–89.e5. ISSN: 00029378. DOI: 10.1016/j.ajog.2009.02.005.

[25] K. Van Delft, S. A. Shobeiri, R. Thakar, N. Schwertner-Tiepelmann, and A. H. Sultan. "Intra- and interobserver reliability of levator ani muscle biometry and avulsion using three-dimensional endovaginal ultrasonography". In: *Ultrasound in Obstetrics and Gynecology* 43 (2 Feb. 2014), pp. 202–209. ISSN: 09607692. DOI: 10.1002/uog.13193.

[26] Kim W.M. van Delft, Abdul H. Sultan, Ranee Thakar, S. Abbas Shobeiri, and Kirsten B. Kluivers. "Agreement between palpation and transperineal and endovaginal ultrasound in the diagnosis of levator ani avulsion". In: *International Urogynecology Journal* 26 (1 Jan. 2015), pp. 33–39. ISSN: 14333023. DOI: 10.1007/s00192-014-2426-6.

[27] National Cancer Institute. *NCI Dictionary of Cancer Terms: Transvaginal Ultrasound*. Accessed on 18/05/2026. URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/transvaginal-ultrasound.

[28] Kourosh Kalayeh, J. Brian Fowlkes, Bryan S. Sack, Jennifer LaCross, Stephanie Daignault-Newton, Payton Schmidt, Haowei Tai, William W. Schultz, James A. Ashton-Miller, and John O. DeLancey. "A New Automated Ultrasound Quantification of Urethral Mobility for Stress Urinary Incontinence: A Feasibility Study". In: *Journal of Ultrasound in Medicine* 44 (7 July 2025), pp. 1213–1227. ISSN: 15509613. DOI: 10.1002/jum.16676.

[29] Hans Dietz. "Pelvic floor muscle trauma". In: *Expert Review of Obstretrics and Gynecology* 5 (July 2010), pp. 479–492. DOI: 10.1586/eog.10.28.

[30] Hans Peter Dietz. "Axial Plane Imaging". In: *Atlas of Pelvic Floor Ultrasound*. London: Springer London, 2008, pp. 76–90. ISBN: 978-1-84628-584-4. DOI: 10.1007/978-1-84628-584-4_6.

[31] H. P. Dietz, K. L. Shek, and G. K. Low. "All or nothing? A second look at partial levator avulsion". In: *Ultrasound in Obstetrics & Gynecology* 60.5 (2022), pp. 693–697. DOI: https://doi.org/10.1002/uog.26034.

[32] Hans Peter Dietz, Maria Jose Bernardo, Adrienne Kirby, and Ka Lai Shek. "Minimal criteria for the diagnosis of avulsion of the puborectalis muscle by tomographic ultrasound". In: *International Urogynecology Journal* 22 (6 2011), pp. 699–704. ISSN: 14333023. DOI: 10.1007/s00192-010-1329-4.

[33] Yan Xu, Rixiang Quan, Weiting Xu, Yi Huang, Xiaolong Chen, and Fengyuan Liu. "Advances in Medical Image Segmentation: A Comprehensive Review of Traditional, Deep Learning and Hybrid Approaches". In: *Bioengineering* 11 (Oct. 2024). DOI: 10.3390/bioengineering11101034.

[34] Elena Escobar-Linero, Francisco Luna-Perejón, Luis Muñoz-Saavedra, José Luis Sevillano, and Manuel Domínguez-Morales. "On the Feature Extraction Process in Machine Learning: An Experimental Study About Guided Versus Non-Guided Process in Falling Detection Systems". In: *Engineering Applications of Artificial Intelligence* 114 (Sept. 2022). DOI: 10.1016/j.engappai.2022.105170.

[35] Vladimir Nasteski. "An Overview of the Supervised Machine Learning Methods". In: *HORIZONS.B* 4 (Dec. 2017), pp. 51–62. DOI: 10.20544/horizons.b.04.1.17.p05.

[36] Samreen Naeem, Aqib Ali, Sania Anam, and Muhammad Munawar Ahmed. "Unsupervised Machine Learning Algorithms: Comprehensive Review". In: *International Journal of Computing and Digital Systems* 13 (2023), pp. 911–921. DOI: 10.12785/ijcds/130172.

[37] Igor Kononenko. “Machine Learning for Medical Diagnosis: History, State of the Art and Perspective”. In: *Artificial Intelligence in Medicine* 23.1 (2001), pp. 89–109. DOI: 10.1016/s0933-3657(01)00077-x.

[38] Abeer Aljuaid and Mohd Anwar. “Survey of Supervised Learning for Medical Image Processing”. In: *SN Computer Science* 3 (July 2022). DOI: 10.1007/s42979-022-01166-1.

[39] Yann Lecun, Bernhard Boser, John Denker, Don Henderson, R. Howard, W. Hubbard, and Larry Jackel. “Backpropagation Applied to Handwritten Zip Code Recognition”. In: *Neural Computation* 1 (Dec. 1989), pp. 541–551. DOI: 10.1162/neco.1989.1.4.541.

[40] K. O’Shea and R. Nash. *An Introduction to Convolutional Neural Networks*. arXiv preprint. Nov. 2015. DOI: 10.48550/arXiv.1511.08458.

[41] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. “A Survey on Deep Learning in Medical Image Analysis”. In: *Medical Image Analysis* 42 (Dec. 2017), pp. 60–88. DOI: 10.1016/j.media.2017.07.005.

[42] Koon Meng Ang, El Sayed M. El-kenawy, Abdelaziz A. Abdelhamid, Abdelhameed Ibrahim, Amal H. Alharbi, Doaa Sami Khafaga, Sew Sun Tiang, and Wei Hong Lim. “Optimal Design of Convolutional Neural Network Architectures Using Teaching–Learning-Based Optimization for Image Classification”. In: *Symmetry* 14 (Nov. 2022). DOI: 10.3390/sym14112323.

[43] 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.

[44] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. “Densely Connected Convolutional Networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 2261–2269. DOI: 10.1109/CVPR.2017.243.

[45] Mingxing Tan and Quoc Le. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”. In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 6105–6114. DOI: 10.48550/arXiv.1905.11946.

[46] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. “An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *International Conference on Learning Representations (ICLR)*. 2021. DOI: 10.48550/arXiv.2010.11929.

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. “Attention Is All You Need”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 30. Curran Associates, Inc., 2017. DOI: 10.48550/arXiv.1706.03762.

[48] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. "A ConvNet for the 2020s". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 11966–11976. DOI: 10.1109/CVPR52688.2022.01167.

[49] Yuxin Wu and Justin Johnson. *Rethinking "Batch" in BatchNorm*. arXiv preprint. May 2021. DOI: 10.48550/arXiv.2105.07576.

[50] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. "ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 16133–16142. DOI: 10.1109/CVPR52729.2023.01548.

[51] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. "Masked Autoencoders Are Scalable Vision Learners". In: (2022), pp. 15979–15988. DOI: 10.1109/CVPR52688.2022.01553.

[52] Candelaria Mosquera, Luciana Ferrer, Diego H. Milone, Daniel Luna, and Enzo Ferrante. "Class imbalance on medical image classification: towards better evaluation practices for discrimination and calibration performance". In: *European Radiology* 34 (Dec. 2024), pp. 7895–7903. ISSN: 14321084. DOI: 10.1007/s00330-024-10834-0.

[53] Hao Guan, Pew Thian Yap, Andrea Bozoki, and Mingxia Liu. *Federated learning for medical image analysis: A survey*. July 2024. DOI: 10.1016/j.patcog.2024.110424.

[54] Rongguang Wang, Pratik Chaudhari, and Christos Davatzikos. "Embracing the disharmony in medical imaging: A Simple and effective framework for domain adaptation". In: *Medical Image Analysis* 76 (Feb. 2022). ISSN: 13618423. DOI: 10.1016/j.media.2021.102309.

[55] Elsevier. *Scopus: Abstract and citation database*. Online. [Online]. Accessed: 2025-09-22. 2025. URL: https://www.scopus.com/.

[56] Suruchi Kumari and Pravendra Singh. "Deep Learning for Unsupervised Domain Adaptation in Medical Imaging: Recent Advancements and Future Perspectives". In: (July 2023). DOI: https://doi.org/10.48550/arXiv.2308.01265.

[57] Sicheng Zhao, Xiangyu Yue, Shanghang Zhang, Bo Li, Han Zhao, Bichen Wu, Ravi Krishna, Joseph E. Gonzalez, Alberto L. Sangiovanni-Vincentelli, Sanjit A. Seshia, and Kurt Keutzer. "A Review of Single-Source Deep Unsupervised Visual Domain Adaptation". In: *IEEE Transactions on Neural Networks and Learning Systems* 33 (Feb. 2022), pp. 473–493. DOI: 10.1109/TNNLS.2020.3028503.

[58] Chenhao Pei, Fuping Wu, Mingjing Yang, Lin Pan, Wangbin Ding, Jinwei Dong, Liqin Huang, and Xiahai Zhuang. "Multi-Source Domain Adaptation for Medical Image Segmentation". In: *IEEE Transactions on Medical Imaging* 43 (Apr. 2024), pp. 1640–1651. DOI: 10.1109/TMI.2023.3346285.

[59] Fares Bougourzi, Feryal Windal Moula, Halim Benhabiles, Fadi Dornaika, and Abdelmalik Taleb-Ahmed. "Ensembling and Test Augmentation for COVID-19 Detection and COVID-19 Domain Adaptation from 3D CT-Scans". In: (Mar. 2024). DOI: https://doi.org/10.48550/arXiv.2403.11338.

[60] Lin Lawrence Guo, Stephen R. Pfohl, Jason Fries, Alistair E.W. Johnson, Jose Posada, Catherine Aftandilian, Nigam Shah, and Lillian Sung. "Evaluation of Domain Generalization and Adaptation on Improving Model Robustness to Temporal Dataset Shift in Clinical Medicine". In: *Scientific Reports* 12 (Dec. 2022). DOI: 10.1038/s41598-022-06484-1.

[61] Lipeng Ning, Elisenda Bonet-Carne, Francesco Grussu, Farshid Sepehrband, Enrico Kaden, Jelle Veraart, Stefano B. Blumberg, Can Son Khoo, Marco Palombo, Iasonas Kokkinos, Daniel C. Alexander, Jaume Coll-Font, Benoit Scherrer, Simon K. Warfield, Suheyla Cetin Karayumak, Yogesh Rathi, Simon Koppers, Leon Weninger, Julia Ebert, Dorit Merhof, Daniel Moyer, Maximilian Pietsch, Daan Christiaens, Rui Azeredo Gomes Teixeira, Jacques Donald Tournier, Kurt G. Schilling, Yuankai Huo, Vishwesh Nath, Colin Hansen, Justin Blaber, Bennett A. Landman, Andrey Zhylka, Josien P.W. Pluim, Greg Parker, Umesh Rudrapatna, John Evans, Cyril Charron, Derek K. Jones, and Chantal M.W. Tax. "Cross-Scanner and Cross-Protocol Multi-Shell Diffusion MRI Data Harmonization: Algorithms and Results". In: *NeuroImage* 221 (Nov. 2020). DOI: 10.1016/j.neuroimage.2020.117128.

[62] Zhizhe Liu, Zhenfeng Zhu, Shuai Zheng, Yang Liu, Jiayu Zhou, and Yao Zhao. "Margin Preserving Self-Paced Contrastive Learning Towards Domain Adaptation for Medical Image Segmentation". In: *IEEE Journal of Biomedical and Health Informatics* 26 (Feb. 2022), pp. 638–647. DOI: 10.1109/JBHI.2022.3140853.

[63] Chenglin Yu and Hailong Pei. "Dynamic Weighting Translation Transfer Learning for Imbalanced Medical Image Classification". In: *Entropy* 26 (May 2024). DOI: 10.3390/e26050400.

[64] Yijun Dong, Yuege Xie, and Rachel Ward. "Adaptively Weighted Data Augmentation Consistency Regularization for Robust Optimization Under Concept Shift". In: (Jan. 2023). DOI: https://doi.org/10.48550/arXiv.2210.01891.

[65] Mingyang Cai, Thomas Klausch, and Mark A. van de Wiel. "A Semi-Supervised CART Model for Covariate Shift". In: (Dec. 2024). DOI: https://doi.org/10.48550/arXiv.2410.20978.

[66] Thomas Ranvier, Haytham Elghazel, Emmanuel Coquery, and Khalid Benabdeslem. *Deep Multi-Source Supervised Domain Adaptation with Class Imbalance*. July 2023. DOI: 10.21203/rs.3.rs-3160713/v1.

[67] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. "Universal Domain Adaptation". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 2715–2724. DOI: 10.1109/CVPR.2019.00283.

[68] Mohammad Areeb Qazi, Anees Ur Rehman Hashmi, Santosh Sanjeev, Ibrahim Almakky, Numan Saeed, Camila Gonzalez, and Mohammad Yaqub. "Continual Learning in Medical Imaging: A Survey and Practical Analysis". In: (Oct. 2024). DOI: https://doi.org/10.48550/arXiv.2405.13482.

[69] Zhiqi Yu, Jingjing Li, Zhekai Du, Lei Zhu, and Heng Tao Shen. "A Comprehensive Survey on Source-free Domain Adaptation". In: (Feb. 2023). DOI: https://doi.org/10.48550/arXiv.2302.11803.

[70] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. "How transferable are features in deep neural networks?" In: (Nov. 2014). DOI: https://doi.org/10.48550/arXiv.1411.1792.

[71] Ibrahem Kandel and Mauro Castelli. "How deeply to fine-tune a convolutional neural network: A case study using a histopathology dataset". In: *Applied Sciences (Switzerland)* 10 (10 May 2020). ISSN: 20763417. DOI: 10.3390/APP10103359.

[72] Roger Bermudez-Chacon, Pablo Marquez-Neila, Mathieu Salzmann, and Pascal Fua. "A domain-adaptive two-stream U-Net for electron microscopy image segmentation". In: *Proceedings - International Symposium on Biomedical Imaging*. Vol. 2018-April. IEEE Computer Society, May 2018, pp. 400–404. ISBN: 9781538636367. DOI: 10.1109/ISBI.2018.8363602.

[73] Weijian Deng, Liang Zheng, Yifan Sun, and Jianbin Jiao. "Rethinking Triplet Loss for Domain Adaptation". In: *IEEE Transactions on Circuits and Systems for Video Technology* 31 (1 Jan. 2021), pp. 29–37. ISSN: 15582205. DOI: 10.1109/TCSVT.2020.2968484.

[74] Jichang Li, Guanbin Li, Yemin Shi, and Yizhou Yu. *Cross-Domain Adaptive Clustering for Semi-Supervised Domain Adaptation*. 2021. DOI: https://doi.org/10.48550/arXiv.2104.09415.

[75] Zizheng Yan, Yushuang Wu, Guanbin Li, Yipeng Qin, Xiaoguang Han, and Shuguang Cui. *Multi-level Consistency Learning for Semi-supervised Domain Adaptation*. 2022. DOI: https://doi.org/10.48550/arXiv.2205.04066.

[76] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. "Attention-based Deep Multiple Instance Learning". In: (June 2018). DOI: https://doi.org/10.48550/arXiv.1802.04712.

[77] Zhilu Zhang and Mert R. Sabuncu. "Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels". In: (Nov. 2018). DOI: https://doi.org/10.48550/arXiv.1805.07836.

[78] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. "Learning Transferable Features with Deep Adaptation Networks". In: (May 2015). DOI: https://doi.org/10.48550/arXiv.1502.02791.

[79] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Alexander Smola, Bernhard Schölkopf, and Alexander Smola GRETTON. *A Kernel Two-Sample Test Bernhard Schölkopf*. Tech. rep. 2012, pp. 723–773. DOI: https://doi.org/10.48550/arXiv.0805.2368.

[80] Baochen Sun, Jiashi Feng, and Kate Saenko. "Correlation Alignment for Unsupervised Domain Adaptation". In: (Dec. 2016). DOI: https://doi.org/10.48550/arXiv.1612.01939.

[81] Werner Zellinger, Edwin Lughofer, Susanne Saminger-Platz, Thomas Grubinger, and Thomas Natschläger. "Central Moment Discrepancy (CMD) for Domain-Invariant Representation Learning". In: (2017). DOI: https://doi.org/10.48550/arXiv.1702.08811.

[82] Nicolas Courty, Remi Flamary, Devis Tuia, and Alain Rakotomamonjy. "Optimal Transport for Domain Adaptation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (9 Sept. 2017), pp. 1853–1865. ISSN: 01628828. DOI: 10.1109/TPAMI.2016.2615921.

[83] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. "Domain-Adversarial Training of Neural Networks". In: (May 2016). DOI: https://doi.org/10.48550/arXiv.1505.07818.

[84] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. "Adversarial Discriminative Domain Adaptation". In: (Feb. 2017). DOI: https://doi.org/10.48550/arXiv.1702.05464.

[85] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. "Conditional Adversarial Domain Adaptation". In: (Dec. 2018). DOI: https://doi.org/10.48550/arXiv.1705.10667.

[86] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. "A Survey on Contrastive Self-supervised Learning". In: (Feb. 2021). DOI: https://doi.org/10.48550/arXiv.2011.00362.

[87] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A Simple Framework for Contrastive Learning of Visual Representations". In: (July 2020). DOI: https://doi.org/10.48550/arXiv.2002.05709.

[88] Zheren Li, Zhiming Cui, Lichi Zhang, Sheng Wang, Chenjin Lei, Xi Ouyang, Dongdong Chen, Xiangyu Zhao, Chunling Liu, Zaiyi Liu, Yajia Gu, Dinggang Shen, and Jie Zhi Cheng. "Domain generalization for mammographic image analysis with contrastive learning". In: *Computers in Biology and Medicine* 185 (Feb. 2025). ISSN: 18790534. DOI: 10.1016/j.compbiomed.2024.109455.

[89] Jun Ma. "Histogram Matching Augmentation for Domain Adaptation with Application to Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Image Segmentation". In: (Dec. 2020). DOI: https://doi.org/10.48550/arXiv.2012.13871.

[90] Yanchao Yang and Stefano Soatto. "FDA: Fourier Domain Adaptation for Semantic Segmentation". In: (Apr. 2020). DOI: https://doi.org/10.48550/arXiv.2004.05498.

[91] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Networks". In: (June 2014). DOI: https://doi.org/10.48550/arXiv.1406.2661.

[92] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks". In: (Aug. 2020). DOI: https://doi.org/10.48550/arXiv.1703.10593.

[93] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. "DualGAN: Unsupervised Dual Learning for Image-to-Image Translation". In: (Oct. 2018). DOI: https://doi.org/10.48550/arXiv.1704.02510.

[94] Jelmer M. Wolterink, Anna M. Dinkla, Mark H. F. Savenije, Peter R. Seevinck, Cornelis A. T. van den Berg, and Ivana Isgum. "Deep MR to CT Synthesis using Unpaired Data". In: (Aug. 2017). DOI: https://doi.org/10.48550/arXiv.1708.01155.

[95] Devavrat Tomar, Manana Lortkipanidze, Guillaume Vray, Behzad Bozorgtabar, and Jean Philippe Thiran. "Self-Attentive Spatial Adaptive Normalization for Cross-Modality Domain Adaptation". In: *IEEE Transactions on Medical Imaging* 40 (10 Oct. 2021), pp. 2926–2938. ISSN: 1558254X. DOI: 10.1109/TMI.2021.3059265.

[96] Eduardo Diniz, Tales Santini, Helmet Karim, Howard J. Aizenstein, and Tamer S. Ibrahim. "Cross-Modality Image Translation of 3 Tesla Magnetic Resonance Imaging to 7 Tesla Using Generative Adversarial Networks". In: *Human Brain Mapping* 46 (9 June 2025). ISSN: 10970193. DOI: 10.1002/hbm.70246.

[97] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. "CyCADA: Cycle-Consistent Adversarial Domain Adaptation". In: (Dec. 2017). DOI: https://doi.org/10.48550/arXiv.1711.03213.

[98] Xiao Liu, Pedro Sanchez, Spyridon Thermos, Alison Q. O'Neil, and Sotirios A. Tsaftaris. *Learning disentangled representations in the imaging domain*. Aug. 2022. DOI: 10.1016/j.media.2022.102516.

[99] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. "Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: (2017).

[100] Shengjia Zhao, Jiaming Song, and Stefano Ermon. "InfoVAE: Balancing Learning and Inference in Variational Autoencoders". In: *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*. 2019, pp. 8989–8996. DOI: https://doi.org/10.48550/arXiv.1706.02262.

[101] Junlin Yang, Nicha C. Dvornek, Fan Zhang, Julius Chapiro, MingDe Lin, and James S. Duncan. "Unsupervised Domain Adaptation via Disentangled Representations: Application to Cross-Modality Liver Segmentation". In: (Aug. 2019). DOI: https://doi.org/10.48550/arXiv.1907.13590.

[102] Agisilaos Chartsias, Thomas Joyce, Giorgos Papanastasiou, Scott Semple, Michelle Williams, David E. Newby, Rohan Dharmakumar, and Sotirios A. Tsaftaris. "Disentangled representation learning in cardiac image analysis". In: *Medical Image Analysis* 58 (Dec. 2019). ISSN: 13618423. DOI: 10.1016/j.media.2019.101535.

[103] Ran Gu, Guotai Wang, Jiangshan Lu, Jingyang Zhang, Wenhui Lei, Yinan Chen, Wenjun Liao, Shichuan Zhang, Kang Li, Dimitris N. Metaxas, and Shaoting Zhang. "CDDSA: Contrastive domain disentanglement and style augmentation for generalizable medical image segmentation". In: *Medical Image Analysis* 89 (Oct. 2023). ISSN: 13618423. DOI: 10.1016/j.media.2023.102904.

[104] Dong-Hyun Lee. "Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks". In: *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)* (July 2013).

[105] Yundong Li, Longxia Guo, and Yizheng Ge. *Pseudo Labels for Unsupervised Domain Adaptation: A Review*. Aug. 2023. DOI: 10.3390/electronics12153325.

[106] Samuli Laine and Timo Aila. "Temporal Ensembling for Semi-Supervised Learning". In: (Mar. 2017). DOI: https://doi.org/10.48550/arXiv.1610.02242.

[107] Antti Tarvainen and Harri Valpola. "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results". In: (Apr. 2018). DOI: https://doi.org/10.48550/arXiv.1703.01780.

[108] Boliang Li, Yaming Xu, Yan Wang, Luxiu Li, and Bo Zhang. "The student-teacher framework guided by self-training and consistency regularization for semi-supervised medical image segmentation". In: *PLoS ONE* 19 (4 April Apr. 2024). ISSN: 19326203. DOI: 10.1371/journal.pone.0300039.

[109] Christian S. Perone, Pedro Ballester, Rodrigo C. Barros, and Julien Cohen-Adad. "Unsupervised domain adaptation for medical imaging segmentation with self-ensembling". In: *NeuroImage* 194 (July 2019), pp. 1–11. ISSN: 10959572. DOI: 10.1016/j.neuroimage.2019.03.026.

[110] David Berthelot, Rebecca Roelofs, Kihyuk Sohn, Nicholas Carlini, and Alex Kurakin. "AdaMatch: A Unified Approach to Semi-Supervised Learning and Domain Adaptation". In: (Mar. 2022). DOI: https://doi.org/10.48550/arXiv.2106.04732.

[111] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. *FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence*. Tech. rep. Google Research, 2020. DOI: https://doi.org/10.48550/arXiv.2001.07685.

[112] Chen Li, Wei Chen, Xin Luo, Yulin He, and Yusong Tan. "ADAPTIVE PSEUDO LABELING FOR SOURCE-FREE DOMAIN ADAPTATION IN MEDICAL IMAGE SEGMENTATION". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Vol. 2022-May. Institute of Electrical and Electronics Engineers Inc., 2022, pp. 1091–1095. ISBN: 9781665405409. DOI: 10.1109/ICASSP43922.2022.9746286.

[113] Jianghao Wu, Guotai Wang, Ran Gu, Tao Lu, Yinan Chen, Wentao Zhu, Tom Vercauteren, Sebastien Ourselin, and Shaoting Zhang. "UPL-SFDA: Uncertainty-Aware Pseudo Label Guided Source-Free Domain Adaptation for Medical Image Segmentation". In: *IEEE Transactions on Medical Imaging* 42 (12 Dec. 2023), pp. 3932–3943. ISSN: 1558254X. DOI: 10.1109/TMI.2023.3318364.

[114] Yundong Li, Longxia Guo, and Yizheng Ge. *Pseudo Labels for Unsupervised Domain Adaptation: A Review*. Aug. 2023. DOI: 10.3390/electronics12153325.

[115] Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. "Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning". In: (June 2020). DOI: https://doi.org/10.48550/arXiv.1908.02983.

[116] Jian Liang, Ran He, and Tieniu Tan. "A Comprehensive Survey on Test-Time Adaptation under Distribution Shifts". In: (Dec. 2024). DOI: https://doi.org/10.48550/arXiv.2303.15361.

[117] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. "Revisiting Batch Normalization For Practical Domain Adaptation". In: (Nov. 2016). DOI: https://doi.org/10.48550/arXiv.1603.04779.

[118] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. "Tent: Fully Test-time Adaptation by Entropy Minimization". In: (Mar. 2021). DOI: https://doi.org/10.48550/arXiv.2006.10726.

[119] Shuhao Fu, Yongyi Lu, Yan Wang, Yuyin Zhou, Wei Shen, Elliot Fishman, and Alan Yuille. "Domain Adaptive Relational Reasoning for 3D Multi-Organ Segmentation". In: (July 2020). DOI: https://doi.org/10.48550/arXiv.2005.09120.

[120] Haoyu Dong, Nicholas Konz, Hanxue Gu, and Maciej A Mazurowski. *Medical Image Segmentation with InTEnt: Integrated Entropy Weighting for Single Image Test-Time Adaptation*. Tech. rep. 2024. DOI: https://doi.org/10.48550/arXiv.2402.09604.

[121] Jiayi Zhu, Bart Bolsterlee, Yang Song, and Erik Meijering. "Improving cross-domain generalizability of medical image segmentation using uncertainty and shape-aware continual test-time domain adaptation". In: *Medical Image Analysis* 101 (Apr. 2025). ISSN: 13618423. DOI: 10.1016/j.media.2024.103422.

[122] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. *Efficient Test-Time Model Adaptation without Forgetting*. Tech. rep. 2022. DOI: https://doi.org/10.48550/arXiv.2204.02610.

[123] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. "Continual Test-Time Domain Adaptation". In: (Mar. 2022). DOI: https://doi.org/10.48550/arXiv.2203.13591.

[124] JASP Team. *JASP (Version 0.95.4)[Computer software]*. 2025. URL: https://jasp-stats.org/.

[125] Ross Wightman. *PyTorch Image Models*. https://github.com/huggingface/pytorch-image-models. 2019.

[126] M. Jorge Cardoso, Wenqi Li, Richard Brown, and et al. "MONAI: An open-source framework for deep learning in healthcare". In: (2022). DOI: https://doi.org/10.48550/arXiv.2211.02701.

[127] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[128] Kuniaki Saito, Donghyun Kim, Piotr Teterwak, Stan Sclaroff, Trevor Darrell, and Kate Saenko. *Tune it the Right Way: Unsupervised Validation of Domain Adaptation via Soft Neighborhood Density*. 2021. DOI: https://doi.org/10.48550/arXiv.2108.10860.

[129] Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. *Large sample analysis of the median heuristic*. 2018. DOI: https://doi.org/10.48550/arXiv.1707.07269.

[130] Andrew P. Bradley. "The use of the area under the ROC curve in the evaluation of machine learning algorithms". In: *Pattern Recognition* 30.7 (1997), pp. 1145–1159. ISSN: 0031-3203. DOI: https://doi.org/10.1016/S0031-3203(96)00142-2.

[131] Laurens van der Maaten and Geoffrey Hinton. "Viualizing data using t-SNE". In: *Journal of Machine Learning Research* 9 (Nov. 2008), pp. 2579–2605.

[132] Frank Wilcoxon. "Individual comparisons by ranking methods". In: *Biometrics Bulletin* 1.6 (1945), pp. 80–83. DOI: 10.2307/3001968.

[133] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. 2nd. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.

[134] Gail M. Sullivan and Richard Feinn. "Using Effect Size—Or Why the P Value Is Not Enough". In: *Journal of Graduate Medical Education* 4.3 (2012), pp. 279–282. DOI: 10.4300/JGME-D-12-00156.1.