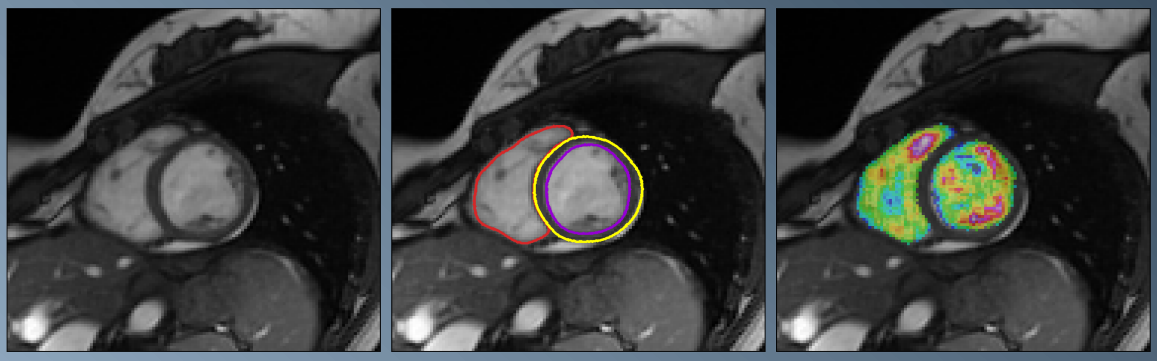
UNIVERSITAT DE BARCELONA

Generalizability in multi-centre cardiac image analysis with machine learning

Víctor Manuel Campello Román

Generalizability in multi-centre cardiac image analysis with machine learning

Víctor Manuel Campello Román



Universitat de Barcelona

European Commission

UNIVERSITAT DE BARCELONA

DOCTORAL THESIS

Generalizability in multi-centre cardiac image analysis with machine learning

Author:
Víctor Manuel CAMPELLO ROMÁN

Supervisors:
Dr. Karim LEKADIR
Dr. Santi SEGUÍ MESQUIDA

Tutor:
Dr. Álex HARO PROVINCIALE

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy in Mathematics and Informatics

in the

Departament de Matemàtiques i Informàtica

September 29, 2023

UNIVERSITAT DE BARCELONA

"This ability to generalize is the key to learning. What would be the point of a machine that could recognize a picture that it has already seen, or win a game of Go that it has already played? Obviously, the real aim is to recognize any picture, or to win against any player [...]."

Stephané Dehaene. *How we learn: The new science of education and the brain*. Penguin UK, 2020.

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica
Departament de Matemàtiques i Informàtica

Doctor of Philosophy in Mathematics and Informatics

Generalizability in multi-centre cardiac image analysis with machine learning

by Víctor Manuel CAMPELLO ROMÁN

The field of Artificial Intelligence (AI) has undergone a revolution in recent years with the advent of more efficient computing hardware and well-documented software for model development. Many fields are being transformed. Medicine is one of the fields that has seen the appearance of models that can solve complex tasks such as automatic image segmentation or diagnosis. However, there are important challenges that need to be overcome for a successful application in clinical practice. One important challenge is the generalization of models to unseen domains independently of other factors, such as the scanner manufacturer, the scanning protocol, the sample size or the image quality. In this thesis, we aim to investigate the effects of the domain shift in medical imaging, specifically for cardiovascular studies, which present a particular challenge since the heart is a moving organ. Furthermore, we aim to contribute to methods to overcome or reduce the model performance gap.

First, we establish a collaboration with clinical researchers from six different centres from three countries and assemble a large multi-centre dataset to tackle one of the greatest challenges in research: the domain gap problem. We process and annotate the data and develop a benchmark study by organizing an international competition to compare and analyse different techniques to bridge the generalization gap. The dataset is later open-sourced to foster innovation within the research community, becoming the first open multi-centre cardiac dataset.

Then, we perform an exhaustive comparison of domain generalization and adaptation methods, including the best-performing methods in the aforementioned competition, for late gadolinium-enhanced image segmentation for the first time. We show that extensive data augmentation is very important for generalization and that model fine-tuning can reach or even surpass multi-centre models.

Finally, we investigate the effects of differences in image appearance for the first time in a multi-centre study with cardiovascular imaging and compare several harmonisation techniques both at the feature and image levels for improved diagnosis. We show that histogram matching-based harmonisation results in image features (radiomics) that are more generalizable across centres.

Acknowledgements

This long and challenging adventure, like a pilgrimage, was full of pain and joy and many individuals were part of it.

First, I would like to express my gratitude to my supervisor Dr Karim Lekadir for bringing me into this adventure by allowing me to be part of the BCN-AIM research group and by sharing his great experience with me. Seemingly, I am very grateful to my co-supervisor, Dr Santi Seguí, for his scientific and personal support and advice during these years, and for having the patience to teach me the way. I thank both of them for their priceless academic guidance during this PhD programme. Additionally, I would like to thank Dr Sotirios A. Tsaftaris for welcoming me to Edinburgh and his research group with open arms. I certainly enjoyed every bit of the Scottish part of this adventure.

There are always good people along the way that give you a hand or stop by to have a drink or a 'tapa' with you, or organize after-work activities and create a human community. Among these people, I need to first thank my fellow pilgrims: Carlos, Cristian and Mireia, that walked most of the way with me and were an important support. However, I am thankful to many other nice people that made this adventure an unforgettable one: Jordi, Xabi, Amelia, Óscar, Akshay, Katharina, Kaisar, Malik, Anna, Carla, Lidia, Richard, Pedro, Sam, and Xiao, among others.

Lastly, to my good friends Javi, Álex and Guille and to my loved ones, my parents, my brother and Violeta. Thank you for always standing by me. Gracias. Gràcies.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	**A**rtificial **I**ntelligence
CMR	**C**ardiac **M**agnetic **R**esonance
CT	**C**omputed **T**omography
CNN	**C**onvolutional **N**eural **N**etwork
DA	**D**omain **A**daptation
DC/DSC	**D**ice (**s**imilarity) **c**oefficient
DG	**D**omain **G**eneralization
DL	**D**eep **L**earning
ED	**E**nd **d**iastole or **E**nd-**d**iastolic
ES	**E**nd **s**ystole or **E**nd-**s**ystolic
GAN	**G**enerative **A**dversarial **N**etwork
GLCM	**G**ray-**l**evel **c**o-ocurrence **m**atrix
GLDM	**G**ray-**l**evel **d**ependence **m**atrix
GLRLM	**G**ray-**l**evel **r**un **l**ength **m**atrix
GLSZM	**G**ray-**l**evel **s**ize **z**one **m**atrix
HCM	**H**ypertrophic **c**ardio**m**yopathy
HD	**H**ausdorff **d**istance
JSD	**J**ensen-**S**hannon **d**ivergence
LGE	**L**ate **G**adolinium **E**nhanced
LGE-CMR	**L**ate **G**adolinium **E**nhanced **C**ardiac **M**agnetic **R**esonance
LV	**L**eft **v**entricle
ML	**M**achine **L**earning
MRI	**M**agnetic **R**esonance **I**maging
MYO	**M**yocardium
PLHM	**P**iecewise **L**inear **H**istogram **M**atching
ROI	**R**egion **o**f **i**nterest
RV	**R**ight **v**entricle

Dedicado a mis padres, Manuel y Pilar.

Chapter 1

Introduction

The field of Artificial Intelligence (AI) has undergone a revolution in recent years with the development of more efficient computing hardware and well-documented software for model development. AI has become a tool with the potential to be applied in many and varied fields, such as transportation, art design or medicine. In some of the potential applications, such as medicine, these methods must meet specific requirements before they can be widely implemented in clinical practice. In this Thesis, we analyse a list of challenges for AI methods in the field of medical imaging and present a collection of results that focus on the challenge of domain shift for multi-centre cardiovascular datasets.

1.1 What is artificial intelligence?

AI can be defined as the intelligence demonstrated by a computer to solve a task. This *intelligence* can simply be a set of rules defined under human supervision, such as in a decision tree, or a method that extracts information from experience (training data) to build a predictive model. Techniques that follow the second paradigm are known as machine learning (ML) methods, which are superior to other methods because they automatically estimate a model or function that predicts the expected output given some input data. This is done through a process called *training*, where the function parameters are adjusted iteratively until the prediction error reaches a desired threshold. Among ML methods, deep learning (DL) refers to models that are inspired by the brain's neural network and that depend on a large number of parameters that need to be adjusted. These models are characterized by having greater expressivity, which can be defined as the capacity of the model to estimate increasingly complex functions. As a result, DL models have attracted the community's attention and have become the best-performing method in many applications. This Thesis focuses on ML methods, so we will hereafter refer only to ML or DL methods.

1.2 Cardiovascular imaging

Cardiovascular images are mostly acquired in clinical practice when a precise assessment of the heart's function and disease is needed. There are different modalities depending on the specific application. Some examples of image modalities are magnetic resonance imaging (MRI), computed tomography (CT) or ultrasound images. Some modalities are used because of their higher spatial resolution, others because they do not involve radiation or because are cheap and fast. In this Thesis, we focus on one of these modalities: MRI. The advantage of using MRI over other modalities is that is a non-invasive technique without radiation and the resulting images have a good temporal resolution. As a disadvantage, the exact MRI signal intensity is not

defined by the tissue properties – contrary to CT – and the spatial resolution is limited to around 1 millimetre (mm) and depends on the MRI magnet's strength. These characteristics make MRI scans suitable for functional and viability assessments of the heart [126]

MRI studies are acquired in two-dimensional slices that are stacked to form a 3D volume with a slice thickness of around 10 mm. Additionally, images are acquired repeatedly at different time points to obtain the whole cardiac cycle with a good temporal resolution. The whole process involves several periods when the patients are holding their breath to avoid respiratory movement causing artefacts in the images. Commonly, four different views are acquired that correspond to the heart's natural planes: the short axis – slices that are orthogonal to the heart's longitudinal axis – and three different views for the long axis – where the different chambers of the heart are visualized in the heart's longitudinal direction. Two of these views are depicted in Figure 1.1. A collection of sequences exists depending on the study being conducted (more details about sequences and specific protocols can be found in Pons-Lladó [126]). After the images have been acquired with acceptable quality, the clinical experts analyse them visually and quantitatively. The quantitative analysis involves a time-consuming annotation process that consists of identifying specific time points and delineating important areas of the heart for the extraction of key biomarkers.

In this Thesis, we will work with two different modalities. On one hand, we will use steady-state free precession or cine MRI sequences that are suited for the functional study of the heart. In these images, the blood is bright and muscular tissues, like the myocardium, are darker. On the other hand, we will consider contrast-enhanced MRI sequences, that are used for heart viability assessment and present greater image appearance variability. These images are qualitatively similar to cine MRI, but one may encounter bright spots within the myocardium depending on the viability of the tissue.

1.3 Artificial intelligence in medical imaging

In the last ten years, AI has been used to automatize a great variety of complex tasks in fields such as natural language processing [168] or computer vision [57] achieving impressive performance, and frequently surpassing state-of-the-art methods. Among the different fields, medical imaging is very active with a great number of publications demonstrating the added value of AI and an increasing number of AI-enabled devices being licensed and approved by the Food and Drug Administration in the United States for clinical practice applications [38]. These recent advancements anticipate a promising future for AI in healthcare with the potential to reduce the most time-consuming and tedious tasks as well as support expert clinicians in their day-to-day clinical practice. Among other applications, AI methods have been used for speeding up the processing of medical data, increasing the level of precision for diagnosis, discovering new disease biomarkers or decreasing medical errors in clinical workflows [169].

Multiple examples of the application of ML methods in medical imaging can be found in the literature and for a wide variety of tasks within the clinical workflow [71]. Starting with image acquisition, ML models have been used to speed up the acquisition process by reconstructing the final image with fewer input data or for improving the quality of the final image when a smaller dose of the contrast agent is injected into the patient [54]. Once an image is acquired, ML methods can be

used to improve its resolution [114, 125]. These models are trained to approximate the distribution of high-resolution images from a set of low-resolution images. As a result, one can generate a scan with full coverage of an organ from a limited number of slices or leverage images acquired with a lower radiation dose by improving their spatial resolution.

When analysing medical images, several applications have been proposed. For instance, DL methods are the current state-of-the-art for the segmentation of regions of interest in radiological images [5]. Important biomarkers, such as volume, are extracted from these regions of interest and then used for the patient diagnosis. Similarly, these tools have also been employed for the localization of points of interest and the detection of objects in medical images, such as nodules in breast mammograms or cells in histological imaging. Another potential application tested in the literature is the registration of images, which consists of finding the transformation that aligns two images based on the underlying anatomical structures [43]. This task is particularly useful for comparing images acquired with different modalities or at different moments in time and may require the application of non-trivial deformation mappings to obtain a good registration.

At the end of the pipeline, ML methods are also able to extract and utilize features from medical images or regions of interest previously segmented for the diagnosis of diseases [106]. These features are usually referred to as radiomics. The predictive power of these models lies in the ability to extract non-linear relationships between input data and the outcome without human intervention for feature selection. In some cases, these methods can even extract information from patterns in images that were not previously identified by human experts, such as patient race from different chest imaging modalities [52] or patient sex from fundus images [77].

Other emerging applications are appearing as well inspired directly by the specific needs of ML methods, such as the generation of realistic synthetic data for training better and more robust models [39]. Additionally, recent publications are also targeting the prediction of patient prognosis [23] or the determination of the risk of suffering an adverse event [128].

Overall, the ML community's activity is buoyant in the field of medical imaging and new and better ways to solve diverse medical tasks are being proposed at a very fast speed every year. In the next subsections, we are going to delve into specific details for each of the applications mentioned previously when proposed within the particular field of interest in this Thesis: cardiovascular imaging. Moreover, while medical imaging refers to a range of imaging modalities such as computed tomography, echography or magnetic resonance, the references provided as well as the results of this Thesis will focus on magnetic resonance imaging (MRI). However, the majority of methods presented can be applied to other modalities as well.

1.3.1 Artificial intelligence in cardiovascular imaging

Four main tasks can be found in the literature for cardiovascular image analysis that are summarized in Figure 1.1. For each of them, we provide a brief description and an overview of recent works using ML.

Image reconstruction

Cardiovascular medicine presents a special difficulty within medicine because the heart is a moving organ. This makes the image acquisition process more challenging

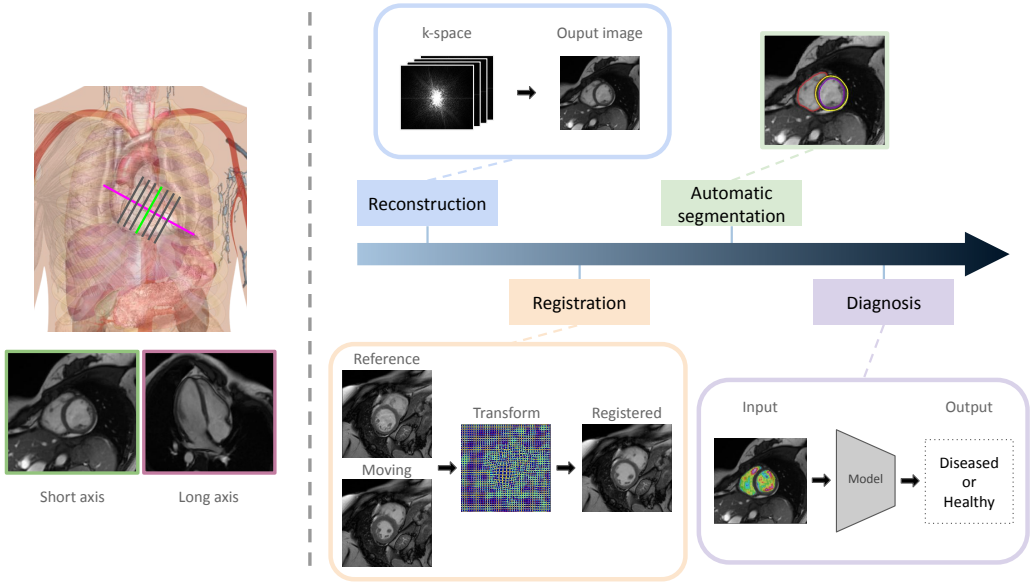


FIGURE 1.1: Visualization of the heart orientation in the chest with the two main MRI views used in the literature and a depiction of four important steps in cardiovascular image analysis where AI has been used.

due to the need to acquire the image under several breath-holds to avoid respiratory-related anatomical deformations. The scanning time may therefore take a long time. One way of reducing scanning time and patient discomfort is to reduce the number or duration of breath-holds, with the consequent decrease in spatial or temporal resolution. Alternatively, one can perform a free-breathing acquisition [159] with a longer scanning time and apply a post-processing alignment later to correct possible slice misalignments.

The versatility of DL methods has enabled the proposal of solutions to the two approaches presented to speed up the acquisition and reduce patient discomfort. On one hand, convolutional neural network (CNN) models can be trained to derive a full-coverage stack from a reconstructed sparse 3D cardiac volume [157]. Additionally, the *k*-space information can be leveraged together with the lower-resolution image to compute a 3D cardiac sequence under a single breath-hold in less than 10 seconds, which results in clinical biomarkers with good agreement with reference values [80]. On the other hand, a model for respiratory motion correction can be applied as post-processing to complement the free-breathing acquisition protocol [51].

Image registration

A second image post-processing step may be applied depending on the specific goal in mind for the acquired imaging modality. For instance, image registration is necessary for perfusion studies where MRI images are acquired during several heartbeats to be able to see the contrast agent washing out. This allows cardiologists to identify damaged regions in the heart walls, called myocardium. However, to compare accurately the images at different time points, these need to be morphologically aligned. DL can accurately register these images with only one optimization step [109], faster than traditional iterative methods such as ELASTIX [75] or ANTS [6].

Another potential application of registration algorithms is the construction of anatomical models of the heart from MRI scans with compromised myocardial tissue regions for revascularization therapy. The tissue information can be extracted from late gadolinium-enhanced (LGE) imaging and later registered to the MRI scan using manual annotations of the left and right ventricles and the left ventricle myocardium [158]. The deformation fields, extracted from an MRI scan, can also be used for extracting local metrics and assessing heart deformation for improved diagnosis [101].

Segmentation

When the MRI scan has been acquired and has undergone all necessary preprocessing steps, the expert clinician needs to manually delineate regions of interest in the heart for biomarker extraction and patient assessment that vary depending on the scan application. For instance, for studying the function of the heart cine sequences are considered and the basic annotated regions are the left and right ventricles and the left ventricular myocardium. Other regions of interest can also be included for a more exhaustive or targeted analysis such as the left and right atria, the left ventricle papillary muscles or the right ventricular myocardium. The whole annotation process takes an important amount of work and it is subject to interrater variability. To optimise time and effort, usually, a clinician only delineates the critical parts, such as specific frames and cardiac regions. In the example of functional assessment, only the time points of maximum contraction (end-systolic phase) and maximum dilation (end-diastolic phase) are annotated while the right ventricle epicardial contour and the papillary muscles are usually left out [126].

Multiple works have been presented in the academic literature for automatic segmentation of cardiac regions of interest and DL has stood out in recent years, becoming the state-of-the-art tool [24]. Diverse models have been proposed for the segmentation of cine sequences and enhancement imaging. The first fully convolutional neural network for short-axis cardiac MRI was proposed in 2016 by Tran [151]. From then on, networks with more advanced training pipelines and more effective architectures have been proposed (see the review by Chen et al. [24] and references therein) and optimised segmentation pipelines have been constructed that can be applied to varied medical image modalities [67]. Other cardiac modalities explored in the literature are late-gadolinium enhanced and T2 imaging [181] or perfusion imaging [140].

All the aforementioned models need in general a relatively large amount of studies annotated by clinical experts for training. To overcome this limitation, several approaches have been proposed to effectively learn from a few annotated samples leveraging also the information from unannotated studies. Some of these methods propose semi-supervised learning techniques, by learning to reconstruct the input images [96] for example, or self-supervised learning, where pretext tasks are created with other available information in a scan, such as orientation information [9].

Diagnosis

ML methods have been applied to cardiovascular imaging for the diagnosis of several cardiomyopathies based on features extracted from the scans. A large number of works have proposed different ML techniques for the diagnosis of diverse cardiovascular diseases such as myocardial infarction, left-ventricular non-compaction,

atherosclerosis or valvular heart disease, among others (see the review by Martin-Isla et al. [106] and references therein). One of the most common approaches in these works is based on the extraction of radiomics features (from regions of interest in the medical image) and the implementation of an ML pipeline that includes feature selection, cross-validation of the model in several folds or partitions of the training set and the selection of the final model to be evaluated. Other methods apply directly DL models to the images to classify them into positive (diseased) or negative (healthy). The latter approaches require less human interaction but need larger datasets to achieve comparative accuracy.

1.4 Challenges

The potential added value of AI tools in the clinical workflow has been widely demonstrated in multiple works, as shown by the variety of applications and the large number of academic contributions presented in the previous section. They can reduce acquisition times to a few seconds while reducing or removing the patient's breath-hold, improve the registration performance, output an automatic segmentation in less than a second or compute a risk score or a suggested diagnosis with high accuracy. However, important implementation challenges remain unsolved. As an example, consider the work conducted by Schaffter et al. [142], where the authors demonstrated that a combination of radiologists' predictions and ML outcomes achieves the highest overall accuracy for mammography screening. However, the specific implementation of these tools is important. Lehman et al. [84] showed that the number of detection failures increases when radiologists interpret mammography scans assisted by an automated system. This example highlights the importance of identifying and solving the challenges necessary for the successful implementation of AI in clinical practice.

In the following, we present some technical challenges related to the implementation of ML models and the interpretation of results. This Thesis will focus on the first one, domain shift, but there are other challenges that are worth introducing to provide more context due to their importance. Furthermore, there are other important challenges related to the infrastructure and knowledge gap of potential users that we do not discuss, as they are out of the scope of this Thesis.

Domain shift

One of the key recipes for a high-performance model is the use of a large number of parameters. In this way, the training algorithm can better fit the training data. This also allows the model to memorize certain patterns of the training data (imaging characteristics, traits for a specific population, biases from the data annotator) that might not be generalizable, limiting their applicability to new datasets [112].

AI models learn from existing datasets and may reproduce, as a consequence, existing biases. For instance, if the dataset has an ethnicity or sex-related imbalance, the model trained on it can show disparities in performance depending on the ethnic group or sex [129, 130]. Thus, creating an unfair treatment if deployed in real-life scenarios. Active research is being conducted to find methods that enforce fairness in model predictions [107], i.e. to improve the generalizability to underrepresented or missing groups.

Another type of bias may come from rapidly changing clinical settings with new scanner models or imaging modalities being introduced for a better assessment of

patients' health [45]. AI tools must be able to adapt to these changes without the need to repeat the training process every time the settings are modified. In this Thesis, we will focus on this particular challenge and will choose cardiovascular imaging as the application.

Interpretability

ML models are characterized by their ability to extract patterns from input data and build relations between features indirectly while training to optimise an objective function (e.g., the prediction accuracy). This property, together with the large number of parameters used in DL models, has resulted in the consideration of these models as 'black boxes', that is, of systems that generate an outcome without any explanation for how it was produced. However, if AI models are to be used in clinical practice they need to be trusted by clinicians by providing explanations and the level of uncertainty for their predictions. A human expert should be able to understand which factors motivated the model decision [143]. This lack of understanding of the decisive factors for a prediction has motivated the research community to focus on the *interpretability* of these methods.

As illustrated previously with an example for breast mammography screening, potential mistakes by automatized systems might be difficult to identify due to the black-box nature of many ML methods [170]. DL models tend to be overconfident in predictions even when these are incorrect. An automated model should be able to model the level of uncertainty in the prediction and output "I am not sure" when necessary.

Limited data and/or annotations

Clinical data is difficult to obtain due to legal constraints and when available, obtaining human annotations might be prohibitively expensive. For this reason, AI tools applied to the medical field must be very efficient. Several techniques are tackling this challenge by designing methods that learn from limited data (few-shot learning) or non-expert annotations (noisy labels), that select challenging samples and ask a clinical expert for feedback (active learning), that synthesize realistic samples using generative models or that fine-tune a model pre-trained with an unrelated dataset to leverage the optimised parameters.

Unbalanced data and rare diseases

AI models tend to overfit to the training set, resulting in several undesired behaviours such as a lack of generalizability to new cohorts or diseases with a limited representation in the whole population. Unfortunately, a balanced dataset or a dataset with the same number of patients for different diseases may not always be available. Therefore, there is a need for techniques that can handle these datasets and be adjusted according to possible imbalances. This is a key aspect to obtain fair models, i.e. models that perform with similar accuracy independently of the population subgroup.

Lack of evidence on prospective datasets

Very often existing academic publications report the performance on a retrospective cohort similar to the one used for training the model. This might lead to over-optimistic results that are not preserved on new cohorts from a different population

or acquired with a new scanner model or an updated acquisition protocol. As a consequence, one gets an undetermined performance gap when the models are implemented. Prospective validations are needed to assess the model performance in different and real clinical settings [170].

Trustworthiness and clinical acceptance

Apart from solving all of the aforementioned challenges, ML models need to obtain the clinical expert's acceptance to be applied in day-to-day practice. For that, strong evidence of its usefulness and added value is necessary.

1.5 Domain shift

In this Thesis, we will focus on one of the aforementioned challenges: domain shift. Domain shift refers to the existing differences between datasets due to different population characteristics, different acquisition protocols or different acquisition devices used, among other factors. These differences may cause the AI models to decrease their performance or simply fail when tested on new target domains. For this reason, the research community has been investigating how to adapt features learnt from one or more source domains to new domains or how to learn generalizable features directly during the training phase.

A clear example of a domain shift that results in poor generalizability can be obtained when training an ML model for skin cancer detection only with a specific population, such as white-skinned people. When the model is applied to a black-skinned person, it fails to detect cancer because many of the patterns learnt were specific to brighter skin images. The same may happen for models that use other types of imaging, such as computed tomography or MRI, where the extracted patterns and the differences among populations are usually less interpretable.

Depending on the availability of data and annotations from source and target domains, the problem of domain shift can be tackled using different methodologies. Let us assume we have always annotated data from the source domains[1] and that we want to train a model with the least performance gap between source and target domains with very limited human effort regarding data acquisition and annotation. Then, we are interested in methods that leverage differences between source domains to build a generalizable model (the problem of **Domain Generalization**) and methods that can be adapted to the new domain with few target samples with or without annotations (the problem of **Domain Adaptation**).

1.5.1 Domain generalization

Domain generalization methods use variations in one or more source domains to fit a model that can perform accurately in an unseen target domain.

A majority of works focus on building a feature representation that is the result of reducing the dissimilarity of features extracted across training domains [50, 62, 92, 110, 119]. The motivation lies in the (more or less justified) intuition that a feature representation with minimal dissimilarities across domains can generalize better to unseen domains. Khosla et al. [74] even disentangles these representations into two components – a domain-specific one and a domain-invariant one – and uses the

[1] If we did not have annotated data for training, we would need to focus on unsupervised learning methods.

latter for training generalizable models. Most of these works embed the representations in a reproducing kernel Hilbert space, but recent approaches minimize other metrics like the Wasserstein distance [4]. Or let the model learn the representation through an encoder-decoder model that reconstructs the original image for different source domains using one single latent representation [49].

Other approaches focus on building an ensemble of models trained with diverse datasets to obtain averaged and robust predictions on unseen domains [163].

Finally, recent proposals rely on adversarial training to enforce the model to learn parameters or latent features that are indistinguishable across domains [87, 91]. Or use meta-learning [36] or metric-learning [177] methodologies to build feature representations that preserve class relationships and enforce the clustering of data points from the same class independently of the domain.

1.5.2 Domain adaptation

The problem of *domain adaptation* (DA) has the largest list of contributions. Early methods were based on the alignment of vectorial representations of images, referred to as shallow DA methods, while more recent methods rely on deep learning-based features or are end-to-end DL methods [31, 174].

Among the shallow DA methods, the majority are based on the application of small perturbations to the data points – to better approximate the target domain distribution – or to the model parameters – to adapt the model decision boundary to the target data representation. Csurka [31] provides a detailed review of methods such as Transfer Adaptive Boosting [33] or Adaptive Support Vector Machine [164] among many others.

The advent of DL methods has resulted in improved feature representations of images as compared to handcrafted features by using the DL model as a feature extractor. When the source and target domains are sufficiently close, these DL-based features perform generally better [35]. However, if the difference between domains is larger, one needs to apply transfer learning techniques to fine-tune the model parameters – and thus the image deep representations – to the new domain using some annotated samples from the target domain [29, 116].

Lastly, recent works have focused on the development of end-to-end techniques fully based on DL models referred to as deep DA. These models leverage the expressivity of complex neural networks to boost performance. We classify them into two groups:

- **Image reconstruction-based approaches.** This approach consists of learning a latent space invariant to domain differences through an encoder-decoder-based model. This has been enforced by reconstructing paired images with small variations (rotations, dilations or different views of them) from a single vectorial representation of an original image [48]. Or by using synthetically generated images with the same class as the original image to simulate alternative domains [11, 12]. Finally, some works have proposed models based on generative adversarial networks (GANs) [53] to learn a domain-invariant representation without the need for paired images [93].

- **Classification-based approaches.** These approaches utilize siamese networks – i.e. two copies of the same network – trained together, one with data from the source domain and the other with data from the target domain. During training, these methods minimize a loss that measures the discrepancy between

model parameters from both domains [99, 146], a classification loss or an adversarial loss to enforce that the model features are not able to predict the original image domain [153, 154] or a combination of several or all of them [46, 61, 98, 100]. In some works, the model parameters are learnt independently but in the majority of cases, the weights are shared partially or entirely across networks.

1.5.3 Domain shift in medical imaging

The problem of domain shift is particularly relevant in computer vision, where images are a representation of the real world and have always an implicit bias, that depends on the acquisition process [150]. Medical imaging in particular poses a narrower problem — meaning that the variability of images is lower — when compared with general computer vision tasks where different objects with varied backgrounds need to be identified or classified. Medical images are in general acquired within a specific range of the target region of interest or organ, with similar orientations and similar devices. However, differences in acquisition parameters of protocols or scanning devices result in subtle differences in the images that affect model generalization, which results in unwanted disparities in prediction accuracy. In general, few public datasets or works existed in the literature to study the domain shift problem in medical imaging, but they have been collected and open-sourced in recent years. For instance, Prados et al. [127] collected a multi-centre and multi-vendor dataset with 80 cases for a spinal cord segmentation challenge, although the authors did not address the domain shift problem explicitly.

Several recent studies have reported on the performance degradation of state-of-the-art ML tools in medical imaging. For instance, Ting et al. [149] showed disparities in performance for disease classification of retinal images when the model was tested on external data, with an area under the curve score ranging from 0.889 to 0.983 depending on the dataset population and acquisition clinic. For radiological images, the first results evidencing the domain gap of CNNs were presented by Al-Badawy, Saha, and Mazurowski [2], Bai et al. [7] and Zech et al. [173]. The authors of these studies demonstrated the level of degradation of CNN-based models for brain MRI, cardiac MRI and chest X-ray datasets, respectively, when tested on external datasets. To address and surmount the performance gap, existing works rely on DA and DG techniques depending on the availability of samples from the target domain during training. We will discuss both approaches separately next.

Domain generalization in medical imaging

To optimise performance on an unseen target domain, Yao et al. [165] collected ten chest X-ray datasets from different clinical centres and demonstrated that combining several datasets during training improves domain generalization. In cardiac imaging, Tao et al. [148] published the first study comparing automatic segmentation results on varied manufacturers from multiple clinical centres. The authors demonstrated that the best-performing model was again trained with a varied dataset combining data from several institutions and manufacturers. However, the results of the study could not be reproduced since the datasets are not publicly available. Later, Chen et al. [25] proposed a method to improve the generalization of the model across domains by using extensive transformations of the input images during training,

usually referred to as data augmentation. This provided more variability to the training data and limited the accuracy drop to around 0.1 in Dice score in two external datasets with the majority of images acquired with scanners from one manufacturer.

Domain adaptation in medical imaging

When data is available from the target domain, a greater variety of proposals exists. For example, some studies have proposed the use of GANs to generate realistic synthetic images based on images from other modalities and increase the number of training samples [14, 26, 104, 105], or directly transform the images from the target domain to the source domain via image-to-image translation models [27, 37, 176].

Other works have used techniques for deep DA (introduced in the subsection 1.5) based on domain classifiers for medical imaging applications [68, 70, 81].

Finally, a different approach by Perone et al. [123] presents a self-ensembling method on the model weights for unsupervised DA for spinal cord MRI image segmentation. This model uses unlabeled images from the target domain and computes the prediction consistency between the teacher and student networks.

1.6 Aim and contributions

In this Thesis, we aim to investigate the effects of the domain shift in medical imaging and to contribute to methods to bridge or reduce the model performance gap. In particular, we focus on the differences found in cardiovascular MRI scans acquired at several institutions with different scanners and protocols and varied cardiac pathologies.

The contributions of this Thesis can be summarized as follows:

- We assembled a multi-centre dataset containing 375 cardiovascular magnetic resonance images from different scanner manufacturers and diverse pathologies and established the largest benchmark for multi-centre generalizable segmentation methods.
- We demonstrated the effect of diverse domain generalization and adaptation techniques for the segmentation of the more challenging late gadolinium-enhanced imaging.
- We investigated the effect of several harmonisation methods on classification models when tested on unseen data from other centres.

More in detail, to tackle the domain gap problem between clinical centres, we established a collaboration with several institutions in Spain, Germany and Canada within the context of the Horizon 2020 EU project euCanSHare to assemble a diverse dataset of cardiac magnetic resonance (CMR) images. A total of 375 studies were collected from six clinical centres from four different scanner manufacturers with a diversity of pathologies that are present in a real clinical setting. The dataset was later curated, annotated and transformed into the appropriate format for subsequent image analysis. The processed data were then used for organizing an international competition in the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2020 conference to compare solutions for the domain generalization problem in a controlled setting. A specific setting was designed for the competition to assess performance on three scenarios depending on the availability of data for each

scanner manufacturer: 1) annotated data is available, 2) unannotated data is available and 3) no data is available. A total of 14 international teams submitted a final proposal for the challenge. Further details of the competition settings as well as the results and conclusions extracted from the participating methods were published in the journal paper:

> Campello, Víctor M. et al. (2021). "Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge". In: *IEEE Transactions on Medical Imaging*. Doi: 10.1109/TMI.2021.3090082.
> JCR IF: 10.6 Q1.

The collected dataset was then released, becoming the first open-source multi-centre and multi-vendor dataset in cardiovascular imaging. This contribution increases the amount of openly accessible cardiac datasets for further validation and quantification of the generalization gap of ML models.

Subsequently, having addressed the problem in cine MRI we studied the more challenging contrast-enhanced modality. The best-performing methodologies from the M&Ms Challenge were compared against transfer learning approaches for model generalization of late enhanced CMR image segmentation for the first time. Late-enhanced imaging is a technique that uses MRI to visualize regions of interest when a ferromagnetic contrast agent has been injected into the bloodstream. This allows for the detailed analysis of tissue viability, that is to assess whether the blood flow is normal within a tissue. These images present greater variability due to the disparities in time elapsed between the image acquisition and the time of contrast injection. We show that single-centre models can generalize well to new domains when trained with appropriate data augmentation and that transfer learning achieves comparable accuracy to the best models with fewer computational resources. The whole analysis has been published in the journal paper:

> Sendra-Balcells, Carla, Campello, Víctor M., et al. "Domain generalization in deep learning for contrast-enhanced imaging." *Computers in Biology and Medicine* 149 (2022): 106052. Doi: 10.1016/j.compbiomed.2022.106052.
> JCR IF: 7.7 Q1.

Finally, having addressed generalizability in segmentation we focus on model generalization on imaging features, that are useful for diagnosis. We studied the effect of centre-related variability on features extracted from CMR images. These features are combined with clinical biomarkers to improve diagnostic performance. However, they can be affected by subtle differences in scanner manufacturers or acquisition protocols that are not related to biological factors. To reduce the impact of scanner-related variability, image and feature harmonisation techniques were compared to obtain standardized features from images across different centres with different scanner manufacturers. The quality of the harmonisation was assessed in terms of feature similarity across centres for groups of healthy and pathological subjects separately and in terms of diagnosis generalization to unseen domains. For the first time, the effect of image harmonisation on cardiovascular multi-centre studies was reported. The results have been published in the journal paper:

> Campello, Víctor M., et al. "Minimising multi-centre radiomics variability through image normalisation: a pilot study." *Scientific Reports* 12, 12532 (2022). Doi: 10.1038/s41598-022-16375-0.
> JCR IF: 4.6 Q2.

1.7 Thesis outline

This dissertation is based on a compendium of research works published in top-tier journals that are presented as separate chapters. Preceding these works, in Chapter 1 the motivation of this Thesis has been presented together with an introduction to AI, cardiovascular imaging and a description of different applications in cardiovascular image analysis. Also, a list of existing challenges for AI in medical imaging is introduced and a review of state-of-the-art applications for the specific challenge of domain shift is provided. This chapter intends to provide a general overview of the state-of-the-art of the two connecting fields of this Thesis (computer science and medical imaging) so that the document is self-contained. Finally, the contributions of this Thesis have been presented.

Chapter 2 describes the details of the M&Ms Challenge, including the data preprocessing, design and metrics as well as the solutions proposed by the participating teams. These results are then analysed to extract conclusions about model generalizability.

In Chapter 3, several adaptation techniques for DL segmentation are compared when working with multi-centre LGE images.

In Chapter 4, several pre-processing methods are compared to harmonize features extracted from images from different centres with different scanner manufacturers. Methods that are applied both at the image and feature level are compared and their effects are assessed for a diagnosis task.

Finally, Chapter 5 summarises the results obtained during the PhD programme and highlights future directions to be investigated.

Chapter 2

Domain shift for multi-centre image segmentation

Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge

This chapter contains material from:

> Campello, Víctor M. and Gkontra, Polyxeni and Izquierdo, Cristian and Martín-Isla, Carlos and Sojoudi, Alireza and Full, Peter M. and Maier-Hein, Klaus and Zhang, Yao and He, Zhiqiang and Ma, Jun and Parreño, Mario and Albiol, Alberto and Kong, Fanwei and Shadden, Shawn C. and Acero, Jorge Corral and Sundaresan, Vaanathi and Saber, Mina and Elattar, Mustafa and Li, Hongwei and Menze, Bjoern and Khader, Firas and Haarburger, Christoph and Scannell, Cian M. and Veta, Mitko and Carscadden, Adam and Punithakumar, Kumaradevan and Liu, Xiao and Tsaftaris, Sotirios A. and Huang, Xiaoqiong and Yang, Xin and Li, Lei and Zhuang, Xiahai and Viladés, David and Descalzo, Martín L. and Guala, Andrea and La Mura, Lucia and Friedrich, Matthias G. and Garg, Ria and Lebel, Julie and Henriques, Filipe and Karakas, Mahir and Çavuş, Ersin and Petersen, Steffen E. and Escalera, Sergio and Seguí, Santi and Rodríguez-Palomares, José F. and Lekadir, Karim. "Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge". In: *IEEE Transactions on Medical Imaging* (2021). Doi: 10.1109/TMI.2021.3090082.

2.1 Introduction

Accurate segmentation of cardiovascular magnetic resonance (CMR) images is an important pre-requisite in clinical practice to reliably diagnose and assess several major cardiovascular diseases [19, 106]. Currently, the process typically requires the clinician to provide a significant amount of manual input and correction to accurately and consistently annotate the cardiac boundaries across all image slices and cardiac phases. The automation of such a tedious and time-consuming task has been pursued for a long time by using multiple approaches, such as statistical shape models [3] or cardiac atlases [8]. In the last few years, the advent of the deep learning paradigm has motivated the development of many neural network-based techniques for improved CMR segmentation, as listed in a recent review [24]. However,

most of these techniques have been all too often trained and evaluated using cardiac imaging samples collected from single clinical centres using similar imaging protocols. While these works have advanced the state-of-the-art in deep learning-based cardiac image segmentation, their high performances were reported on samples with relatively homogeneous imaging characteristics.

As an example, the CMR datasets from the Automated Cardiac Diagnosis Challenge (ACDC) dataset [10] have been extensively used to build and test new implementations of deep neural networks for cardiac image segmentation. The top performing technique in the ACDC challenge, proposed by Isensee et al. [66], obtained a very high segmentation accuracy for both the left and right ventricles. However, the ACDC datasets were compiled from 150 subjects scanned at a single clinical centre using the same imaging protocol, which limits the ability of the researchers to develop and test models that can generalize suitably across multiple centres and scanner vendors. Other researchers attempted to encode higher variability by building and testing their models based on much larger datasets obtained from the UK Biobank [20]. For instance, Bai et al. [7] implemented a fully convolutional network that achieved highly accurate results on this large dataset (over 4,875 cases), but the authors concluded that their model might not generalize well to other vendor or sequence datasets.

Some researchers proposed to improve CMR segmentation by training neural networks with images from multiple cohorts [148, 151], but these works do not include methods for addressing domain shifts between training and new unseen cohorts. Other works used data augmentation on models built from single cohorts such as the ACDC [73] or the UK Biobank [25], then tested their techniques on other existing public cohorts, including the Sunnybrook Cardiac Data [131], LV Segmentation Challenge Dataset (LVSC) [145] or RV Segmentation Challenge Dataset (RVSC) [124]. However, these studies are limited by the fact that these different CMR cohorts have been annotated with distinct standard operating procedures (SOPs), which makes it difficult to draw conclusions from the multi-cohort comparative results. Furthermore, such an approach requires a large training dataset from the single centre to model high variability across subjects. Another multi-centre and multi-vendor study conducted by Tao et al. [148] relied solely on private data, which makes it difficult to replicate the results and perform community-driven benchmarking. While these recent works confirmed the difficulties encountered by deep learning models to generalize beyond the training samples, they also support the need for well-defined heterogeneous public datasets that can be used by the community to improve model generalizability through scientific benchmarking.

In this context, the Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation (M&Ms) Challenge was proposed and organized as part of the Statistical Atlases and Computational Modelling of the Heart (STACOM) Workshop, held in conjunction with the MICCAI 2020 Conference. The M&Ms challenge was set up as part of the euCanSHare international project[1], which is aimed at developing interoperable data sharing and analytics solutions for multi-centre cardiovascular research data. Together with clinical collaborators from six different hospitals in Spain, Canada and Germany, a public CMR dataset was established from 375 participants, scanned with four different scanners (Siemens, Philips, General Electric (GE) and Canon) and annotated using a consistent contouring SOP across centres.

To our knowledge, this dataset is the most diverse resource of CMR studies,

[1] euCanSHare project website: www.eucanshare.eu

which is provided as open-access[2] to promote further research and scientific benchmarking in the development and evaluation of future generalizable deep learning models in cardiac image segmentation. In this paper, we also present and discuss the results of the M&Ms challenge in detail, to which a total of 14 international teams submitted a range of solutions, including different strategies of transfer learning, domain adaptation and data augmentation, to accommodate for the differences in scanner vendors and imaging protocols. The obtained results show the extent of the problem, the promise of the proposed solutions, as well as the need for further research to build fully generalizable tools that can be translated reliably and deployed in routine clinical practice across the globe.

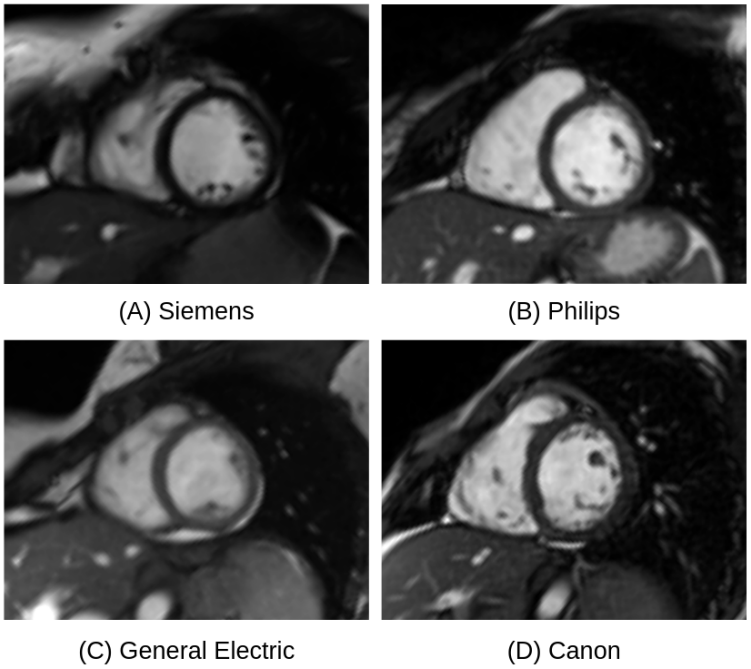


FIGURE 2.1: Visual appearance of a CMR short axis middle slice for anatomically similar subjects in the four different vendors considered.

2.2 Challenge framework

2.2.1 Data preparation

TABLE 2.1: Information from centres included in this work.

	Name	City	Country
1	Hospital Vall d'Hebron	Barcelona	Spain
2	Clínica Sagrada Familia	Barcelona	Spain
3	Universitätsklinikum Hamburg-Eppendorf	Hamburg	Germany
4	Hospital Universitari Dexeus	Barcelona	Spain
5	Clínica Creu Blanca	Barcelona	Spain
6	McGill University Health Centre	Montreal	Canada

[2]The dataset is publicly available at www.ub.edu/mnms

TABLE 2.2: Distribution of the most frequent pathologies and healthy volunteers between centres. The abbreviations correspond to hypertrophic cardiomyopathy (HCM), dilated cardiomyopathy (DCM), hypertensive heart disease (HHD), abnormal right ventricle (ARV), athlete heart syndrome (AHS), ischemic heart disease (IHD) and left ventricle non-compaction (LVNC).

Pathology \ Centre	1	2	3	4	5	6
Healthy vol.	22	33	32	21	14	3
HCM	25	37	14	8	15	4
DCM	37	-	5	-	9	-
HHD	-	4	-	19	1	1
ARV	12	-	-	2	1	1
AHS	-	-	-	3	-	-
IHD	-	-	-	4	1	3
LVNC	-	-	-	-	2	2
Other	-	-	-	18	7	15

TABLE 2.3: Average specifications for the images acquired in the different centres.

Centre	Vendor	Model	Field strength (T)	In-plane resolution (mm)	Slice thickness (mm)	Number of slices	Number of time frames
1	Siemens	MAGNETOM Avanto	1.5	1.32	9.2	12	25
2	Philips	Achieva	1.5	1.20	9.9	10	30
3	Philips	Achieva	1.5	1.45	9.9	11	26
4	GE	Signa Excite	1.5	1.36	10	12	25
5	Canon	Vantage Orian	1.5	0.85	10	13	29
6	Siemens	MAGNETOM Skyra	3.0	0.98	9.7	12	29

A total of six clinical centres from Spain, Canada and Germany (numbered 1 to 6 in this work) contributed to this challenge by providing a different number of CMR studies from different scanner vendors, as detailed in Table 2.1. In total, 375 studies were included in this challenge. The subjects considered for this multi-disease study were selected among groups of various cardiovascular diseases, such as hypertrophic cardiomyopathy, dilated cardiomyopathy, coronary heart disease, abnormal right ventricle, myocarditis and ischemic cardiomyopathy as well as healthy volunteers (see Table 2.2 for more details on the distribution of these cases). The specific scanner manufacturers are 1) Siemens (Siemens Healthineers, Germany), 2) Philips (Philips Healthcare, Netherlands), 3) General Electric (GE, GE Healthcare, USA) and 4) Canon (Canon Inc., Japan). These four manufacturers were coded as A, B, C and D during the challenge, respectively. The CMR images derived from these four vendors are illustrated in Fig. 2.1. More specific details on the studies are given in Table 4.2.

Every CMR study was annotated manually by an expert clinician from the centre of origin, with experiences ranging from 3 to more than 10 years. Following the clinical protocol, short-axis views were annotated at the end-diastolic (ED) and end-systolic (ES) phases, as they correspond to the phases used to compute the relevant clinical biomarkers for cardiac diagnosis and follow-up. Three main regions were considered: the left and right ventricle (LV and RV, respectively) cavities and the left ventricle myocardium (MYO). In order to reduce the inter-observer and inter-centre variability in the contours, in particular at the apical and basal regions, a detailed revision of the provided segmentations was performed by four researchers in pairs. They applied the same SOP across all CMR datasets to obtain the final ground truth. To generate consistent annotations for the research community, we chose to apply

the SOP that was already used by the ACDC challenge, as follows:

a) The LV and RV cavities must be completely covered, including the papillary muscles.

b) No interpolation of the MYO boundaries must be performed at the basal region.

c) The RV must have a larger surface at the ED time-frame compared to ES.

d) The RV does not include the pulmonary artery.

Clinical delineations as well as later corrections were performed using CVI42 software (Circle Cardiovascular Imaging Inc., Calgary, Alberta, Canada). All studies were provided in DICOM format and contours were extracted in cvi42 workspace format (.cvi42ws). An in-house software was then used to extract the contours and transform the images into the NIFTI format, representing the final files delivered to the challenge participants.

TABLE 2.4: Number of studies for each step of the challenge presented by centre and scanner vendor.

	Siemens		Philips		GE	Canon	Total
Label	A		B		C	D	
Centres	1	6	2	3	4	5	
Training	75	0	50	25	25	0	175
Validation	5	5	5	5	10	10	40
Testing	16	24	19	21	40	40	160
Overall	96	29	74	51	75	50	375

2.2.2 Model training

The 375 CMR studies were divided into three sets, namely training, validation and testing, as detailed in Table 2.4. To decide on a particular subdivision, we first estimated the degree of generalizability of models trained from the four vendors, as shown in Figure 2.2. We have thus decided to combine the datasets from vendors A, which generalize relatively well, with datasets from B, which generalize poorly to new vendors, as training datasets. The participants received the 175 training cases on 1st May 2020, including 75 annotated CMRs from vendor A, 75 annotated CMRs from vendor B, 25 CMRs from vendor C but without any annotations (only the raw images) and no datasets from vendor D, to test generalizability to different situations (e.g. image protocol included or not included in the training). Note that in the case of vendor A, the 75 CMRs were included from centre 1 but none from centre 6, to test generalizability across vendors but also across centres for the same vendors. Regarding vendor B, we included more training datasets from centre 2 (50 cases) than from centre 3 (25 cases) to assess the impact of imbalanced training data and fairness in multi-centre cardiac image segmentation. For optimizing the models, the participants were allowed to remotely validate against 40 additional CMRs, i.e. 10 from each of the four vendors. A maximum of 7 submissions were allowed per team during the validation process. Note that during training, it was not allowed to use any external datasets or pre-trained models, to enable a fair comparison between the proposed solutions.

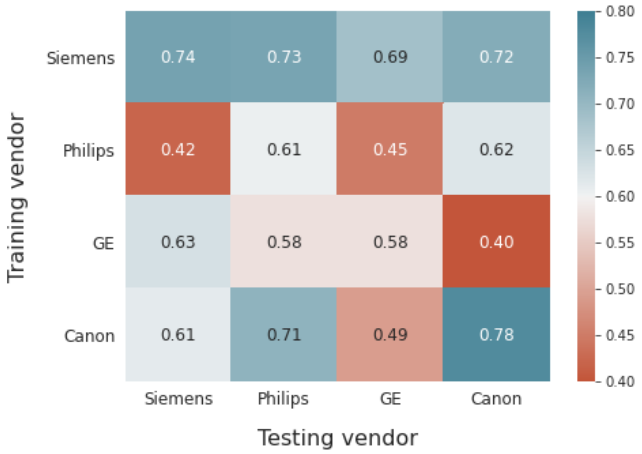


FIGURE 2.2: Degree of generalizability of models trained from the four vendors. Four 2D UNet models [137] were trained with datasets from the four vendors separately (rows) and subsequently tested their segmentation performance on datasets from all vendors (columns). The heatmap shows the Dice similarity coefficient, with a colour scale that goes from blue (good generalizability) to red (poor generalizability). The results are the average of 5 models cross-validated on subsets of 30 training subjects.

2.2.3 Model evaluation

The testing period for the challenge started on 8th June 2020 and concluded on 15th July 2020. The participants had to evaluate their models remotely to ensure the unseen datasets were totally hidden from the segmentation methods. As such, for example, the participants had no prior information on the images provided by vendor D. To evaluate the models, the participants were asked to build a Singularity image[3] and share it with the organizers via a MEGA[4] folder shared by the organizers or by any other secure cloud storage service. This Singularity image allows its execution on a similar architecture machine without the need to install all the diversity of used libraries. The necessary computing power was sponsored by NVIDIA, who provided the organizers with access to an NVIDIA V100 GPU card with 16GB of memory, as well as the Barcelona Supercomputing Center (BSC) who provided access to two K80 NVIDIA GPU cards.

To assess the quality of the automatically segmented masks P with respect to the ground truth G, four measures were proposed, namely:

(i) Dice similarity coefficient (DSC):

$$DSC(P,G) = \frac{2|P \cap G|}{|P| + |G|} \tag{2.1}$$

that measures the degree of overlapping of two volumes.

(ii) Jaccard index (JI):

$$JI(P,G) = \frac{|P \cap G|}{|P \cup G|} = \frac{|P \cap G|}{|P| + |G| - |P \cap G|} \tag{2.2}$$

[3]https://sylabs.io
[4]https://mega.nz

that measures overlapping as well but is more sensitive to results with average performance.

(iii) Average symmetric surface distance (ASSD):

$$ASSD(P,G) = \frac{1}{|P|+|G|}\left(\sum_{p\in P} d(p,G) + \sum_{g\in G} d(g,P)\right)$$
$$d(p,G) := \inf_{g\in G} d(p,g) \tag{2.3}$$

that measures the average distance between the two volumes.

(iv) Hausdorff distance (HD):

$$HD(P,G) = \max\left\{\sup_{p\in P} d(p,G), \sup_{g\in G} d(g,P)\right\} \tag{2.4}$$

that measures the largest disagreement between the volumes and it is useful for identifying small outliers. All these metrics were computed using the public library medpy[5].

These metrics were computed for the three target labels: LV, RV, and MYO, resulting in a total of 12 measures. In case one participant had a prediction missing for a specific subject, a value of zero was assumed for DSC and JI and maximum values of 150 and 50 millimetres were assumed for HD and ASSD, respectively, based on the worst results obtained by the participating methods. Any value above the thresholds on surface distances was set to the maximum value.

To obtain the final ranking for each team, a weighted average was computed giving greater importance to the unlabelled and unseen scanner vendors. Therefore, if v_A and v_B are defined as the labelled vendors, v_C, the unlabelled one and v_D, the unseen one, the weighted sum for a metric M is obtained as follows:

$$M = \frac{1}{6}M_{v_A} + \frac{1}{6}M_{v_B} + \frac{1}{3}M_{v_C} + \frac{1}{3}M_{v_D} \tag{2.5}$$

Then, a min-max normalization was applied across participants for each measure and a final average over the normalized metrics yielded the performance (P) ranging from 0 to 1, being 1 the value that a team would obtain if it had the best results for every metric.

2.3 Participating methods

In total, 80 teams registered to download the M&Ms training dataset, 16 submitted a solution for the final testing phase and 14 teams submitted their methodology as a paper to the STACOM Workshop (see Table 2.5 for details on these teams). All participants used deep learning as their segmentation approach. Table 2.6 summarizes the main characteristics of the submitted techniques, including the backbone architectures and domain adaptation strategies, which are described in more detail in the following subsections. Furthermore, details on the hardware used during training and the times that each method took for training and inference as well as the number of parameters for each model are presented in Table 2.7.

[5]https://github.com/loli/medpy

TABLE 2.5: List and details of the participating teams in the challenge.

Team	Institution	Location	Name during challenge	Reference
P1	German Cancer Research Center (DKFZ)	Heidelberg, Germany	Mountain goat	[44]
P2	Chinese Academy of Sciences	Beijing, China	Dugong	[175]
P3	Nanjing University of Science and Tech.	Nanjing, China	Opossum	[103]
P4	Universitat Politècnica de València	València, Spain	Ox	[120]
P5	University of California	Berkeley, USA	Monkey	[76]
P6	University of Oxford	Oxford, UK	Donkey	[30]
P7	Nile University	Cairo, Egypt	Porpoise	[139]
P8	Technical University of Munich	Munich, Germany	Owl	[88]
P9	Aristra GmbH	Berlin, Germany	Lovebird	[72]
P10	King's College London	London, UK	Mandrill	[141]
P11	University of Alberta	Edmonton, Canada	Muskox	[18]
P12	University of Edinburgh	Edinburgh, UK	Springbok	[94]
P13	Shenzhen University	Shenzhen, China	Seagull	[63]
P14	Fudan University	Shanghai, China	Steer	[90]

TABLE 2.6: Characteristics of participating models. Abbr: rotations (R), flipping (F), scaling (S), deformations (D), histogram matching (HM), Gaussian noise (GN), brightness (B), gamma (G), test time augmentation (TTA).

Method	Backbone architecture	Data augmentation: Spatial augmentations				Data augmentation: Intensity-based augmentations					Others	TTA	Domain adapt.
		R (°)	F	S	D	HM	GN	B	G	Synthesis			
P1	nnUNet	±180	✓	✓	✓		✓	✓	✓		contrast	✓	No
P2	nnUNet	±180	✓	✓	✓	✓			✓		label propagation	✓	No
P3	nnUNet	±180	✓	✓	✓	✓			✓			✓	No
P4	UNet (ResNet-34)	±45	✓		✓						translations		Yes
P5	Attention UNet	±10		✓						CycleGAN	low-level frequency		No
P6	UNet+DA +DUNN	±180	✓								translations		Yes
P7	UNet	±15	✓	✓				✓					No
P8	DRUNet	±15		✓	✓			✓	✓	CycleGAN	blurring		No
P9	nnUNet	±180	✓		✓							✓	No
P10	UNet	±22.5		✓	✓		✓	✓			translations		Yes
P11	UNet++ (ResNet101)												No
P12	SDNet			✓						VAE			No
P13	UNet	±90	✓	✓				✓		WaveCT-AIN [97]	contrast	✓	No
P14	UNet									CycleGAN			No

2.3.1 Backbone architectures

There is a degree of variability in the backbone architectures used between the different participants, as shown in Table 2.6. Four teams used the nnUNet [67] (which includes UNet architectures in 2D and 3D as well as a cascaded UNet) as their baseline segmentation model (P1-P3 & P9). Four participants used a traditional UNet [137] (P6, P10, P13, P14), while other variants of UNets were adopted by the rest of the teams. In particular, UNets combined with residual connections were applied by three teams (P4, P8, P11), with P8 preferring a residual UNet with dilated convolutions (DRUNet) [89]. P5 proposed the use of an attention UNet [113], while P7 developed a modified UNet based on multi-gate and dilated inception blocks to extract multi-scale features. Lastly, one team (P12) proposed a modified Spatial Decomposition Network (SDN) [21] with an AdaIN [64] decoder.

As pre-processing techniques, all models that provided detailed information about this step performed either image normalization to a unit Gaussian distribution or pixel value rescaling to the range [0,1] (only P6 chose the range [0,255] instead). With regards to image resolution, images were resized based on target size or pixel resolution values in 10 out of 14 methods, while the other methods preferred to

keep the original image resolution (P4, P7, P8, P11). In order to obtain squared images, cropping and zero padding were used depending on the desired image size for each case. Additionally, some methods applied intensity clipping between varying ranges to get rid of bright artefacts (P5, P6, P11). Finally, P8 was the only method to apply also a non-local means denoising filter prior to the training process.

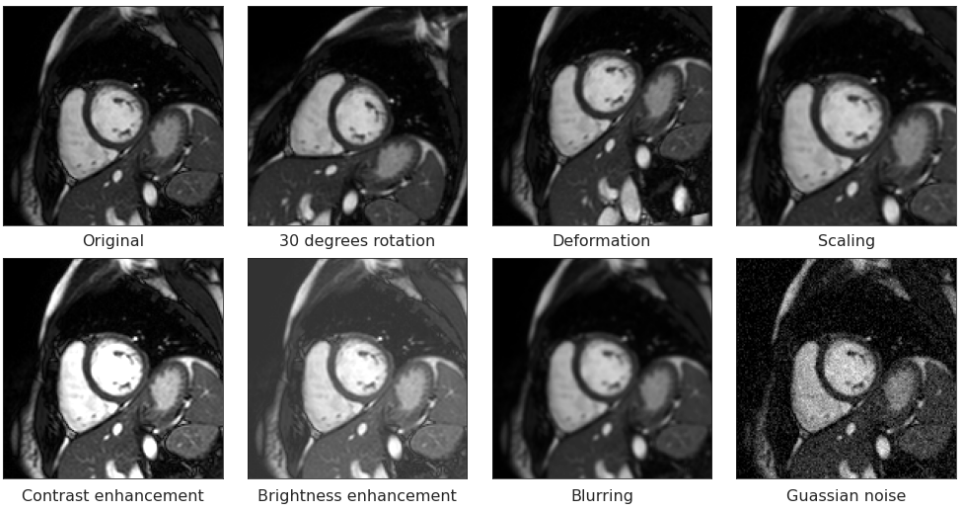


FIGURE 2.3: The effect of data augmentation on a single CMR slice. In the top row, the original image and spatial augmentations are shown. In the bottom row, intensity-based augmentations.

2.3.2 Data augmentation

All participants in the challenge (except P11) used some form of data augmentation to enhance their models. Specifically, two families of data augmentations were considered: (1) spatial transformations to increase sample size through rotation, flipping, scaling or deformation of the original images; (2) intensity-driven techniques, which maintain the spatial configuration of the anatomical structures but modify their image appearance. The second type of augmentation seems particularly relevant for the M&Ms as it may increase the variability in image appearance, with the hypothesis that this may lead to improved adaptation to varying imaging protocols and scanner vendors. Two teams performed data augmentation using only spatial transformations (P4, P6). Eleven teams additionally implemented intensity-based transformations using one of two main approaches: (i) standard image transformations such as histogram matching, blurring, change in brightness, gamma and contrast, or addition of Gaussian noise (P1-P3, P7-P8, P10, P13) (see 2.3 for a visualization of a subset of these transformations on a training slice); (ii) advanced image synthesis by using generative adversarial networks (GANs) (P5, P8, P14) or variational auto-encoders (VAE) (P12). For the latter one, the generation of synthetic images for the unseen vendor D is not feasible since it was not included in the training. Note that the majority of the teams participating in the challenge (10 out of 14) relied solely on data augmentation of the training sample to address the domain-shift problem posed by the M&Ms challenge.

Additionally, some teams (P1-P3, P9, P13) applied test-time augmentation techniques, which consist of passing to the model two or more transformed versions of the same inference image to obtain several predictions. These predictions are then combined to obtain one final outcome, usually by averaging them. This method has

been shown to improve the final performance in small data size scenarios and a net improvement with a scale effect that depends on the model architecture [144].

2.3.3 Domain adaptation

Of all participants, only three teams (P4, P6, P10) implemented a method to explicitly address the differences in the image distributions between the unseen and trained vendors. At training, P4 constructed a classifier to distinguish between scanner vendors and used it to modify the training images (through error propagation) until the classifier could not distinguish between the domain. In other words, this method resulted in training images and a trained model that are less dependent on the specific vendors. P6 and P10 proposed to train two models simultaneously with shared features, one for segmentation and one for classification, such that the classification loss is high while the segmentation loss is low, generating features that are robust to vendor-specific variations as well as optimal for segmentation.

TABLE 2.7: Training and inference time, and hardware used, for all participating methods. h, m, s and Mil. stand for hours, minutes, seconds and millions, respectively.

Team	Training time	Inference time (s)	Model parameters (Mil.)	GPU (NVIDIA)
P1	60 h	26	30	Titan XP
P2	48 h	4.8	30	Tesla V100
P3	96-120 h	n/a	30	Tesla V100
P4	6 h	0.35	36	RTX 2080
P5	11 h	10.4	33	GTX 1080 Ti
P6	15 m/epoch	10	28	Tesla V100
P7	8 h	0.0022	6	GTX 1080 Ti
P8	8 h	10	9	Titan V 12GB
P9	96 h	1.2	30	GTX 1080 Ti
P10	10 h	1	4	Tesla K20
P11	11 h	4.48	38	Tesla P100 12GB
P12	3.4 h	0.014	18	GTX 1080 Ti
P13	3 h	0.087	20	GTX 2080 Ti
P14	n/a	15	24	Titan X GPU

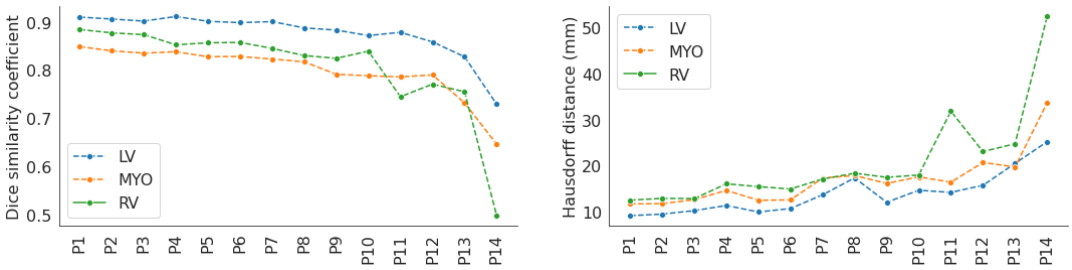


FIGURE 2.4: Weighted average DSC and HD for all participating methods, according to equation (2.5).

2.4 Results

As shown in Table 2.4, a balanced dataset across the four vendors was prepared for evaluating the final submissions (40 CMRs per vendor, total 160 datasets). In this

section, we analyse the obtained results per (1) team, (2) vendor, (3) clinical centre, and (4) show some qualitative results. For analysing the obtained results, we also implemented two baseline models to better appreciate the added value of the data augmentation and domain adaptation techniques used in this challenge:

B1: A 2D UNet without any data augmentation as described in the original reference [137], trained with weighted cross-entropy loss.

B2: The nnUNet pipeline, with a 2D UNet module and default parameters as given in [67] (the best fold according to the validation set was selected).

In particular, B2 differed from those in P1-P3 in that it only included one architecture type [2D UNet] and $\pm$180 degrees rotations, flippings, scalings, deformations, gamma transformations and test-time augmentation as data augmentation. In contrast, P1, P2 and P3 methods included further augmentation techniques such as histogram matching, noise addition, brightness modification, contrast modification and pseudo-label generation by label propagation in time-space.

2.4.1 Analysis per team

Fig. 2.4 displays the results of the challenge for all participants according to two evaluation metrics (DSC and HD). It can be seen that the curves are flat for about half of the participating teams, which indicates comparable performances overall. Note that these methods (P1 to P7) are also the ones that performed better than the baseline methods and we hypothesize that the other models (P8 to P14) suffered from some form of over-fitting (see also the shapes of the curves in Fig. 2.4). Team P1 provided the most consistent results across all metrics. However, the difference to other teams was relatively small and in many cases not statistically significant, as presented in Table 2.8. The three best-performing teams, P1 to P3, used nnUNet as the baseline pipeline, as well as standard intensity-based data augmentation (e.g. blurring, noise addition, histogram matching), but no domain adaptation, showing a significative improvement with respect to the standard nnUNet implementation B2. For similar performance, P5 used an Attention UNet as the backbone architecture and CycleGANs for data augmentation through image synthesis. P4 and P6 also obtained similar performances overall but implemented instead domain adaptation methods and no image-driven data augmentation.

Fig. 2.5 displays the average DSC for all participating teams organised this time per pathology, showing better segmentation performance for healthy cases and dilated cardiomyopathy (DCM), followed by hypertrophic cardiomyopathy (HCM) and other pathologies. It can be seen that the performances of the 14 techniques relative to each other do not change when analysed per pathology.

2.4.2 Analysis per vendor

Fig. 2.6 summarizes the segmentation results for all teams for each vendor separately (A, B, C & D). It can be seen that overall, the differences in the segmentation errors between the vendors are reduced with respect to the results obtained by the two baseline methods as detailed in Table 2.9. Specifically, it can be seen that for the baseline methods, there is a loss of accuracy of up to -6% in the segmentation of images from vendors C and D compared to A and B. However, this loss is reduced, for example, to -1.5% for P1 (e.g. from DSC = 0.92 for vendor A to 0.90 in vendor C and D, for the LV), -2.1% for P2 (e.g. from DSC = 0.87 in vendor B to 0.82 in vendor

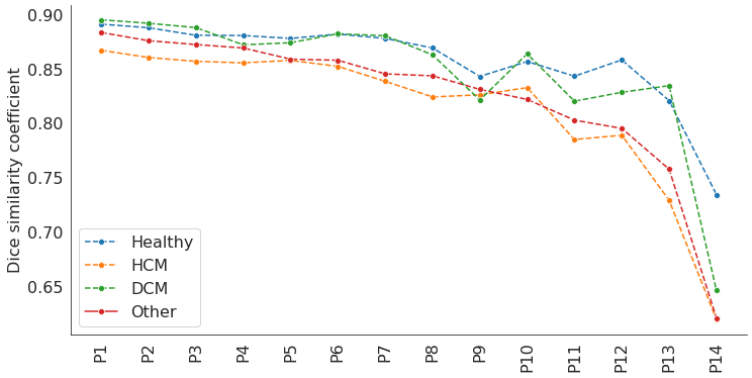


FIGURE 2.5: Average DSC for all participants for the most common pathologies in the dataset. HCM and DCM stand for hypertrophic and dilated cardiomyopathy, respectively.

TABLE 2.8: DSC and HD for the final submissions of all participants and the two baseline models. Boldface numbers are the best results for each column and blue numbers are non-significantly lower results when compared to the P1 results (p-value > 0.01 for the Welch's t-test). HD is measured in millimeters.

Method	ED						ES					
	LV		MYO		RV		LV		MYO		RV	
	DSC	HD	DSC	HD	DSC	HD	DSC	HD	DSC	HD	DSC	HD
P1	**0.939**	**9.1**	**0.839**	**12.8**	**0.910**	**11.8**	**0.886**	**9.1**	**0.867**	**10.6**	**0.860**	**12.7**
P2	0.938	9.3	0.830	12.9	0.909	12.3	0.880	9.5	0.861	10.8	0.850	13.0
P3	0.935	9.5	0.825	13.3	0.906	12.3	0.875	10.5	0.856	11.6	0.844	13.0
P4	**0.939**	11.3	0.826	15.2	0.886	15.4	0.884	11.4	0.856	14.0	0.829	16.7
P5	0.931	10.0	0.816	13.7	0.893	14.3	0.877	9.8	0.850	11.3	0.827	15.2
P6	0.927	11.2	0.815	14.0	0.892	13.6	0.877	9.7	0.852	11.1	0.834	15.0
P7	0.933	13.4	0.812	17.1	0.876	15.7	0.867	14.0	0.839	18.2	0.815	18.1
P8	0.922	15.5	0.809	18.0	0.867	16.6	0.857	17.5	0.836	17.2	0.802	19.1
P9	0.914	12.1	0.768	17.2	0.850	17.5	0.853	12.0	0.814	15.2	0.794	17.0
P10	0.905	13.6	0.772	17.2	0.876	16.2	0.848	15.5	0.820	17.5	0.809	19.6
P11	0.913	14.5	0.776	17.8	0.791	30.7	0.851	13.0	0.809	14.5	0.732	32.9
P12	0.889	16.0	0.785	22.1	0.814	22.1	0.835	14.2	0.808	18.9	0.758	22.0
P13	0.896	15.7	0.761	17.9	0.820	21.0	0.772	23.0	0.721	20.2	0.698	29.5
P14	0.797	21.9	0.668	31.6	0.552	49.1	0.716	25.8	0.673	33.0	0.517	52.0
B1	0.918	12.9	0.801	15.5	0.881	15.7	0.866	11.5	0.842	12.6	0.817	16.3
B2	0.930	10.8	0.817	15.7	0.889	14.8	0.863	13.2	0.835	14.8	0.818	16.8

TABLE 2.9: DSC results stratified by vendor and heart substructure. The last two columns are the average DSC loss for vendors C and D with respect to the combined average DSC results from vendors A and B.

	Vendor A			Vendor B			Vendor C			Vendor D			DSC % loss for ven. C	DSC % loss for ven. D
	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV		
P1	**0.923**	**0.857**	**0.887**	0.915	**0.876**	**0.888**	0.903	**0.842**	**0.884**	0.909	**0.838**	**0.882**	-1.7	-1.6
P2	0.919	0.848	0.885	**0.916**	0.872	0.887	0.899	0.834	0.876	0.903	0.827	0.871	-2.0	-2.4
P3	0.915	0.843	0.877	0.914	0.868	0.879	0.894	0.827	0.873	0.898	0.824	0.870	-2.0	-2.1
P4	0.908	0.831	0.864	0.913	0.867	0.879	**0.906**	0.833	0.870	**0.918**	0.833	0.816	-0.9	-2.4
P5	0.912	0.834	0.869	0.910	0.859	0.870	0.891	0.817	0.819	0.903	0.820	**0.882**	-3.8	-0.8
P6	0.912	0.837	0.880	0.912	0.858	0.877	0.893	0.816	0.861	0.892	0.823	0.833	-2.6	-3.4
P7	0.891	0.804	0.820	0.904	0.859	0.870	0.898	0.821	0.838	0.908	0.817	0.853	-0.7	+0.1
P8	0.889	0.821	0.817	0.900	0.854	0.877	0.880	0.799	0.842	0.889	0.815	0.802	-2.3	-2.9
P9	0.879	0.765	0.800	0.889	0.816	0.827	0.881	0.787	0.831	0.885	0.797	0.829	+0.5	+1.0
P10	0.894	0.812	0.860	0.887	0.822	0.841	0.849	0.753	0.803	0.877	0.796	0.865	-6.1	-0.8
P11	0.885	0.781	0.778	0.899	0.846	0.846	0.875	0.787	0.773	0.869	0.758	0.650	-3.3	-9.8
P12	0.831	0.769	0.795	0.909	0.860	0.867	0.859	0.786	0.792	0.847	0.771	0.690	-3.1	-8.3
P13	0.820	0.712	0.684	0.885	0.823	0.858	0.868	0.779	0.803	0.762	0.650	0.691	+2.5	-12.1
P14	0.805	0.668	0.492	0.872	0.818	0.794	0.822	0.740	0.703	0.528	0.456	0.147	+2.3	-50.9
B1	0.908	0.834	0.861	0.901	0.850	0.865	0.863	0.790	0.800	0.894	0.813	0.870	-6.0	-1.3
B2	0.905	0.832	0.860	0.902	0.846	0.857	0.890	0.806	0.836	0.886	0.821	0.861	-2.7	-1.3

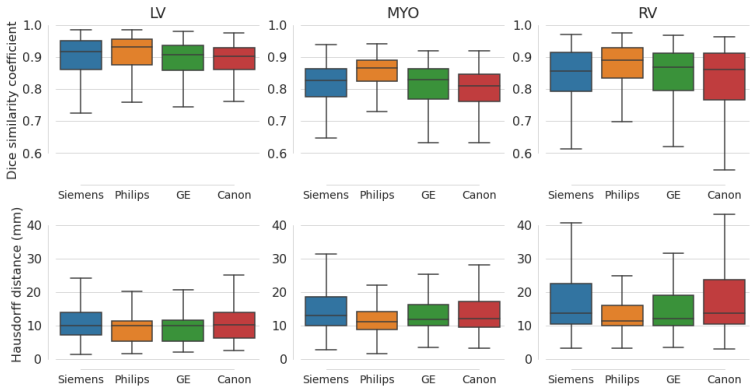


FIGURE 2.6: Boxplots with vendor-wise results for DSC and HD when all participants' predictions are considered. Vendors are presented in order: Siemens (A), Philips (B), GE (C) and Canon (D).

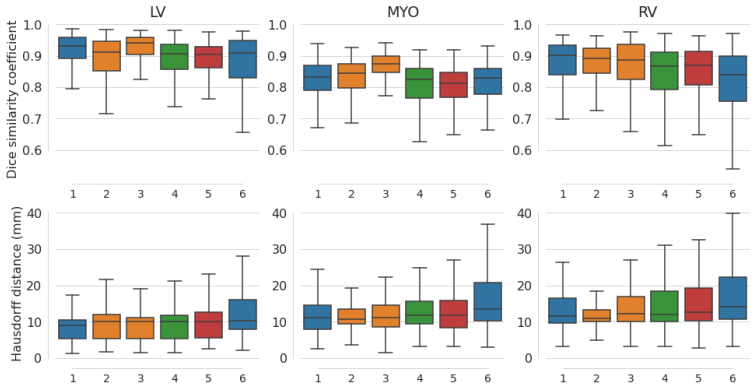


FIGURE 2.7: Boxplots with centre-wise results for DSC and HD when all participants' predictions are considered. The same colour-coding as in Fig. 2.6 is used for scanner vendors.

D, for the RV), and almost to 0% for P7. This indicates that while there is a need for further research to bring segmentation accuracy in unseen and unlabelled vendors at the same level as the one obtained in trained vendors, data augmentation and data adaptation enable to close the gap and improve the generalizability of deep learning models.

2.4.3 Analysis per centre

In the previous subsection, centres were combined in the analysis despite having different machines or scanning protocols. In doing so, possible variabilities between centres using the same scanner may be overstated, making it necessary to consider also Fig. 2.7, where the segmentation results are summarized according to the six clinical centres. Here too, it can be seen that there remains some degree of variation in the segmentation of the CMR images from the different centres. In more detail, there is a decrease in segmentation accuracy between centres 1 and 6 even though their images are from the same scanner vendor A. However, this difference can be explained by two facts: 1) the scanners in these two centres are different models and have different field strengths, as shown in Table 4.2, and 2) all the 75 datasets included during training for vendor A were from centre 1 (Spain) and none from

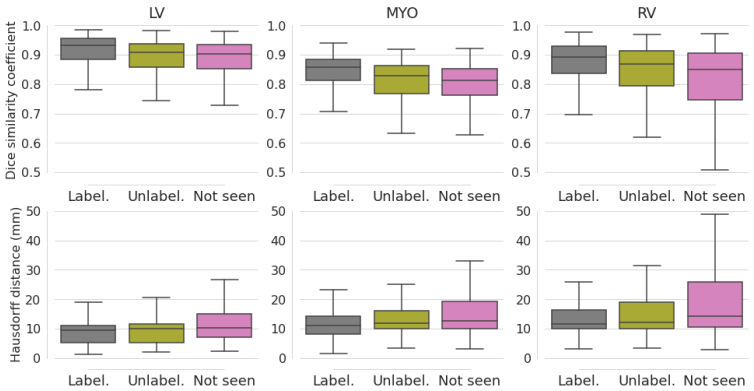


FIGURE 2.8: Boxplots for DSC and HD results for centres that had labelled samples in the training set, unlabelled samples in the training set and no samples at all.

centre 6 (Canada). In this case, even though the images are from the same vendor, differences in scanner specifications resulted in the lack of generalizability. In contrast, images from both centres 2 and 3 were included in the training of vendor B, which resulted in segmentation accuracies for these two centres that are comparable. Finally, the datasets from centres 4 and 5 correspond to vendors C and D, respectively, which were not included in the training, which explains the loss of accuracy compared to centres 1, 2 and 3. In Fig. 2.8, the results are grouped for all centres according to their inclusion (or not) in the training. It can be seen that the segmentation accuracy is the highest for centres that are part of the training together with their labels, followed by those with images but no labels, and finally, the performance is the lowest and most variable for images from fully unseen centres. This result confirms the need for further developments to optimise the generalizability of deep learning solutions in future tools for cardiac image segmentation.

2.4.4 Qualitative results

Fig. 2.9 presents the effect of the slice position in the final segmentation DSC for the top three performing teams, quantifying the loss of accuracy, especially prominent in the apical and basal slices. To illustrate this, Fig. 2.10 provides some visual examples from team P1 to further show the added value of the implemented techniques, as well as their limitations when applied to unseen vendors. In the two examples above, the segmentation techniques enabled us to accurately identify the cardiac boundaries even though these imaging protocols were not included in the training set. However, in the two examples below, despite the use of data augmentation and domain adaptation, the models were unsuccessful in the segmentation of these unseen cases and diverged more notably from the ground truth in basal slices. These examples illustrate the need for future work to further improve the generalizability of deep learning models in cardiac image segmentation.

2.5 Discussion

In this paper, we presented a comprehensive analysis of a range of deep learning solutions for the automated segmentation of multi-centre, multi-vendor and multi-disease CMR datasets. Roughly speaking, the 14 participants in the challenge developed varying workflows combining a baseline neural network, intensity-based

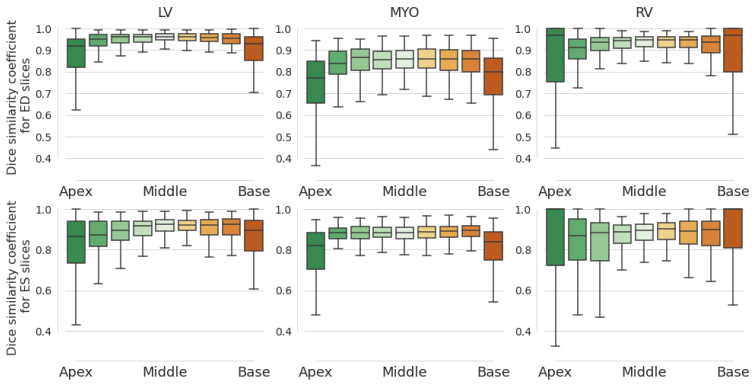


FIGURE 2.9: Boxplots for DSC results for the top 3 performing methods depending on different cardiac structures (LV, MYO and RV) and different slice positions for both ED and ES. The apex and the base are defined as the last and first annotated slices, respectively. The middle slice is the slice located in between the apex and base slices. The remaining slices are defined based on their relative position with respect to the middle slice.

and/or spatial data augmentation, and in some cases a data adaptation strategy. In addition to a relatively large sample of 175 cases for training, the authors were given a total of seven attempts to optimise the parameters and characteristics of their models during the validation process, to ensure an optimal design of the solutions.

2.5.1 Analysis of the methods

The obtained results, first of all, indicate that data augmentation, though its primary purpose is to increase training size and reduce over-fitting, can perform well in addressing some of the differences in image appearance between vendors. In particular, by varying the parameters and types of intensity transformations (e.g. histogram matching, contrast modification, noise addition, image synthesis), one can generate new training images that enhance the generalizability of the models. As an example, one can look at the performance of the baselines models B1 and B2 and augmented models, such as P1, P2 and P3. While for the baseline models, the results do not differ significantly for specific cases, such as at ES, P1-P3 used many more data augmentation types, such as histogram matching, noise addition, brightness modification and contrast modification, and obtained a more marked improvement (*e.g.* the DSC for the myocardium at ES increased from 0.84 for B1 to 0.86 for P1, the DSC for the RV at ES increased from 0.81 for B1 to 0.84 for P3). This indicates the added value of more advanced image-driven data augmentation for multi-vendor image segmentation as well as that the domain shift between different scanners or protocols can be potentially solved by using an exhaustive set of image transformations during training. However, the results also clearly show that the obtained segmentations remain generally more stable in trained vendors compared to unseen vendors, as intensity-driven data augmentation alone cannot enable full coverage of the variety of imaging protocols that can exist across clinical centres.

As for domain adaptation, while it is theoretically suitable for multi-vendor image segmentation, as it can adapt on the spot to the imaging distribution of unseen images, it did not result in better segmentations than when using exhaustive data augmentation alone. In fact, the three first techniques in the ranking did not use any domain adaptation, though it is important to reiterate that the first seven solutions

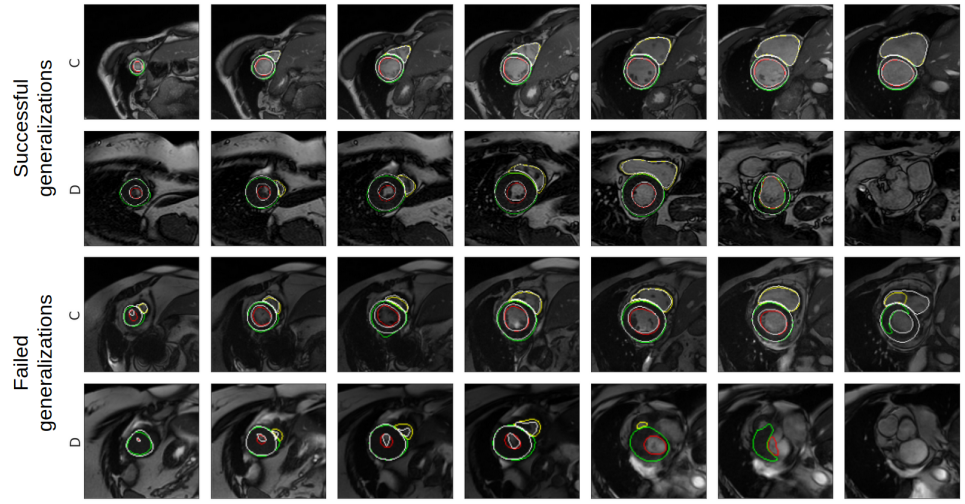


FIGURE 2.10: Prediction examples for method P1 for vendors C (GE) and D (Canon). The top two rows show satisfactory results, while the two bottom rows present some errors in the final contours. Colour correspondence: left ventricle endocardium (red), left ventricle epicardium (green) and right ventricle endocardium (yellow). Ground truth is drawn in white colour.

obtained relatively similar results overall. It is worth noting that the choice of the baseline model may play a role, as again the first three techniques used the same model, namely the nnUNet. Finally, while the results indicate the potential of data augmentation and domain adaption, they also show that there is still a loss in segmentation accuracy when segmenting labelled versus unlabelled or unseen image samples. Note also that training and testing a model on two datasets from the same vendor does not guarantee good generalizability. This is particularly true if the two sets of images are from two different centres and scanner types, such as 1.5T (e.g. centre 1) and 3T (e.g. centre 6) as shown in Figure 2.7.

The results also show that advanced workflows integrating, for instance, data augmentation or generative adversarial networks, are not guaranteed to lead to robust segmentations. In fact, half of the submitted techniques had a lower performance than the two baselines implemented for comparison. This shows that overfitting remains a challenge that requires special attention during the calibration and validation of complex deep-learning solutions for cardiac image segmentation, in particular in the presence of highly heterogeneous data.

Lastly, the presented methods show a vast diversity in hardware performance, with training times ranging from 6 to 100 hours and inference times from tenths of seconds to almost half a minute. However, the amount of training and inference time does not correlate well with the final accuracy, indicating excessive use of computational power for some techniques. For example, the methods implemented by P1 and P2, despite using the same baseline model as P3, needed around half the time for training and obtained slightly better results (1.2% average improvement in DSC), while P4 used around one tenth of computing time for similar loss of accuracy with respect to P1 (1.6% average loss in DSC). Furthermore, clinical centres usually lack dedicated hardware for deep learning models thus increasing, even more, the segmentation time. In this sense, a good equilibrium between accuracy and processing time needs to be attained, with methods such as P4 serving as a good example with

competitive performance and a prediction rate of around 3 images per second.

In summary, the main findings are:

a) Exhaustive data augmentation reduced considerably the domain gap, although the results were still more stable within the domains used during training.

b) Domain adaptation did not result in better performance when compared to nnUNet models trained with spatial and intensity-driven data augmentation.

c) Complex workflows did not always lead to better results, resulting sometimes in excessive use of computing resources.

2.5.2 Analysis of the segmentation results

Compared to other publicly available and annotated multi-structure (LV, MYO, RV) datasets in the field of CMR segmentation, M&Ms is the largest as well as the most diverse (375 cases from four vendors, six centres and three countries, vs. 150 cases for ACDC from one centre). However, given that ACDC is an established database, we selected to use its contouring SOP in this challenge to derive standardized annotations for the community, as well as to enable the combination of these datasets in future studies.

Note that our study, while it focuses on multi-scanner generalizable segmentation, confirms several of the results already obtained by the ACDC challenge and other previous works. Specifically:

a) The segmentations at ED were more accurate than at ES for LV and RV cavities, but not for the myocardium, which becomes thicker and therefore easier to segment when the heart contracts.

b) The segmentation accuracy according to the DSC was the highest for the LV blood pool, followed by the RV and MYO, in this order, but it was the lowest for the RV for the distance-based measures, given its shape complexity.

c) The segmentation accuracy was at its maximum at the mid-ventricular slices, while the performance decreased for the apical and basal slices, where there is higher variability and complexity.

On average, the best-performing method in this challenge obtained 0.88 as DSC and 11 mm as HD versus the values 0.93 and 9 mm obtained in the ACDC challenge, respectively, with the greatest difference shown at ES. This gap can be easily explained by the single-centre nature of the ACDC studies in comparison to a multi-centre scenario in this work, although other effects such as the training size may play a role and should be assessed (150 vs. 100 studies, respectively).

2.5.3 Future work

In addition to the results and analyses presented in this paper on multi-scanner cardiac image segmentation, we also provide the M&Ms dataset open-access for the community, which can be downloaded from the M&Ms website[6]. It represents one of the most heterogeneous datasets ever compiled in cardiac image analysis, comprising CMRs from a variety of imaging protocols and cardiology units, and including a range of cardiovascular diseases as distinct as coronary heart disease, cardiomyopathies, abnormal right ventricle or myocarditis. We thus hope the dataset

[6] www.ub.edu/mnms

will be of high value for the community to address several research topics in the field, such as multi-scanner image registration, multi-structure segmentation, cardiac quantification, motion analysis and image synthesis.

It is important to note that a follow-up challenge is being organised on multi-centre, multi-vendor and multi-disease cardiac diagnosis. The diagnoses for the 375 cases are being gathered from the different hospitals in a legally compliant manner and the clinical information will be made available after the end of the next challenge, thus allowing the community to work on cardiac image analysis as well as on computer-aided diagnosis in a multi-centre setting. Note that the participants had less than three months to implement, optimise and test their techniques, which did not allow them to go beyond the existing state-of-the-art techniques in data augmentation and domain adaptation. With more time at their disposal beyond the constraints of the challenge, we expect that researchers will have a valuable resource with the M&Ms dataset to investigate, develop and test new theories and frameworks for addressing the difficulties posed by domain-shift in cardiac image analysis.

2.5.4 Conclusions

The M&Ms challenge is the first study to evaluate a range of deep learning solutions for the automated segmentation of multi-centre, multi-vendor and multi-disease cardiac images. The results show the promise of existing data augmentation and domain adaptation methods but also call for further research to develop highly generalizable solutions given the inherent heterogeneity in cardiac imaging between centres, vendors and protocols. More generally, there is a need for more research and development to realise the much-needed shift from single-centre image analysis towards multi-domain approaches that will enable the wider translation and usability of future artificial intelligence tools in cardiac imaging and clinical cardiology.

Chapter 3

Domain shift for contrast-enhanced imaging segmentation

Domain shift in deep learning for contrast-enhanced imaging

This chapter contains material from:

> Sendra-Balcells, Carla and Campello, Víctor M. and Martín-Isla, Carlos and Viladés, David and Descalzo, Martín L. and Guala, Andrea and Rodríguez-Palomares, José F. and Lekadir, Karim. "Domain generalization in deep learning for contrast-enhanced imaging." *Computers in Biology and Medicine* 149 (2022): 106052. Doi: 10.1016/j.compbiomed.2022.106052.

3.1 Introduction

3.1.1 Problem and motivation

Over the last years, the domain shift problem has attracted increased attention in the medical image analysis community [56]. Several studies have evaluated the level of generalization of deep learning techniques across domains [24]. For example, a recent challenge on this topic was organized in the cardiac magnetic resonance imaging (MRI) domain at the 2020 Medical Image Computing & Computer-Assisted Intervention Conference (MICCAI 2020), in collaboration with six Spanish, German and Canadian clinical centres. Entitled "Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation (M&Ms)", the study demonstrated that single-centre, single-vendor neural networks do not generalize naturally when segmenting cine-MRI images with distinct imaging domains [15]. The lack of generalizability of neural networks to unseen domains limits their clinical applicability at scale. Thus far, several approaches have been attempted to address this problem in non-contrast imaging, such as methods based on extensive spatial- and intensity-based data augmentation [25], the use of synthetic images from generative models [76], explicit domain adaptation (by forcing the model to learn a similar representation across domains) [30, 120, 141], transfer learning [28, 79] and meta-learning [86, 95]. However, it is unclear whether such approaches can improve generalizability in the case of complex imaging modalities, such as contrast-enhanced imaging, which is the subject of this paper.

In many clinical applications, contrast-enhanced imaging is applied to further improve the visibility of internal body structures and lesions in MRI [16], Computed

Tomography [122] or Ultrasound [118] imaging. For example, late gadolinium enhancement MRI (LGE-MRI) is an essential imaging modality for several applications such as angiography [136], neuroimaging [40], oncology [115], hepatology [161] and cardiology [155]. Contrast-enhanced imaging is faced with additional challenges, compared to non-contrast imaging, due to the intensity heterogeneity arising from the accumulation of the contrast agent in the target areas and the artefacts introduced, which reduce the quality of the images and modify the data distributions. Furthermore, the time between contrast injection and image acquisition can vary greatly between patients and centres, typically between 7 minutes up to a total of 10 minutes, resulting in differences in contrast wash-out and image formation. As a result, the final image appearance, both globally and locally, can have marked differences as clearly illustrated in Figure 3.1 based on images from four different hospitals. At the same time, the limited numbers of available LGE-MRI datasets in existing open-access cohorts compared to non-contrast MRI images, combined with legal and organizational obstacles across centres and countries, has made access to interoperable multi-centre LGE-MRI datasets more difficult. Hence, there is a need for new tools for generalizing single-domain, single-centre deep learning models across new unseen domains and clinical centres in contrast-enhanced imaging such as in LGE-MRI.

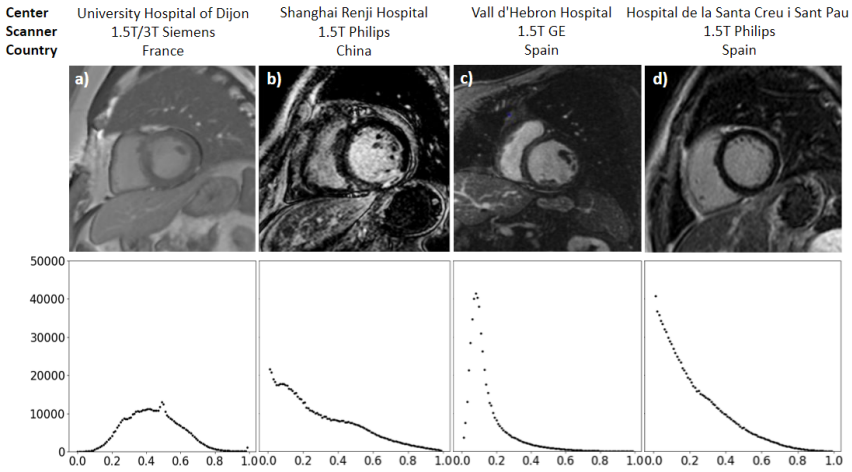


FIGURE 3.1: Four LGE-MRI cardiac images acquired in four different hospitals, together with the average intensity distribution of each dataset. Each histogram has a very different shape and shows marked variability between centres in terms of intensity distributions.

3.1.2 Goals and contributions

In this paper, we present an exhaustive evaluation of deep learning techniques to achieve generalizability to unseen clinical centres for contrast-enhanced imaging. To this end, several techniques are investigated, optimised and systematically evaluated, including data augmentation, domain mixing, transfer learning and domain adaptation. To demonstrate the potential of domain generalization for contrast-enhanced imaging, the methods are evaluated for ventricular segmentation in cardiac LGE-MRI [34]. For this important clinical application, existing deep learning techniques have been almost systematically trained and validated with an LGE-MRI sample from a single clinical centre ([78, 171, 172, 181]). As a result, while many research and commercial tools are already in use for non-contrast cardiac MRI, image

segmentation in cardiac LGE-MRI still relies on labour-intensive manual delineation in clinical practice. Our work is based on a unique multi-centre cardiac LGE-MRI dataset acquired with three distinct scanner vendors (Siemens, Philips and General Electric) in four hospitals located in three countries (France, Spain and China).

3.2 Methods

In this section, an end-to-end pipeline is investigated for generalizable image segmentation in multi-centre LGE-MRI datasets. It is applied for deep learning-based segmentation of the left ventricle (LV), including the blood pool and the myocardium, in multi-centre LGE-MRI cardiac images. To this end, four different approaches are explored to enhance the generalizability across clinical sites of existing deep neural networks for LGE-MRI segmentation, as schematically represented in Figure 3.2. These include:

1. Data augmentation techniques to artificially extend the data distribution captured by the trained models.
2. Image harmonisation to align the data distributions of the training and testing images.
3. Transfer learning to adjust the neural network to the new clinical centre based on very few unseen images.
4. Multi-centre models directly trained with data from multiple clinical centres, which are used for comparative evaluation of the different generalization mechanisms.

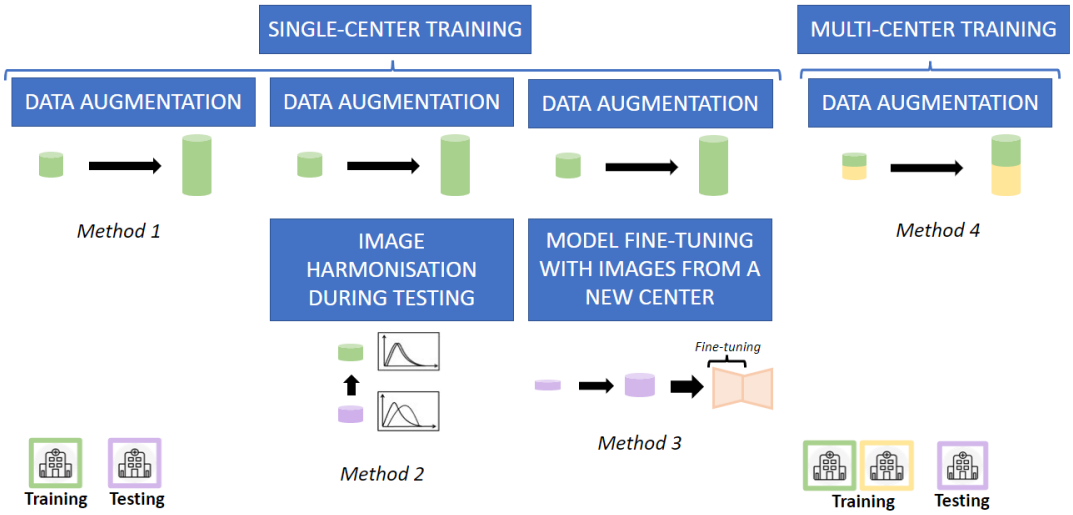


FIGURE 3.2: The four different approaches implemented in this work to enhance the generalisability of LGE-MRI segmentation models across distinct clinical sites.

We confirm that all experiments were performed in accordance with relevant guidelines and regulations.

TABLE 3.1: Details of the multi-centre LGE-MRI datasets and characteristics of the acquired images used in this work. Imaging time = Acquisition time after contrast injection.

Dataset	Clinical centre	Country	MRI scanner	Imaging time (mins)	In-plane resolution (mm)	Slice thickness (mm)	Number of slices	Sample size
EMIDEC	University Hospital of Dijon	France	1.5T and 3T Siemens	10	1.37-1.88	8-13	4-10	100
MSCMR	Shanghai Renji Hospital	China	1.5T Philips	-	0.75	5	10-18	45
VH	Vall d'Hebron Hospital	Spain	1.5T GE	10	1.48-1.68	10	8-15	41
STPAU	Sant Pau Hospital	Spain	1.5T Philips	7-10	1.18	5	18-24	30

3.2.1 Datasets

The multi-centre and multi-vendor dataset used in this study consists of 216 cardiac LGE-MRI datasets acquired in four different clinical centres as detailed in Table 3.1. Two out of four samples are publicly available datasets from France and China, while the two other samples correspond to new LGE-MRI images acquired in two different hospitals in Spain. The subjects have been scanned by using a range of scanner vendors from Siemens, Philips or General Electric (GE). In addition to having distinct intensity distributions as observed in Figure 3.1, the multi-centre LGE-MRI images also differ in the image resolution (0.75-1.88 mm), slice thickness (5-13 mm), and acquisition time after contrast injection (7 to 10 minutes). The samples from each clinical site are described in more detail in the next subsections.

EMIDEC dataset: University Hospital Dijon, France

This dataset was compiled as part of the automatic Evaluation of Myocardial Infarction from Delayed-Enhancement Cardiac MRI challenge (EMIDEC) [82]. The EMIDEC volunteers included 33 healthy and 67 diseased subjects, for a total of 100 studies. The data was acquired at the University Hospital of Dijon, France, using Area 1.5 T as well as Skyra 3T Siemens MRI scanners. Slice thickness and in-plane spatial resolution varied greatly, being comprised of between 8 and 13 mm and 1.37 and 1.88 mm, respectively. The manual segmentation of the LV blood pool and myocardium was performed by a cardiologist with over 10 years of experience. It is the largest of the four samples and hence it was used as the reference sample for training the single-centre neural networks.

MSCMR dataset: Shanghai Renji Hospital, China

The MSCMR dataset was obtained from the Multi-sequence Cardiac MR Segmentation Challenge and it comprises a total of 45 patients suffering from various cardiomyopathies ([179], [180]). The images were acquired at the Shanghai Renji Hospital, China, which will allow us to evaluate generalizability across countries as well as continents in this study. Compared to EMIDEC, the MSCMR dataset has a higher image resolution (in-plane resolution = 0.75 mm, slice thickness = 5 mm for all scans) and all images were acquired with a 1.5 T Philips scanner. The manual delineations were initially performed by trainees and later validated by expert cardiologists.

VH dataset: Vall d'Hebron Hospital, Spain

The VH dataset consists of 41 LGE-MRI datasets acquired at the Vall d'Hebron University Hospital, located in Barcelona, Spain. In addition to covering a new geographical location, namely Spain, the VH sample has several differences with EMIDEC and MSCMR, including the disease group (MI) and the MRI scanner (1.5 T GE scanner). Manual annotations of the LV boundaries were generated by a trained rater using the cvi42 software. The study was approved by the ethics committee of the Vall d'Hebron Hospital and written informed consent was obtained from all participants.

STPAU dataset: Sant Pau Hospital, Spain

The STPAU dataset comprises 30 LGE-MRI cases acquired at the Sant Pau Hospital in Barcelona, Spain. While the clinical centre is located in the same region as for the VH sample, the dataset covers a different disease group (ischemic and non-ischemic cardiomyopathy) and was acquired using an MRI scanner from a different vendor (Philips Achieva 1.5T) and a higher-resolution imaging protocol. Furthermore, the time delay between contrast injection and image acquisition varies between 7 and 10 minutes, which adds extra variability. The manual annotations were also performed using cvi42, as in the previous case. All patients signed the informed consent, the study protocol was approved by the Ethical Committee for Clinical Research of our region, and it follows the ethical guidelines of the Declaration of Helsinki.

3.2.2 Single-centre model with data augmentation

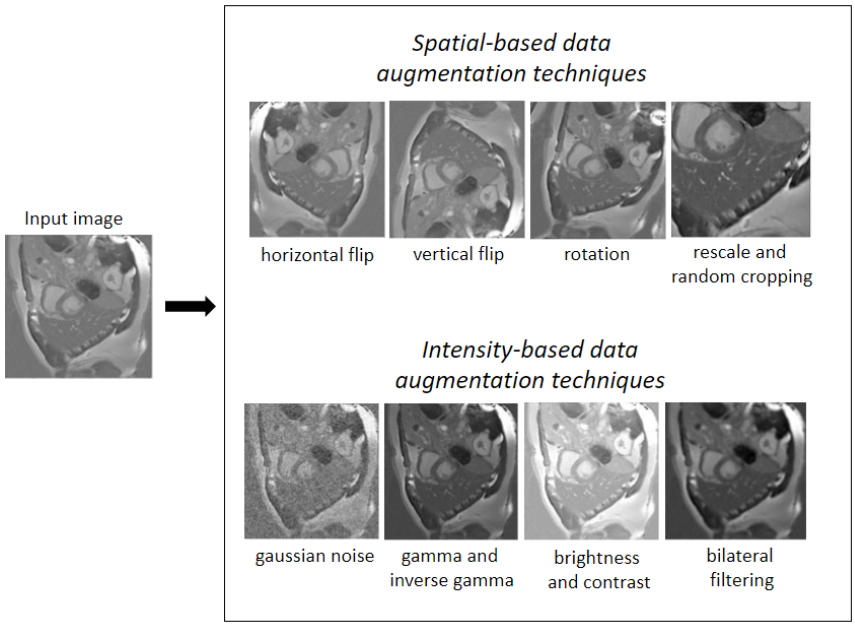


FIGURE 3.3: Both spatial and intensity-based data augmentation techniques are applied together with a probability of 0.2 each. From only one slice many samples can be generated, increasing the size of the original dataset significantly.

In this work, we first investigate the potential of data augmentation to enhance the generalizability of LGE-MRI segmentation models (Method 1 in Figure 3.2). Data

augmentation has been widely used to create more robust neural networks by increasing the size as well as the heterogeneity of training samples synthetically. However, the promise of data augmentation is yet to be examined for LGE-MRI, where there is higher complexity due to inherent variability in scar characteristics and contrast appearance.

In this work, we investigate several operators for data augmentation in the context of LGE-MRI as illustrated in Figure 3.3 and described as follows.

1. **Spatial-based data augmentation**: In addition to the natural variability between cardiac anatomies, especially across countries and ethnic groups, patients undergoing LGE-MRI typically suffer from regional remodelling of the ventricles due to the presence of scar tissue. Hence, spatial-based data augmentation is proposed using the following operators:

    - Horizontal and vertical flips to generate images with different orientations.
    - Random rotations of up to ± 30 degrees, to simulate different positions of the heart.
    - Random rescaling in the [0.75, 1.88] mm range so that the model can process images and hearts that vary in size. This range is defined by the minimum and maximum voxel size of our multi-centre dataset.
    - Random cropping, such that the training images have the same dimensions of 256x256 pixels but with a variation in the position of the heart in the image.

2. **Intensity-based data augmentation:** Because the LGE-MRI appearance can vary between images acquired using different MRI scanners and scanning protocols, such as due to differences in acquisition time after contrast injection, we implemented the following intensity-based data augmentation techniques:

    - Bilateral filtering to generate blurred and less detailed copies of the original images.
    - Gaussian noise with a standard deviation ranging between [0, 0.03] to generate artificial noise and image artefacts.
    - Gamma and inverse Gamma function with magnitude [0.7, 1.5] to generate synthetic images with different lighting.
    - Brightness and contrast with magnitude [-0.5, 0.5] to support brightness and contrast variations in the training images.

Each data augmentation technique is applied with a probability of 0.2 during the training of the model. Then, this data augmentation pipeline is evaluated by measuring the final generalization ability of the network (Method 1 in Figure 3.2). Table 3.2 summarizes the split of the data used for the training, validation and testing of the model in each experiment.

3.2.3 Image harmonisation at testing

While the data augmentation operations focused on improving model generalizability at training, we propose to apply image harmonisation at the testing stage to further reduce the discrepancies between the multi-centre LGE-MRI images (Method

TABLE 3.2: Number of subjects for each of the four datasets used during the training, validation and testing phases when data augmentation is implemented in a single-centre setting.

Dataset	Training	Validation	Testing
EMIDEC	68	17	15
MSCMR	24	6	15
VH	21	5	15
STPAU	12	3	15

TABLE 3.3: Number of samples used for the training and validating each CycleGAN model built to harmonise the imaging properties from the different clinical centres.

Dataset		Training		Validation	
Source	Target	Source	Target	Source	Target
EMIDEC	MSCMR	24	24	6	6
EMIDEC	VH	21	21	5	5
EMIDEC	STPAU	12	12	3	3
MSCMR	VH	21	21	5	5
MSCMR	STPAU	12	12	3	3
VH	STPAU	12	12	3	3

2 in Figure 3.2). Image harmonisation enables the transformation of testing LGE-MRI images from a new clinical centre such that their intensity distribution matches as much as possible the imaging characteristics of the single centre used to train the baseline neural network. Concretely, two main image harmonisation techniques were implemented:

1. **Histogram matching**: It consists of transforming the testing images from the unseen centre such that the histogram of the pixel intensity values is superimposed as much as possible with the corresponding histogram extracted from the training images from the training clinical centre. The transformation from the testing data (B: target data) to the training data (A: source data) is illustrated in Figure 3.4(i).

2. **CycleGAN**: Another strategy to address the domain shift between multiple centres is domain adaptation, which can be used to learn the image translation from the source domain to the target domain. To this end, we choose to implement a CycleGAN architecture [178], based on an unpaired image-to-image translation. Given that CycleGAN uses cycle consistency, it would learn the translation from the target domain (B) to the source domain (A), and vice versa (Figure 3.4(ii)). Both target-to-source and source-to-target generators are saved in each implementation, reducing to 6 the number of implementations needed. The number of samples used to train each of the CycleGAN models is summarised in Table 3.3, adjusting for each case the percentage of images from each centre so that it is adequately balanced (50% source and 50% target data).

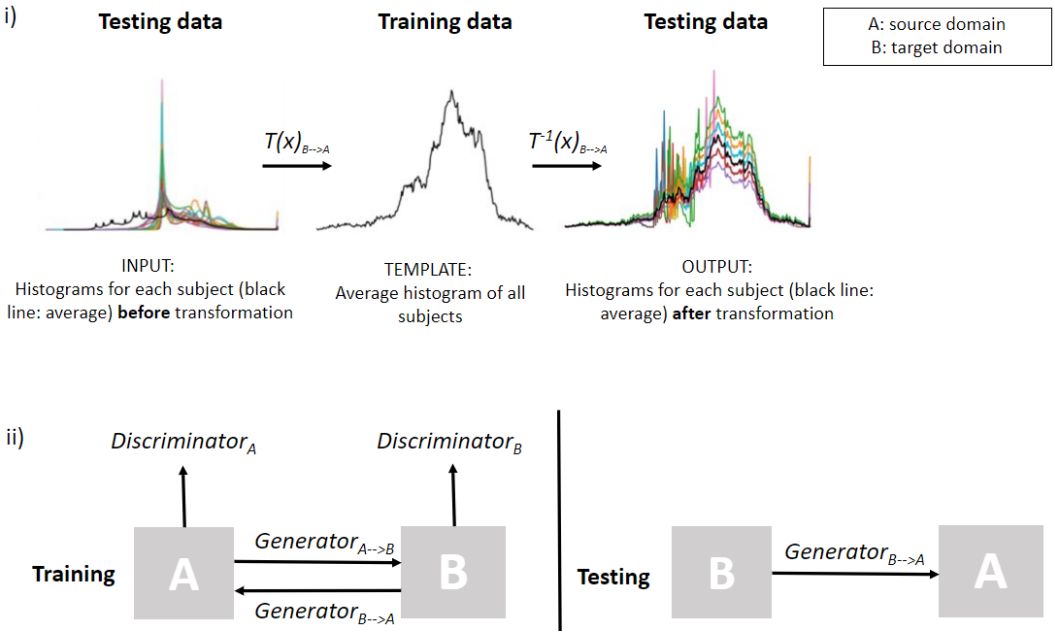


FIGURE 3.4: Schematic illustration of the image harmonisation techniques used in this work to make the intensity distributions from the different clinical sites as aligned as possible. First, histogram matching is used to learn a transformation of the histogram of each image from the unseen clinical sites (B) onto the original training clinical centre (A). Second, CycleGAN architecture is used to learn the mapping between the training and the testing clinical centre.

3.2.4 Transfer learning from the original to the new clinical site

Another strategy investigated in this work to improve the scalability of single-centre models consisted of applying the so-called transfer learning paradigm, by fine-tuning specific layers of the neural network with a reduced number of LGE-MRI images from the new clinical site (Method 3 in Figure 3.2). The approach has shown promise for multi-centre image segmentation in cardiac cine-MRI [102] but is yet to be demonstrated for multi-centre LGE-MRI imaging, where there is increased variability. The following steps are implemented in this work:

1. Initiate the training of the neural network with the EMIDEC dataset, then evaluate the minimum number of fine-tuned layers, in both the decoder and the encoder, that are needed during transfer learning to obtain the maximal segmentation performance on the new LGE-MRI datasets from the remaining clinical centres.

2. Compare the previous results with the segmentations obtained based on a multi-centre model directly trained each time with images from two clinical centres (EMIDEC and the new centre).

3. Estimate the minimum percentage of images needed from the second clinical centre during the fine-tuning to obtain the desired level of performance.

4. Implement the same approach from the previous point but this time by using a model pre-trained on a large dataset (n=350) from cine-MRI (M&Ms dataset), to evaluate transfer learning from a related cardiac MRI modality for which data is abundantly available.

TABLE 3.4: List and number of samples used for training and validating multi-centre models in this study.

Dataset	Training	Validation
EMIDEC	42	10
EMIDEC+MSCMR	21+21	5+5
EMIDEC+VH	21+21	5+5
EMIDEC+MSCMR+VH	14+14+14	3+3+3
EMIDEC+MSCMR+VH+STPAU	11+11+11+11	3+3+3+3

3.2.5 Multi-centre model

A fourth and last modelling strategy, i.e. training the neural networks directly from multiple centres (Method 4 in Figure 3.2), is used for comparative evaluation of the three extended single-centre models described in the previous section, i.e. enriched with data augmentation, image harmonisation and transfer learning. In this study, we investigated the number of new centres/domains that are needed to bridge the domain gap in LGE-MRI segmentation, by using a balanced dataset with the same number of subjects for each multi-centre data combination, namely EMIDEC, EMIDEC + MSCMR, EMIDEC + VH, EMIDEC + MSCMR + VH and ALL centres. The samples used for training the multi-centre models in each combination of datasets/centres are listed in Table 3.4. In all experiments, the same testing dataset is used for comparative evaluations (n=15).

3.2.6 Baseline workflow

Pre-processing

Min-max normalization is used after cropping the image to keep the same intensity range in images from the same dataset.

Post-processing

Post-processing is applied to all predictions generated by the model by keeping only the largest connected component of the segmentation volume. This step is commonly used in medical image segmentation, especially in organ imaging, to help in the detection of false positives.

Network architecture

As a baseline model, a U-Net architecture is implemented to perform the LV boundary segmentations in LGE-MRI based on some of the modifications proposed by [67] for improved model training as follows. First, Leaky ReLU is used as the activation function, and then instance normalization is applied after each hidden convolutional layer to stabilise the training. Deep supervision is included to allow gradients to be injected deeper into the network and facilitate the training of all layers. Furthermore, a 2D architecture is selected as it is suitable to address the differences in slice thickness between clinical centres, as well as slice misalignment due to respiratory and cardiac motion artefacts. The encoder and decoder architecture of the model are illustrated in Figure 3.5.

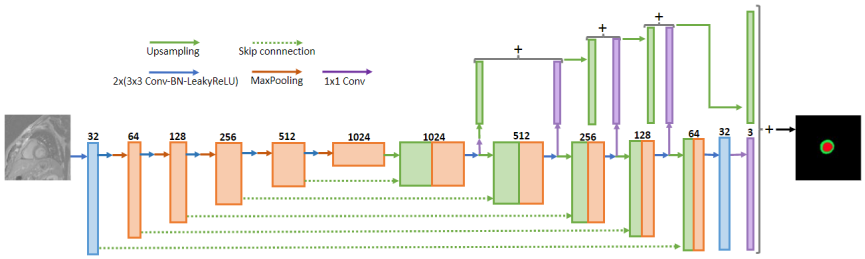


FIGURE 3.5: U-Net architecture composed by 6 layers, increasing progressively the number of feature maps until 1024. Additionally, deep supervision layers are included in the decoder.

Implementation details

PyTorch is the open-source machine learning library for Python used for the implementation of the model learning process. Stochastic gradient descent (SGD) optimization is performed with Adam and the batch size of 16 slices is constrained by the 8 GB of memory of the NVIDIA GeForce RTX 2080 Ti GPU. The learning rate is kept to $1{\cdot}10^{-3}$ during every training, while the dice and cross-entropy losses are calculated at every iteration to optimise the network parameters. The neural network is trained 250 epochs each time and takes half an hour approximately to converge. During testing, each LGE-MRI image segmentation takes less than one second. The main criterion followed to split each dataset into subgroups is 80% for the training and 20% for the validation while keeping 15 subjects for the testing.

Performance evaluation

For all experiments and results, the performance of each method will be assessed with the average 3D Dice Coefficient (DC), which calculates the overlap ratio between the automatically generated and ground truth segmentations. The measure is estimated by:

$$\mathrm{DC} = \frac{2 \cdot (X \cap Y)}{X + Y} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}, \tag{3.1}$$

where X and Y are the set of pixels from the automated and true labels of the target structures, while TP, FP, and FN are the corresponding true positives, false positives and false negatives, respectively.

3.3 Results

This section presents detailed experimental results obtained by evaluating and comparing the different strategies proposed for enhancing model generalizability in LGE-MRI segmentation. Four experiments are proposed to study model generalizability: (1) effect of data augmentation, (2) image harmonisation, (3) transfer learning and (4) multi-centre training. The results are summarized in Table 3.5, where a similar limited generalization performance is achieved for experiments (1) and (2) while experiments (3) and (4) show a significant improvement. Each experiment is analysed in detail next.

TABLE 3.5: Dice score coefficient for the different domain generalization experiments performed. The results are averaged over five runs of models. All models used EMIDEC for training. In experiment 3, every model is transferred to the corresponding target centre and in experiment 4, every model is trained with EMIDEC and a training set from the corresponding target centre. Standard deviation is presented as a subscript for five independent runs of each model.

Test Centre	*Experiment 1* Effect of data augm. (single-centre training)			*Experiment 2* Image harmonisation		*Experiment 3* Transfer learning	*Experiment 4* Multi-centre training
	No augm.	Spatial	Spatial & intensity	CycleGAN	Hist. match.		
EMIDEC	$0.85_{0.05}$	$0.88_{0.03}$	$0.78_{0.08}$	-	-	-	-
MSCMR	$0.30_{0.15}$	$0.62_{0.19}$	$0.72_{0.12}$	$0.64_{0.17}$	$0.78_{0.07}$	$0.87_{0.03}$	$0.89_{0.03}$
STPAU	$0.54_{0.16}$	$0.61_{0.12}$	$0.68_{0.09}$	$0.70_{0.08}$	$0.68_{0.08}$	$0.85_{0.04}$	$0.85_{0.04}$
VH	$0.32_{0.21}$	$0.26_{0.23}$	$0.62_{0.13}$	$0.53_{0.17}$	$0.58_{0.12}$	$0.78_{0.11}$	$0.82_{0.06}$

3.3.1 Experiment 1: Effect of data augmentation

In the first experiment, the added value of the different types of data augmentation is evaluated, including spatial and intensity-based data augmentations. Figure 3.6 shows the comparative results obtained by three different models: (i) a single-centre model without data augmentation (blue line), (ii) a single-centre model with spatial data augmentation (orange), and (iii) a single-centre model enriched with both spatial and intensity based data augmentations (green). As observed in the results, data augmentation consistently improves the segmentation performance for LGE-MRI independently of the clinical centre used for training, increasing the DC value up to 0.6 units when compared to the baseline model without data augmentation. Furthermore, the results in Figure 3.6 show that while most of the improvement can be achieved by spatial data augmentation (orange line), intensity-based data augmentation adds value to the approach, in particular when training on the largest sample (EMIDEC) and testing on smaller samples (MSCMR, VH and STPAU). Having demonstrated the added value of data augmentation, all subsequent experiments are performed using spatial- and intensity-based data augmentation.

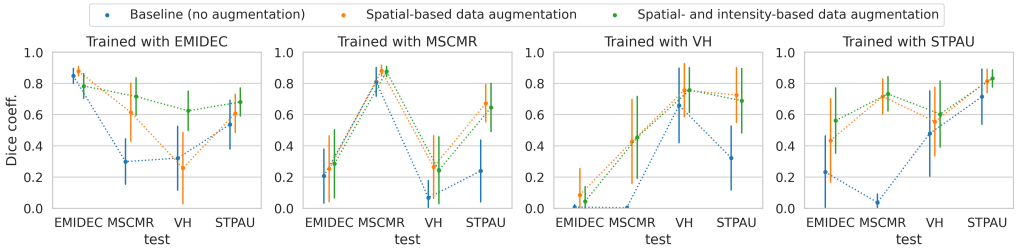


FIGURE 3.6: Comparison of the subject-wise Dice coefficient obtained for models trained with a single-centre with and without data augmentation, including spatial- and intensity-based augmentations. Models are tested on subjects from the testing set for every centre. The results are averaged over five different runs of each model.

3.3.2 Experiment 2: Effect of image harmonisation

Here, the impact of image harmonisation is evaluated when applied to match the intensity distribution and appearance of LGE-MRI images from a new clinical centre to that of the training set. Specifically, we evaluate three approaches, namely (i) the baseline model with data augmentation from Experiment 1 but without any normalization, (ii) the baseline model with histogram matching, and (iii) the baseline model with CycleGAN normalization. The results are given in Figure 3.7, clearly showing that, overall, the two harmonisation operations (green and orange lines) do not improve significantly the LGE-MRI segmentations over the baseline model without harmonisation (blue line). There are, however, few cases where the mean Dice score is slightly improved, as for histogram matching when training with EMIDEC and testing in MSCMR or when using CycleGAN for models trained with VH.

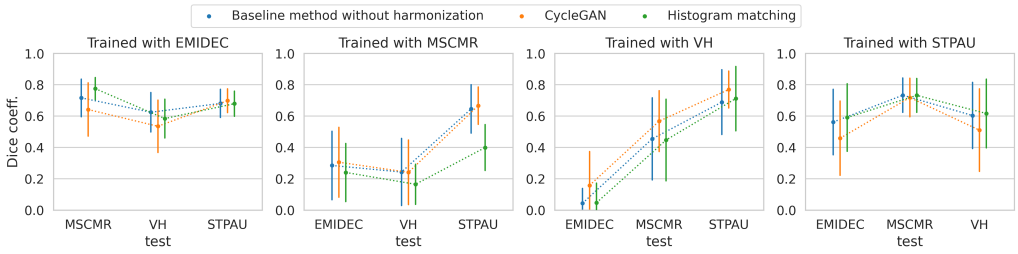


FIGURE 3.7: Effect of histogram matching and CycleGAN harmonisation for LGE-MRI segmentation in unseen clinical centres. X corresponds to each testing centre not included in the training dataset and the results are averaged over five different runs of each model.

3.3.3 Experiment 3: Effect of transfer learning

This section evaluates the potential value of fine-tuning a model pre-trained on a larger dataset (such as EMIDEC) via transfer learning and the effect of the sample size used during the tuning process. Figure 3.8 shows the performance of transfer learning for a model pre-trained with EMIDEC and fine-tuned for each new clinical centre (MSCMR, VH and STPAU). The red and blue lines in the figure show the segmentation accuracy when the fine-tuning is performed on the encoder and decoder of the neural network, respectively, while the remaining parts of the model are frozen. The black line corresponds to a model trained and tested in the same centre. The results show an increase in DC with the number of fine-tuned blocks and the maximum is obtained when 5 or all blocks of the encoder are fine-tuned, reaching nearly the same performance as the single-centre model of the new centre (black line). Furthermore, in Figure 3.9, the single-centre models fine-tuned based on 5 encoding blocks are directly compared to multi-centre models trained based on all images from the original and new clinical centres. Based on the results, fine-tuned models (green bars) –despite being fine-tuned on the new LGE-MRI images– achieve similar segmentation performances to models directly trained from multi-centre image data (orange bars). This shows the potential of transfer learning to adjust and optimise a few layers of the existing single-centre model based on unseen LGE-MRI images from a new clinical centre.

However, transfer learning requires manual annotations of some images from the new clinical sites. Hence, ideally, the number of new annotated images required to suitably adapt the existing model to the new centre should be minimal. In Figure

3.10, we evaluated the impact of the number of new LGE-MRI images used for fine-tuning. The results indicate that the fine-tuning of single-centre models with a small percentage of the target data is sufficient to reach a desirable segmentation accuracy. Such generalization is achieved, for example, in the case of a single-centre model pre-trained with the EMIDEC dataset and fine-tuned using only 10% (about 1-3 subjects) from the new dataset. A similar pattern is found when training with a different modality, as shown by the model pre-trained with cine-MRI images from the M&Ms dataset (grey line), except for better performance when testing on the VH centre.

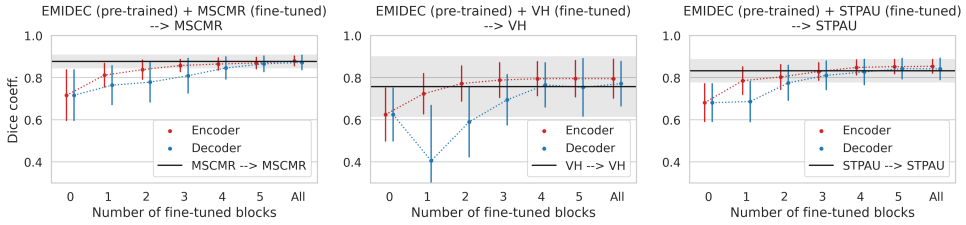


FIGURE 3.8: Evaluation of a single-centre model pre-trained on the EMIDEC dataset and fine-tuned with a new clinical dataset (MSCMR, VH or STPAU). Red: Fine-tuning of several blocks in the encoder. Blue: Fine-tuning of several blocks in the decoder. Black: Model trained from scratch with data from the same centre. The bars and the grey band stand for the standard deviation of the five independent model runs.

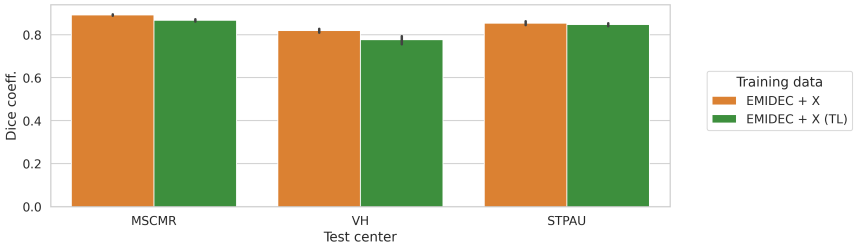


FIGURE 3.9: Model trained from scratch using EMIDEC and a second dataset (X), which can be MSCMR, VH or STPAU (orange). Then, a model was pre-trained with EMIDEC, and fine-tuned and evaluated on X (green). The black bars represent the standard deviation for five independent runs of each model. TL: transfer learning.

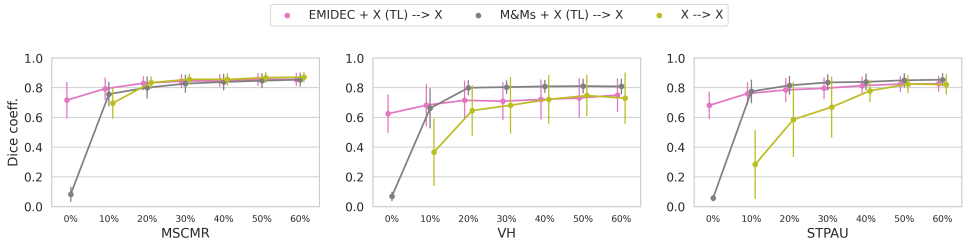


FIGURE 3.10: Impact of sample size (percentage) of a new LGE-MRI dataset used for fine-tuning existing single-centre models for different training datasets (fuchsia and grey) and compared to single-centre models (yellow). Results are averaged over five independent model runs.

3.3.4 Experiment 4: Comparison to a multi-centre scenario

In this last experiment, the added value of training multi-centre models for LGE-MRI segmentation is evaluated by including training images from multiple clinical sites (i.e. from 1 to 4 centres). In Figure 3.11 different combinations of the four datasets considered in this study were explored, either by using a baseline model, data augmentation or histogram matching. As observed in the results, when the model is trained with no data augmentation (baseline), the multi-centre data enhanced the generalization ability as demonstrated by the increase in average DC values and the reduction of the standard deviation. However, when data augmentation is included in the pipeline, no gain is found by adding new clinical sites to the training stage, as the data augmentation alone is sufficiently powerful for training the model with reduced over-fitting when tested in new centres. The results also confirm that histogram matching does not show a significant positive impact on the final performance.

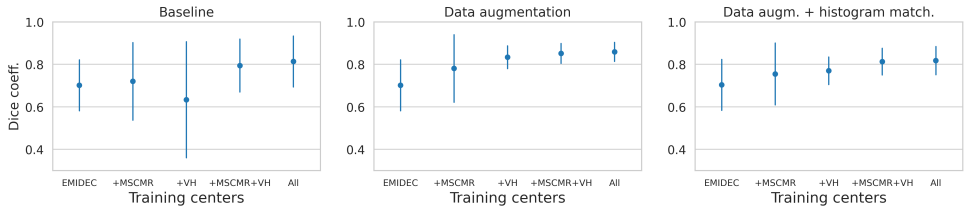


FIGURE 3.11: Average DC achieved by models trained on different combinations of clinical datasets, with and without data augmentation, as well as with histogram matching. The first model is initially trained with the EMIDEC dataset, and then new datasets are included progressively from the three other clinical sites. Results are averaged over five independent model runs.

3.3.5 Qualitative analysis

Finally, we show a qualitative comparison in Figure 3.12 of model predictions (coloured overlay) for selected cases that demonstrate the common mistakes of the models as compared to the ground truth (white delineations). For instance, the first three columns show how data augmentation improves the model's ability to identify and segment the left ventricle while for some cases (like for the second row, with the VH sample), it is still insufficient. The fourth and fifth columns show the effect of the image harmonisation experiments, which help in segmenting failing cases but do not improve significantly the accuracy of the segmentation as observed in the disagreement between ground truth and predictions. Finally, the last two columns show the predictions for transfer learning and multi-centre models, respectively. These final predictions are the most accurate among all the columns, but one can still identify some disagreements in challenging regions annotated with orange arrows where scars can be found.

3.4 Discussion

In this work, several strategies were implemented and evaluated for generalizable segmentation of left ventricular anatomy in multi-centre LGE-MRI. The pipeline was built with the purpose of training single-centre models that can maintain a good

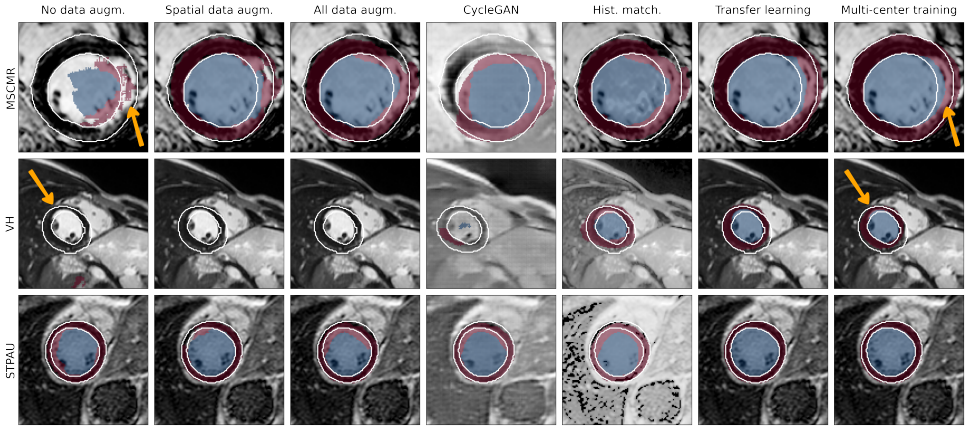


FIGURE 3.12: Qualitative comparison of model predictions for selected slices of test subjects. The ground truth is delineated with white lines while the overlayed colour represents the model prediction. Each row corresponds to a different dataset and each column corresponds to each model as presented in Table 3.5. Challenges regions with scars are highlighted with orange arrows in the first and last columns.

level of performance when used to segment out-of-sample images from new hospitals. The results highlight the importance of using data augmentation, including both spatial and intensity-based transformations, in particular when there is a high domain shift between the training and unseen clinical site, e.g. EMIDEC in our results. After applying adequate data augmentation to existing single-centre models, it was found that neither multi-centre training nor image harmonisation techniques are needed to obtain additional generalizability, confirming the results obtained by [15] in the M&Ms study for a multi-centre and multi-vendor cine-MRI. This finding shows that single-centre LGE-MRI models can generalize well if appropriately enriched with data augmentation, which results in an important practical benefit: Multi-centre training is difficult in practice as there is a lack of labelling harmonisation between centres, in addition to the legal and other obstacles that make difficult cross-site data sharing. Moreover, multi-centre models are still specific to those clinical centres that contributed data, whereas there is a need for models that can generalize well beyond the training data.

Regarding domain adaptation, which theoretically is a promising solution, existing research has shown that histogram matching could lead to hidden noise in some images after the post-processing [47], while CycleGANs would typically require substantial training data from the new clinical centre to achieve a good model performance. In addition to data augmentation, the results demonstrated that transfer learning can positively impact the model performance across sites. This method is based on the fine-tuning of an existing model initially pre-trained on a single-centre dataset and adjusted with a few datasets from the new clinical site. The obtained results indicate that fine-tuning the first 5 blocks of the encoder of the model with the 10% of the dataset, ranging from 1 to 3 subjects, is sufficient to achieve the desired LV segmentation accuracy in LGE-MRI. For example, a neural network pre-trained based on the EMIDEC dataset and fine-tuned with one subject/image only from STPAU (DC: 0.76 ± 0.07) performs similarly when compared to a model trained from scratch with 100% of the STPAU images (DC: 0.79 ± 0.13). In terms of computational time, the first model is completely trained in half an hour and the posterior

fine-tuning requires only 5 minutes.

In addition to transfer learning focused on LGE-MRI, we evaluated the potential of fine-tuning a pre-existing model trained on larger cine-MRI datasets from the M&Ms dataset, which consists of 350 training images. Despite the different imaging characteristics between cine and LGE-MRI images, in particular the additional presence of scar tissue and contrast enhancements in the LGE-MRI images, the results showed that such cross-modality transfer learning results in enhanced generalizability. This can be easily explained by the fact that such pre-trained multi-centre and multi-disease model encodes additional inter-subject variability which aids generalizability also in multi-centre and multi-disease LGE-MRI context.

Finally, to illustrate the success of data augmentation and transfer learning to build models with good generalization ability, Figure 3.13 provides two examples of challenging LGE-MRI cases, with varying imaging and anatomical characteristics. Even though these images are from two different clinical centres and vary greatly in the appearance, size, shape and location of the scar tissues, the proposed enriched models are capable of accurately identifying the LV boundaries consistently across the LGE-MRI examples.

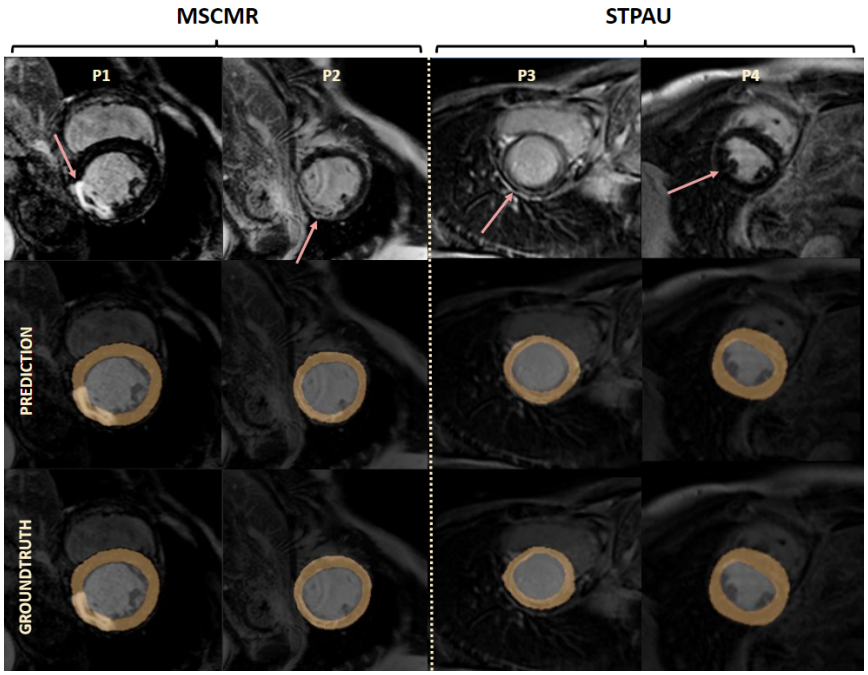


FIGURE 3.13: Challenging cases leading to good model predictions on two patients from two different hospitals. First row: original image, second row: prediction, third row: ground truth. Each of the two columns corresponds to images obtained from MSCMR or STPAU datasets respectively. The red arrows highlight the infarct or scar tissue.

Compared to other multi-centre existing studies, such as the M&Ms challenge that comprises 350 cine-MRI cases, the present multi-centre LGE-MRI study has a lower sample size. This is because the LGE-MRI datasets are less abundant and more difficult to compile for research studies. Nevertheless, the results in this work are generated based on 216 datasets from four clinical centres, three vendors (Siemens, Philips and GE) and three countries from two different continents.

Another limitation is that this work was focused on the segmentation of the LV anatomy and did not consider the more challenging task of segmenting the scar tissues. This is because the clinical annotations for the scar tissues were not available for the two clinical centres in Spain. Future multi-centre studies in LGE-MRI should also investigate the generalizability of neural networks for scar tissue segmentation. However, our work is an important first step in this direction and one that will encourage the development of more generalizable models based on data augmentation and transfer learning, in LGE-MRI but also other cardiac and non-cardiac imaging modalities.

While the proposed framework shows promise for generalizability across multi-centre LGE-MRI datasets with challenging and heterogeneous conditions, it can fail to accurately identify the LV boundaries in a few exceptions. As illustrated in Figure 3.14, several failures have been observed in the presence of low-quality images with artefacts due to suboptimal contrast wash-out or highly complex scar appearance. Furthermore as reported in previous works in cardiac cine-MRI segmentation [25], apical and basal slices are also more error-prone than mid-ventricle slices in LGE-MRI segmentation. Even experienced cardiologists can disagree on the segmentation of the LV borders closer to the apex and base, which generates inter-operator variability that can confuse neural networks, as illustrated in Figure 3.15.

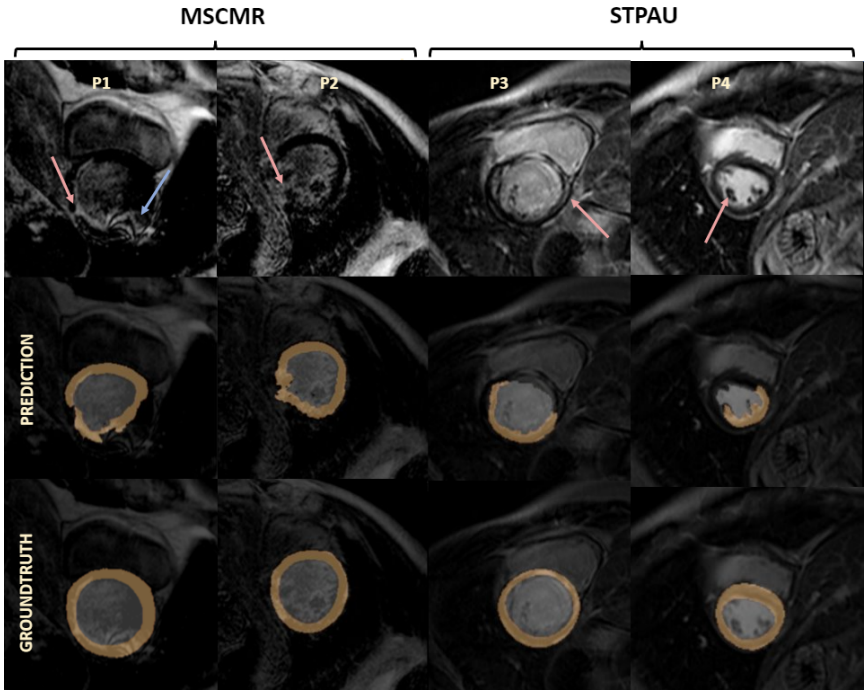


FIGURE 3.14: Segmentation failures obtained due to artefacts and highly complex scars. First row: original image, second row: prediction, third row: ground truth. Each of the two columns corresponds to images obtained from MSCMR and STPAU datasets respectively. The blue arrow shows an image artefact, while the red arrows point to the infarct or scar tissue.

3.5 Conclusions

This work was motivated by the need for new deep learning-based solutions that generalize well across domains, centres and scans, in non-contrast as well as in

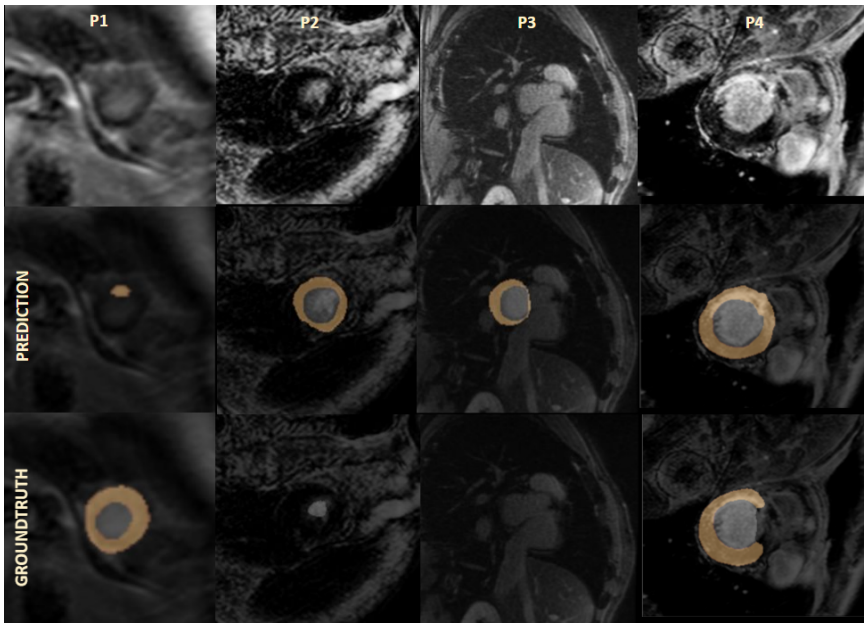


FIGURE 3.15: Examples of segmentation failures obtained at the apical and basal slices. First row: original image, second row: prediction, third row: ground truth. The first and second columns show two similar cases where both apical slices are segmented differently. The third and fourth columns are two heterogeneous segmentation at the basal region.

contrast-enhanced imaging. Data augmentation extended the image distribution in single-centre settings and proved to be an effective technique to generate models with a prominent generalization ability to new clinical centres. In contrast, image harmonisation did not improve the capability of single-centre models when tested on unseen clinical sites. Furthermore, the exploitation of transfer learning based on fine-tuning pre-trained models with as little as one additional subject from an unseen clinical site translated into a substantial improvement in the model's generalizability. This paper showed that single-domain neural networks enriched with suitable generalization procedures can reach and even surpass the performance of multi-centre, multi-vendor models in contrast-enhanced imaging, hence eliminating the need for comprehensive multi-centre datasets to train generalizable models.

Chapter 4

Domain generalization for cardiomyopathy diagnosis

Minimising multi-centre radiomics variability through image normalisation: A pilot study

This chapter contains material from:

> Campello, Víctor M. and Martín-Isla, Carlos and Izquierdo, Cristian and Guala, Andrea and Palomares, José F. Rodríguez and Viladés, David and Descalzo, Martín L. and Karakas, Mahir and Çavuş, Ersin and Raisi-Estabragh, Zahra and Petersen, Steffen E. and Escalera, Sergio and Seguí, Santi and Lekadir, Karim "Minimising multi-centre radiomics variability through image normalisation: a pilot study." *Scientific Reports* 12, 12532 (2022). Doi: 10.1038/s41598-022-16375-0.

4.1 Introduction

For the last decade, there has been a great amount of research devoted to identifying and improving quantitative image biomarkers for precise diagnosis, risk assessment and patient stratification for different pathologies. In particular, radiomics seems to be a promising technique to quantify image-derived biomarkers based on shape, intensity and higher-order texture patterns for a region of interest defined a priori, since it can characterise image patterns that are hardly visible to the naked eye.

These computer-extracted features have the potential to perform an exhaustive analysis of medical images as shown in the literature, predominantly in oncology [1] but also more recently for neurodevelopmental disorders [147] or cardiovascular disease [19]. However, radiomic features have proven to be highly sensitive to changes in scanning protocols and scanner manufacturers, resulting in limited reproducibility [83, 156] (see also the exhaustive reviews by Yip and Aerts [167] and by Traverso et al. [152]) and thus posing an important problem that needs to be solved before implementing these techniques in clinical practice. Despite this, the majority of previous research considered single-institution datasets, due in part to the difficulty in obtaining imaging studies from multiple centres. More recently, several works using multi-centre studies have assessed the robustness of this technique (see for example Raisi-Estabragh et al. [132], for a test-retest study). Several works have proposed harmonisation guidelines for computed tomography (CT) or positron emission tomography in multi-centre scenarios, while no guideline is available for magnetic resonance imaging (MRI), where the lack of a standard intensity grayscale – such as

Hounsfield units in CT – poses further difficulty (see Da-Ano, Visvikis, and Hatt [32] and references therein).

All previous multi-centre MRI radiomics studies focused either on brain or cancer imaging. Due to the lack of multi-centre cardiac imaging radiomics literature, a detailed introduction to brain and cancer imaging is presented. Two types of techniques are used to standardise features across institutions: image- and feature-based transformations.

At the image level, the most common techniques are image intensity normalisation (mean subtraction and division by the standard deviation) or image intensity rescaling to a fixed range (usually from 0 to 1). Other more sophisticated techniques exist, such as bias field correction, isotropic resampling, histogram matching and piecewise linear histogram matching (PLHM). Finally, some techniques are inherently defined for brain imaging and were not considered in this study. Um et al. [156] used T1-weighted MRI brain scans to assess radiomics variability across two different institutions after five image preprocessing techniques were applied, including global and region of interest (ROI) rescaling, bias correction, isotropic resampling and histogram matching. They concluded that histogram matching is the best technique for reducing feature variability and successfully discriminating between different patient subgroups with glioblastoma. Isaksson et al. [65] evaluated the effect of four normalisation techniques on classification performance to identify prostate cancer in T2-weighted MRI. The normalisation method that resulted in the best classification accuracy was the PLHM transformation using intensities from healthy prostate as a reference instead of the whole image to extract landmarks. Finally, Carré et al. [17] standardised brain MRI studies using three different intensity normalisation techniques to find their effects on radiomics robustness, being image intensity normalisation the technique that yielded the best results.

At the feature level, Chatterjee et al. [22] improved the robustness of radiomics from images of primary uterine adenocarcinoma by applying feature normalisation for each institution dataset independently. Orlhac et al. [117], instead, used the empirical Bayes harmonisation method – also referred to as ComBat[69]– to remove inter-centre variability. The transformed features resulted in a sensitivity increase for distinguishing between Gleason grades in prostate cancer studies and in similar distributions for features from brain scans for 1.5T and 3T machines.

In this work, a multi-centre cardiac MRI dataset was considered to analyse the effect of several image- and feature-based normalisation techniques over radiomic features variability and model generalisation across institutions.

4.2 Material and methods

4.2.1 Data and feature extraction

A subset of 218 cardiac magnetic resonance studies from the Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge (M&Ms) dataset was considered [15]. In particular, healthy subjects as well as patients with hypertrophic cardiomyopathy (HCM) were selected from the five available centres. The exact distribution across the five centres is presented in Table 4.1. All the scanners considered had a field strength of 1.5T and the averaged in-plane resolution ranged from 0.85 to 1.45 millimetres. More detailed information about the scanners used can be found in Table 4.2.

Each study consisted of a short-axis cine cardiac magnetic resonance volume. Segmentations of three anatomical ROIs, the left and right ventricle cavities (LV and

TABLE 4.1: Distribution of diseases per centre considered in the analysis

	Creu Blanca	Dexeus	Sagrada Familia	Universitätsklinikum Hamburg-Eppendorf	Vall d'Hebron	Total
Vendor	Canon	General Electric	Philips	Philips	Siemens	
Healthy	14	11	33	32	22	112
HCM	15	5	37	14	25	106

HCM: hypertrophic cardiomyopathy

TABLE 4.2: Average specifications for the studies acquired in the five different centres.

Centre	Vendor	Model	In-plane resolution (mm)	Slice thickness (mm)	Number of slices	Intesities range
Vall d'Hebron	Siemens	Magnetom Avanto	1.32	9.2	12	0 – 1193
Sagrada Familia	Philips	Achieva	1.20	9.9	10	0 – 357
Universitätsklinikum Hamburg-Eppendorf	Philips	Achieva	1.45	9.9	11	0 – 3725
Dexeus	General Electric	Signa Excite	1.36	10	12	0 – 3030
Creu Blanca	Canon	Vantage Orian	0.85	10	13	0 – 14442

RV, respectively) and the left ventricle myocardium (MYO), were provided for two temporal phases, ES and ED. The delineations were revised to follow the same Standard Operating Procedure to avoid the introduction of further bias due to inter-observer variability.

Radiomic features were extracted using the PyRadiomics library [55], version 3.0.1. Before the extraction, all images were resized to match the same spatial resolution of $1 \times 1\ mm^2$, since radiomic features have been shown to intrinsically depend on voxel size and the number of voxels [58]. Fixed bin widths of 25 and 0.05 were used during feature extraction for images before and after normalisation, respectively. This resulted in a good balance between the number of bins and computing requirements. The number of bins after normalisation ranged between 20 and 80, depending on the intensity values for each ROI. Only images without normalisation gave a large variability in terms of number of bins (from 14 to 570).

A total of 100 features were extracted per ROI. They include shape features and first and second-order texture features. In this work, only texture features were used, since shape depends only on the ROI segmentation and not on the image intensity. First-order texture features refer to commonly used statistical metrics to describe the histogram of intensity values such as mean, minimum, maximum, kurtosis, skewness, entropy and energy, among others. Second-order texture features are statistical measures extracted from the four texture matrices considered in this library: Gray Level Co-occurrence Matrix (GLCM), Gray Level Size Zone Matrix (GLSZM), Gray Level Run Length Matrix (GLRLM) and Gray Level Dependence Matrix (GLDM). These features account for different details in the spatial coarseness, variability, heterogeneity and symmetry of textures. A complete list of the features considered is included in the supplementary material.

4.2.2 Normalisation techniques

Four normalisation techniques were considered at the image level:

- R: image intensity rescaling to the range 0 – 1,
- N: image intensity normalisation (mean subtraction and division by the standard deviation),
- HM: histogram matching using scikit-image [160], version 0.17.2,
- PLHM: piecewise linear histogram matching [111], also referred to as Nyúl-Udupa normalisation.

For the histogram matching transformation, an image intensity histogram is interpolated so that it matches a template histogram. In this work, a subject was selected visually from Sagrada Familia as the template after ensuring that the image did not present artefacts. For the PLHM transformation, the code implementation by Reinhold et al. [134] was employed. In this case, a batch of images from one centre was needed to obtain the averaged histogram deciles (landmarks) that were then used as a reference for the transformation of new image histograms. The landmarks were computed for studies from Sagrada Familia. All transformations were applied both to the whole image and at the ROI level, independently. Data from Sagrada Familia were used as a reference since it was the centre with the greatest number of scans.

Regarding feature-based normalisation techniques, the empirical Bayes harmonisation method proposed by Johnson, Li, and Rabinovic [69] Johnson and Rabinovic (ComBat) was considered. This method assumes that the contributions to the final feature values can be separated into biological covariates (*e.g.*, pathology) and centre effects (*e.g.*, different scan manufacturers). Then, the empirical Bayes method is used to estimate the distributions for these terms from the original data and adjust the final feature values to remove centre effects. The ComBat method is robust against outliers and does not need large sample sizes for each centre batch, which makes it a good option for the current study. However, feature distributions are assumed to follow normal distributions for each centre separately, a requirement not always satisfied by the data. For this reason, a quantile transformation (scikit-learn [121], version 0.23.2) had to be applied to all radiomic features for each institution independently before ComBat could be used (we used 20 as the number of quantiles). The Python implementation of ComBat by Fortin et al. [41], available at github.com/Jfortin1/ComBatHarmonization, was used. Five batches were used during the harmonisation process, one for each centre. A parametric adjustment was chosen for fitting the batch effect parameters [69], and the alignment was performed over a virtual reference frame instead of over one of the five batches. No covariates were used along with ComBat harmonisation.

4.2.3 Variability assessment

Radiomics variability across centres was assessed by computing the Jensen-Shannon divergence (JSD) between pairs of feature distributions obtained for healthy subjects within the different ROIs. The HCM pathological group was not considered in this analysis since the possible existence of different HCM sub-groups could introduce uncontrolled bias to the results. Moreover, to avoid redundancy of features in the results, a prior sequential feature selection step was conducted to remove features that showed a square cross-correlation coefficient greater than or equal to 0.9 with

any previous feature following the ordering provided by PyRadiomics (see Table A.1 in the supplementary material). The JSD gives a positive measure of how similar two distributions are, with 0 as the value obtained when the two distributions are identical. A threshold of 0.01 JSD was selected based on the median of the overall distribution as the relative point where changes in feature proportions were to be assessed.

Then, to analyse model generalisation, two tasks were proposed. First, the amount of centre-encoded information after the application of each normalisation technique was measured by training Random Forest (RF) models to identify the source centre for each feature set. The hypothesis was that features with less information about their centre of origin should be more difficult to differentiate and thus, more similar between centres, enhancing the generalisation. Secondly, model generalisation was assessed directly by training RF models for patient classification into healthy or HCM groups, for each normalisation technique. RF were chosen over other techniques due to their simplicity to train and their effectiveness to model non-linear relations between input and output. For all cases, a random seed was fixed before training each model to make the results reproducible.

For the centre identification task, models were trained either with first-order or second-order features as input variables for each ROI separately (LV, MYO and RV). A five-fold cross-validation was used for obtaining an estimate of the average classification accuracy with reduced bias. Notably, in this case, a lower accuracy represents that features carry less centre information.

For the patient classification task, models were trained with a combination of first and second-order features from all three ROIs, so that RF was able to select the most predictive features during training. Five runs of the same cross-validation scheme were considered in this case (five different random seeds) since the variability in models was higher and the accuracy estimates showed greater bias. In particular, these models were trained with features from only one dataset (Vall d'Hebron ($n = 38$) or Sagrada Familia $n = 56$, as these centres had a greater number of samples) and tested on the other four. No feature selection was conducted before model training. The most important features for the best performing models were obtained with the mean impurity decrease method [13], also called Gini importance. For these models, a greater accuracy represents better generalisation. All models were assessed with balanced accuracy, given the imbalanced nature of the dataset.

4.3 Results

4.3.1 Feature variability

Texture features variability showed a great disparity depending on the preprocessing method under consideration (see Fig. 4.1). For both the end-diastole (ED) and the end-systole (ES) frames, the percentage of features with similar distributions across institutions (Jensen-Shannon divergence, JSD, below 0.01) was obtained after the removal of highly correlated features ($R^2 \geq 0.9$). After this step, the amount of first/second order ED (ES) features remaining were 7/38 (8/47) for LV, 9/48 (9/49) for MYO and 9/42 (8/47) for RV (the square correlation heatmaps are shown in Figure A.2 in the supplementary material). The highest percentage of features with similar distributions across institutions was obtained when applying ROI-based PLHM or ROI-based rescaling as shown in Figure 4.1 (with a maximum of 74% for first-order features and of 66% for second-order features). More specifically, these two

methods showed a significant difference in distribution similarity for first-order features when compared to other methods, and only ROI-based histogram matching showed comparable results for second-order ED features (all tests with p-values below 0.01, Mann-Whitney U test).

In contrast, the proportion of features below the 0.01 JSD threshold was the lowest for the methods applied at the whole image level, except for rescaling, and for images without any normalisation (N, HM, PLHM and O in Fig. 4.1). No significant difference was found between these normalisation methods and original images for both ED and ES features (p-values greater than 0.01, Kruskal-Wallis test). The proportion of features below the given threshold was reduced to less than 51%, indicating that feature distributions were less similar for these methods. Additionally, the large standard deviation, represented by the black horizontal bars in Figure 4.1, was associated with differences depending on the ROIs and, especially, on the centre pairs being compared (see Fig. A.1 and A.3 in the supplementary material for a detailed comparison of these factors).

The application of the ComBat harmonisation method had an averaging effect, reducing the proportion of features with similar distributions for the most robust methods in the previous paragraph and increasing it for the least robust methods, as shown in Figure 4.2. Thus, smaller differences were found between methods after the application of ComBat. Specifically, no significant difference was found between four methods for ED features (R.R, R.N, R.PLHM and O in Fig. 4.2), and for three methods for ES features (R.R, R.N and R.PLHM in Fig. 4.2, p-values greater than 0.01, Kruskal-Wallis test).

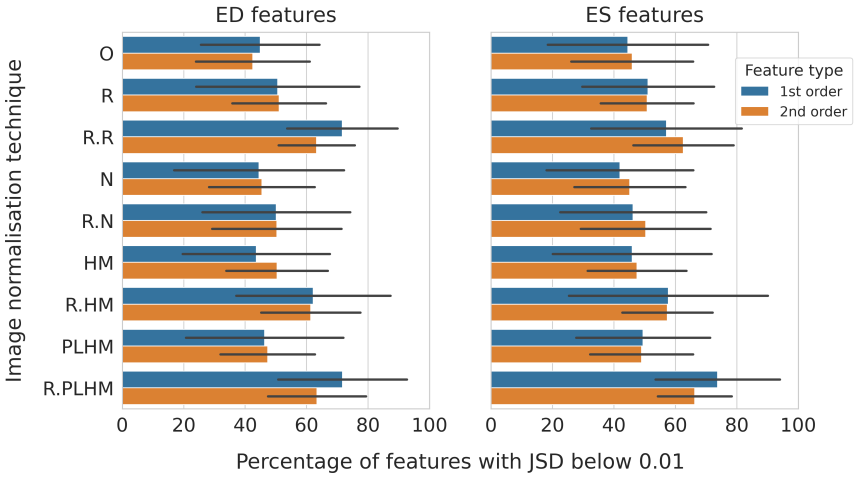


FIGURE 4.1: Percentage of first and second order features below the 0.01 JSD threshold for healthy subjects. Results are averaged over centre pairs and ROI and presented separately for ED and ES frames. Only features with square cross-correlation below 0.9 were considered. The black lines represent the standard deviation. O: original images (without normalisation), R: image intensity rescaling, N: image intensity normalisation, HM: histogram matching and PLHM: piecewise linear histogram matching. An "R." in front of a method means that it is applied at the ROI level.

Among all feature families, Gray Level Size Zone Matrix (GLSZM), Gray Level Run Length Matrix (GLRLM) and Gray Level Dependence Matrix (GLDM) presented the highest dissimilarities among distributions after the application of ROI-based PLHM normalisation in general, as demonstrated by the greater JSD values

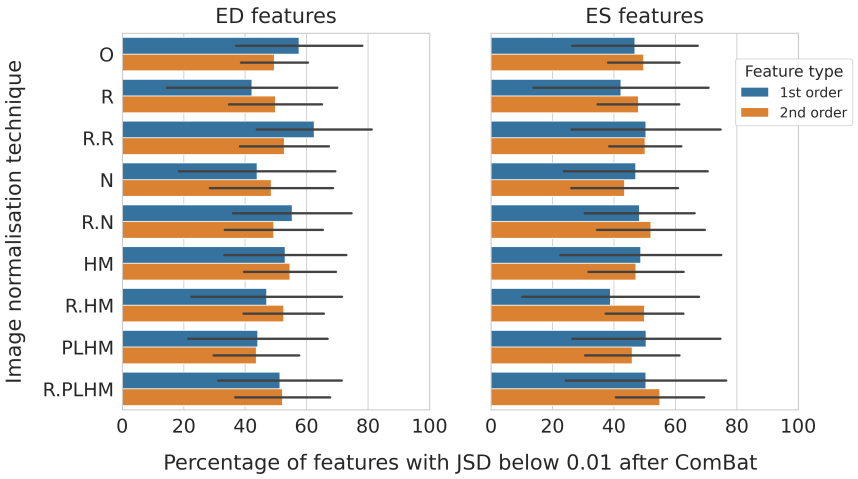


FIGURE 4.2: Percentage of first and second order features below the 0.01 JSD threshold for healthy subjects after the application of the feature-based harmonisation tool ComBat. Results are averaged over centre pairs and ROI and presented separately for ED and ES frames. Only features with square cross-correlation below 0.9 were considered. The black lines represent the standard deviation. O: original images (without normalisation), R: image intensity rescaling, N: image intensity normalisation, HM: histogram matching and PLHM: piecewise linear histogram matching. An "R." in front of a method means that it is applied at the ROI level.

in Table 4.3. Gray Level Co-occurrence Matrix (GLCM) and first-order features obtained the best similarity scores. As noted above, the JSD was averaged over all families of features to an approximate value of 0.011 after the application of ComBat. The features found with the most dissimilar distributions (standard deviation of the JSD distribution greater than 0.01) in both cardiac time frames, ED and ES, before the application of ComBat were zone variance, large area emphasis and large area low grey level emphasis (GLSZM), kurtosis (1st order) and grey level non-uniformity (GLDM). Some examples of the effects of ComBat and PLHM over the different distributions per centre are presented in Figure A.4 in the supplementary material.

4.3.2 Centre identification

When assessing the centre information encoded in the extracted features, second-order texture features carried more information in general than first-order features, as demonstrated by the differences in balanced accuracy for classifiers trained with healthy subjects in Figure 4.3 (orange and blue boxes). Features from original images (without normalisation) were the most discriminative features with testing accuracy above 0.87 ($\pm$ 0.07–0.11) for the three ROIs under consideration and for both feature types, first order and texture features. When comparing normalisation techniques at the whole image level, no clear method showed a greater reduction in the centre information consistently across ROIs and feature types. (Fig. 4.3, orange and blue boxes in the top row).

When normalisation was applied at the ROI level, larger differences appeared depending on the method and the ROI under consideration (Fig. 4.3, orange and blue boxes in the bottom row). Regarding methods that did not use ComBat, ROI-based PLHM consistently reduced the ability of models to infer the centre of origin for each sample for first-order features extracted from LV and MYO, and for

TABLE 4.3: Mean and standard deviation (in parenthesis) for JSD for distributions of features obtained after the application of R.PLHM normalisation on healthy patients. Results are presented separately for ED and ES frames and each feature family before and after the application of ComBat harmonisation. Only features with square cross-correlation below 0.9 were considered. Values are averaged over ROI. Numbers in blue stand for non-significant differences in the JSD distributions when compared to first-order features according to the Mann-Whitney U test at the 0.01 level.

Family	Without Combat		With Combat	
	ED	ES	ED	ES
1st order	0.009 (0.009)	0.008 (0.007)	0.012 (0.011)	0.011 (0.008)
GLCM	0.008 (0.007)	0.009 (0.009)	0.013 (0.012)	0.012 (0.013)
GLDM	0.011 (0.010)	0.010 (0.008)	0.013 (0.013)	0.011 (0.010)
GLRLM	0.012 (0.011)	0.010 (0.007)	0.011 (0.010)	0.011 (0.009)
GLSZM	0.011 (0.011)	0.011 (0.010)	0.013 (0.012)	0.011 (0.010)

R.PLHM: ROI-based piecewise linear histogram matching; ED: End-diastole; ES: End-systole.

second-order features from LV, achieving the lowest performance (p-values below 0.01, Mann-Whitney U test). For the RV, however, three methods (R.R, R.HM and R.PLHM) showed comparable accuracy (p-value greater than 0.01, Kruskal-Wallis test). Finally, ComBat harmonisation was able to remove centre information from features almost entirely for most normalisation techniques and original images, as shown by the red and green boxes in Figure 4.3.

When models were trained only with HCM patients, the general behaviour between methods observed for healthy subjects was reproduced, but the accuracy for identifying the centre was reduced for all methods before using ComBat harmonisation. See Figure A.5 in the supplementary material, for more details.

4.3.3 Generalisation

With regards to the patient classification task into healthy and hypertrophic cardiomyopathy (HCM) groups in unseen centres (see Fig. 4.4), models trained with features from original images (without normalisation) showed the worst performance. With regards to models trained with features from images normalised at whole image level (Fig. 4.4, upper row), N and PLHM methods were significantly better than other methods and performed similarly when trained with studies from Vall d'Hebron, while PLHM was significantly better when trained with studies from Sagrada Familia (all p-values below 0.01, Mann-Whitney U test, after the Bonferroni correction for multiple comparison).

When images underwent ROI-based normalisation, ROI-based rescaling and ROI-based normalisation performed on par and significantly better than other models when trained with Vall d'Hebron studies (p-values below 0.01, Mann-Whitney U

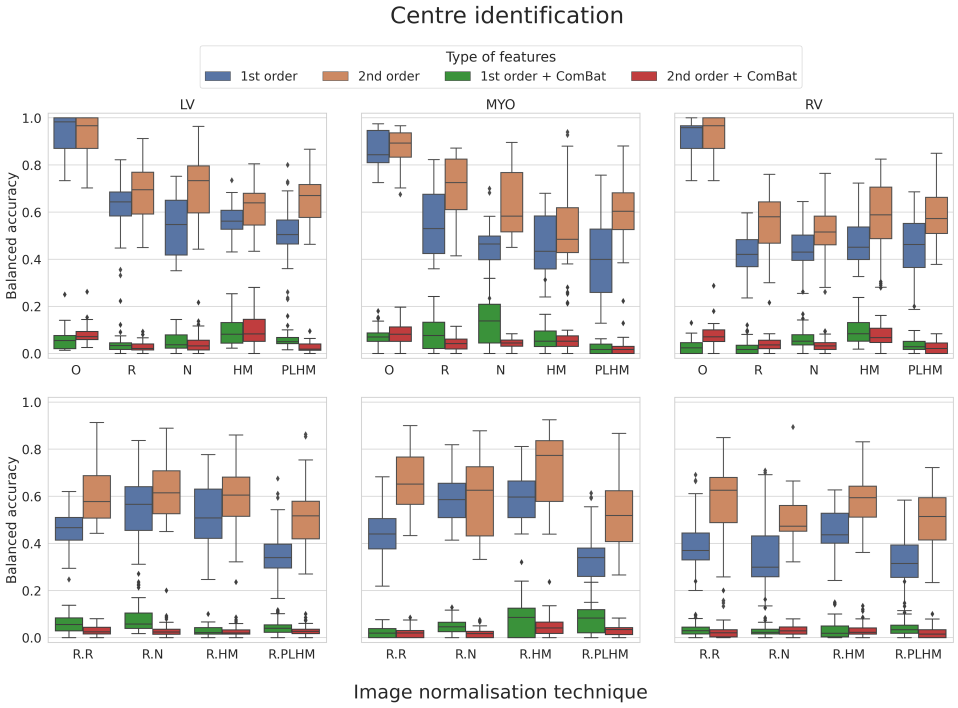


FIGURE 4.3: Balanced accuracy of random forest models when predicting the centre of origin of healthy subjects for first and second-order texture features before and after the application of ComBat harmonisation. The row above corresponds to image preprocessing techniques applied at the whole image level, while in the row below they are applied at the ROI level. O: original images (without normalisation), R: image intensity rescaling, N: image intensity normalisation, HM: histogram matching and PLHM: piecewise linear histogram matching. An "R." in front of a method means that it is applied at the ROI level.

test), while no method was significantly better than others when trained with studies from Sagrada Familia (p-value greater than 0.01, Kruskal-Wallis test).

The application of ComBat reduced the accuracy slightly in general, but the difference was only significant for rescaling, ROI-based normalisation and ROI-based histogram matching when training with Vall d'Hebron studies, and for the whole image and ROI-based histogram matching and ROI-based PLHM when training with studies from Sagrada Familia (p-values below 0.01, Mann-Whitney U test, after Bonferroni correction for multiple comparisons).

For both types of models, trained with Vall d'Hebron and Sagrada Familia studies, the best accuracy was obtained when using features extracted after applying the PLHM transformation and without ComBat harmonisation: $78.3\% \pm 8.4$ and $79.2\% \pm 8.8$, respectively.

In more detail, for models trained with features from Vall d'Hebron studies, the highest accuracy was 0.783 (median: 0.792 [0.745, 0.845]), obtained after PLHM without the application of ComBat. When ComBat harmonisation was used, the highest accuracy was obtained after the application of the same image normalisation technique but was reduced to 0.771 (median: 0.775 [0.694, 0.826]). For models trained with features from Sagrada Familia studies, the best accuracies were again obtained for PLHM and were 0.783 (median: 0.792 [0.728, 0.850]) and 0.762 (median: 0.762

[0.712, 0.811]) before and after the application of ComBat harmonisation, respectively. For these models, features mostly from the myocardium (MYO) were among the most important features for the model prediction according to the Gini importance [13]. The top 20 most important features contained mean and median intensity, kurtosis and skewness (1st order), joint average and autocorrelation (GLCM) and run length non-uniformity and long-run high grey level emphasis (GLRLM).

When comparing the accuracy between validation (same institution) and testing (unseen institutions) sets, models that obtained the highest accuracy on validation generalised worse to new unseen centres (Fig. 4.5). Importantly, models trained with features from ROI-based normalisation methods showed relatively similar generalisation performance among them, even though some suffered from overfitting. Within normalisation methods at the whole image level, features extracted after PLHM obtained the best testing accuracy despite their lower performance in validation when compared to other techniques.

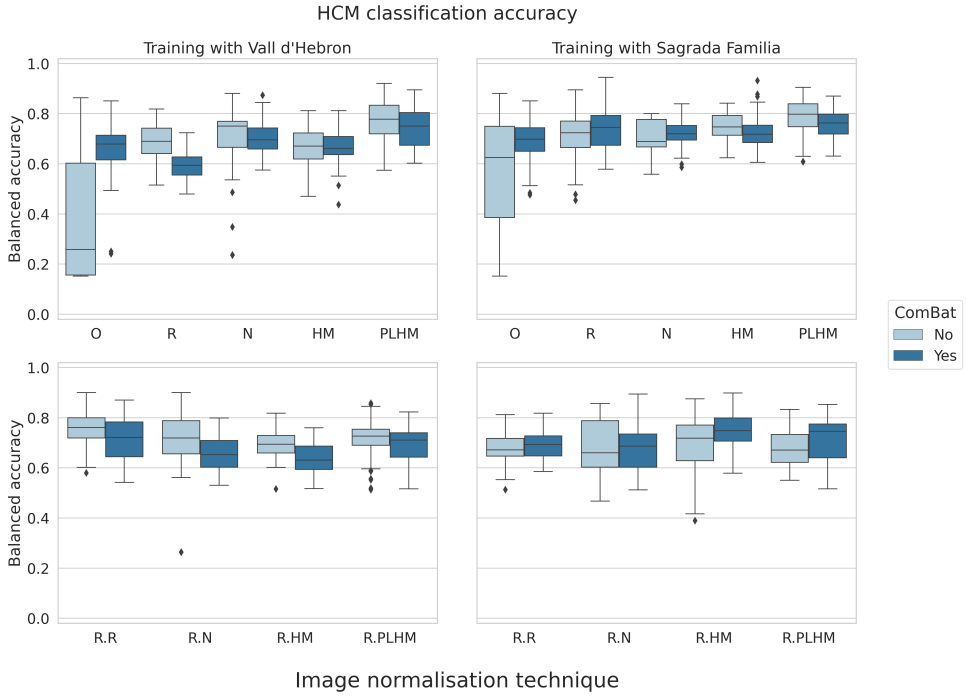


FIGURE 4.4: Balanced accuracy of random forest models on unseen centres for classification of HCM versus healthy patients. All models were trained with a combination of first and second-order texture features from all ROIs. The first column corresponds to models trained with features extracted from Vall d'Hebron studies, while models in the second column were trained with features from Sagrada Familia studies. The row above corresponds to image preprocessing techniques applied at the whole image level, while in the row below they are applied at the ROI level. HCM: Hypertrophic cardiomyopathy, O: original images (without normalisation), R: image intensity rescaling, N: image intensity normalisation, HM: histogram matching and PLHM: piecewise linear histogram matching. An "R." in front of a method means that it is applied at the ROI level.

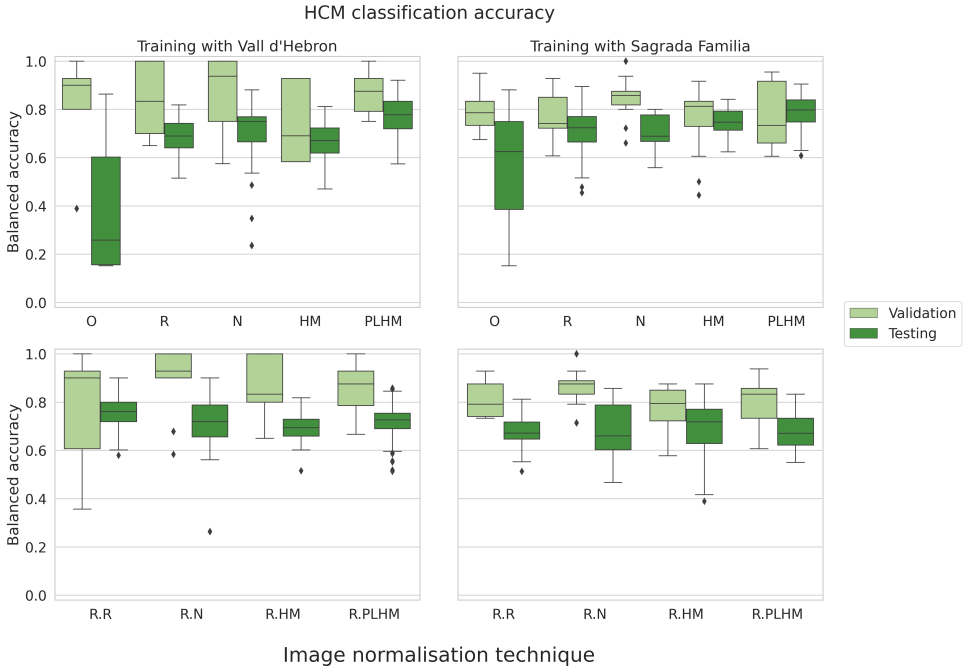


FIGURE 4.5: Balanced accuracy of random forest models on the validation set (same domain) versus the testing set (unseen centres) for classification of HCM versus healthy patients. Results are presented without ComBat harmonisation. All models were trained with a combination of first and second-order texture features from all ROIs. The first column corresponds to models trained with features extracted from Vall d'Hebron studies, while models in the second column were trained with features from Sagrada Familia studies. The row above corresponds to image preprocessing techniques applied at the whole image level, while in the row below they are applied at the ROI level. HCM: Hypertrophic cardiomyopathy, O: original images (without normalisation), R: image intensity rescaling, N: image intensity normalisation, HM: histogram matching and PLHM: piecewise linear histogram matching. An "R." in front of a method means that it is applied at the ROI level.

4.4 Discussion

Radiomic features are promising biomarkers for better disease characterisation. However, their variability across centres makes it difficult to establish reproducible biomarkers based on them [167]. In this study, a comprehensive analysis was carried out to assess feature variability across centres as well as model generalisation for a classification task after the application of several image normalisation techniques and a feature-based harmonisation technique (ComBat).

Based on the results presented, ROI-based PLHM is a good normalisation technique to preserve similar feature distributions across domains (see Fig. 4.1) and to reduce the amount of centre-related information encoded in radiomic features compared to original images (see Fig. 4.3). In brain MRI literature, however, the transformations that yielded less feature variability and more similar distributions were histogram matching and image intensity normalisation [17, 156], although Um et al. [156] did not consider PLHM in their work.

At the feature normalisation level, ComBat satisfactorily removed centre-related information to the point that models were not able to discriminate between features

depending on the institution of origin of each scan. As a drawback, the final feature distributions at different centres were less similar than before the application of ComBat according to the JSD.

With regards to generalisation ability, models trained with features from original images resulted in poor performance for differentiation of healthy subjects from HCM patients in unseen centres, highlighting the importance of normalisation techniques for multi-centre studies. Models trained with features extracted after PLHM normalisation obtained the highest accuracy. In this method, average landmarks are obtained for a reference population, while for histogram matching the reference was only one subject. This could explain why these methods showed differences in performance despite relying on the same principle since defining a template using only one subject could introduce unwanted bias in the analysis. Moreover, the selection of a particular population or subject as a reference template in these methods may affect the results, especially for histogram matching (see Fig. A.6 and A.7 in the supplementary material).

Importantly, successful centre-related information removal from radiomic features does not imply greater generalisation ability. In fact, ROI-based PLHM and ComBat harmonisation methods were not among the best generalisation techniques for the HCM classification task (Fig. 4.4, bottom row). When compared with the brain MRI literature, Orlhac et al. [117] did find an improvement in sensitivity for differentiating between low and high-risk patient groups when using ComBat harmonisation, although the authors did not compare different image normalisation techniques. Lastly, the model trained after PLHM, which showed the best generalisation ability, obtained medium performance on the validation set signalling a reduction in overfitting. The most important features of this model were predominantly features from the MYO, which made sense for the classification task at hand since HCM is most evident when looking at the myocardium.

This work presents several limitations. First, the dataset was not perfectly balanced across the five centres and the population was not controlled by age, sex, body size, or myocardial volume, which could result in dissimilarities across feature distributions. However, no significant differences were found under a Mann-Whitney U test in the volumes of the different ROIs between centres.

The choice of HCM classification as a metric for generalisation has some drawbacks since the heart suffers morphological changes and some texture features are known to be correlated with the shape [58]. This could contribute to overestimating the generalisation ability. The inclusion of other pathologies that greatly affect the myocardium, such as myocarditis or infarction, would potentially result in a less biased generalisation loss estimation.

Finally, according to the Imaging Biomarker Standardization Initiative, ISBI, second-order features from different texture matrices may be modelled better with different intensity discretisation levels (*e.g.* GLSZM are better characterised for low discretisation levels while it is the opposite for GLCM) [182]. In this work, the same discretisation level was used for all features.

4.5 Conclusions

In summary, this study showed that centre-related information removal does not imply good generalisation ability for classification. ComBat harmonisation was able to

remove centre-related information from radiomic features satisfactorily while showing limited generalisation ability. PLHM normalisation resulted in the best generalisable model for the classification of healthy subjects from HCM patients. The choice of reference template when performing histogram matching may affect the results. PLHM was robust against a change in the reference population. Finally, the radiomic features from GLSZM, GLDM and GLRLM families showed greater variability than first-order and GLCM features. Further studies with a larger sample size are needed to replicate the results presented and to assess the effect of different biological covariates.

Chapter 5

Conclusions

In this Thesis, we have contributed to three main aspects of the domain shift challenge in cardiovascular imaging: a) We have assembled and open-sourced the first multi-centre and multi-vendor dataset, including six clinical centres and four scanner manufacturers, and established a benchmark for generalizable cardiac image segmentation, b) We have compared DG and DA approaches for LGE image segmentation and c) We have studied the effects of different pre-processing and harmonisation steps on model generalizability for a classification task.

5.1 Summary of findings

The main findings of this Thesis are the importance of exhaustive data augmentation on real data to improve model training and generalizability to new unseen domains, the competitiveness of transfer learning compared to multi-centre models and models that use extensive data augmentation and finally, the need to harmonise images before feature extraction for improved model diagnosis in multi-centre settings. These findings are discussed separately and in more detail next.

Multi-centre cardiac image segmentation

In chapter 2, we presented the results for the M&Ms Challenge, a competition for benchmarking segmentation models for multi-centre cardiac image segmentation. It became the first open-source benchmark with multiple domains to test DA and DG proposals. Several findings were extracted from the participants' proposals that were in line with the existing literature:

- Deep learning models generalize better when trained with a heterogeneous dataset combining studies from different institutions and acquired with different scanners.
- Data augmentation with varied spatial and intensity-based transformations is a simple but appropriate solution to increase the diversity in the training images and obtain generalizable models.

Other findings established a direct comparison between DG and DA techniques:

- Deep DA techniques, such as the one used by Corral-Acero [30], are not able to outperform optimised pipelines with extensive data augmentation, such as nnUNet [67], but are promising tools.
- Other techniques that proposed explicit domain adaptation modules, such as the approach followed by Liu et al. [94], obtained worse results during the

M&Ms Challenge but were later refined [96] to outperform the nnUNet framework, representing another potential method to be considered in future studies.

Domain shift for LGE image segmentation

In chapter 3, the domain shift challenge was studied for LGE cardiac image segmentation. This modality poses an increased difficulty compared to non-contrast MRI due to the existence of more factors affecting the final scan outcome and we were able to derive the following conclusions:

- Transfer learning is a good DA approach to obtaining a good-performing model when few annotated samples are available from the target domain, outperforming models trained with extensive data augmentation and obtaining comparative performance to models trained from scratch.
- Data augmentation can increase the training set variability to avoid using histogram matching on new samples or training costly domain-translation generative models based on CycleGAN.
- Using auxiliary related datasets, such as non-contrast MRI, for pretraining is also effective for the transfer learning approach.

Multi-centre image harmonisation for classification

Finally, in chapter 4, we compared feature and image transform techniques for the harmonisation of imaging features. Then, we assessed the effects on the domain gap for a classification task. This study allowed us to derive the following findings:

- ROI-based PLHM is a good harmonisation technique to enforce feature distribution similarity across centres.
- ComBat harmonisation makes it impossible to determine the centre of origin from features, but the distribution similarity is worse than for other techniques.
- Models trained with features after PLHM harmonisation obtained the best generalization performance.
- Successful removal of centre-related information from imaging features does not guarantee a good generalization ability for diagnosis.

5.2 Future work

Several topics related to the content of this Thesis still need to be investigated further. Here we highlight some possible avenues to pursue to advance the understanding of multi-centre data analysis and obtain generalizable models.

Importance of multi-centre datasets

The M&Ms Challenge data is a clinical dataset with a heterogeneous cohort, so it is complicated to stratify the domain shift due to different factors, such as disease, age or body mass index. A well-curated dataset with a balanced representation of subgroups may be useful to disaggregate sources of performance degradation. Despite being one of the largest open datasets in cardiovascular research, its size is still

limited when compared to other organs. Larger datasets are needed to improve the significance of the results and obtain stronger evidence of the conclusions extracted. For instance, deep DA models Corral Acero et al. [30] and Liu et al. [94, 96] have shown promising results and a detailed analysis is needed to understand when are these techniques stronger and how to improve their performance.

Similarly, the harmonisation results derived from the work in this Thesis were extracted with a limited-size dataset. Replication studies are key as well as extensions to image-based classification models. For the last point, we need a dataset specifically designed to contain a balanced group of pathological cases across institutions.

Segmentation of the scar tissue in LGE images

The study conducted on LGE imaging did not consider scar tissue annotations, which are more challenging to delineate. The variability in the brightness and spatial distribution of these tissues make it harder to generalize to new centres and even new cohorts with different lesions. It is necessary to evaluate the performance degradation for scar tissue annotation for images with varied diseases and different centres. A recent approach has been proposed to synthetically augment the training dataset considering the factors of variation mentioned above [162].

The transfer learning approach turned out to be the best-performing solution in the existence of a few annotated samples from the target domain, even when the model was pre-trained with non-contrast MRI scans. This highlights the usefulness of foundational (i.e. pre-trained) models both in terms of performance and training cost. Recent works have confirmed the boost in performance when the pre-trained model is trained with a dataset that is closer to the final target domain data [108]. Therefore, future studies are needed to investigate the optimal way to fine-tune a pre-trained model to improve generalization, obtain fair models and even, incorporate new knowledge – such as new diseases or pathologies – into an already fine-tuned model.

Other general aspects

Towards trustworthy AI

The successful implementation of AI in clinical practice depends on the implementation of guidelines and compliance with basic principles. FUTURE-AI [85] is an initiative that assesses these aspects via six building blocks referred to as Fairness, Universality, Traceability, Usability, Robustness and Explainability. Next, we discuss the building blocks that are more relevant within the context of this Thesis and we introduce other key aspects that we think are important for the objective of trustworthy AI.

Fairness. AI techniques are sensible to specific patterns found in the training dataset, such as specific details from one scanner manufacturer, skin tone, sex, ethnicity or other factors within the training cohort. Developing AI solutions that are fair and perform equitably across subgroups is crucial and multi-centre data collection is key to increasing the diversity of these datasets.

Traceability. Model deployment in clinical settings requires the introduction of a monitoring system to log any potential change in performance due to concept or data drifts, and updating the model if necessary to account for them. Additionally, a model should be accompanied by metadata that defines the model scope, different

settings and statistics about its performance, enabling AI accountability and risk awareness.

Robustness. AI models deal with different sources of variations in medical images. In this Thesis, we have worked with variations related to the scanner manufacturer and the clinical centre where the image was acquired. However, other variations are due to the disparities between annotators or data cohorts. All these variations need to be analysed and accounted for, when possible, by using generalizable methodologies, such as the ones discussed in this Thesis, to reduce the domain gap error. For instance, we have seen that exhaustive data augmentation combined with model ensembling is a good recipe for improved segmentation performance on unseen vendors.

Interpretability. AI tools are commonly referred to as "black boxes" since the prediction is obtained after multiple transformations of the input data using the model weights and parameters in a process that is usually not interpretable. However, decisions in clinical practice cannot be simply motivated by the model output and need to be reasoned, especially for identifying potential model failures, but also for clinicians to understand the factors that were involved in the final prediction. Several tools exist to highlight the most relevant regions in an image for a given prediction (e.g. Class Activation Maps), but these techniques do not include any reasoning and often the regions are not meaningful for the expert.

Large-scale validation. To reduce uncertainty on model performance and robustly assess the model reliability against different factors of variation in datasets, it is very important to evaluate AI methods on large-scale datasets with great variability in terms of sex, ethnicity, pathologies, age, scanners, country of origin, among other factors. Such resources are difficult to obtain, especially in medical imaging, and in this Thesis, we have contributed the largest multi-centre open cardiac dataset to date. Furthermore, our current efforts are also targeting underrepresented African countries where old devices are often used resulting in images with lower quality as compared to standard European datasets.

Multi-modal models with context

Very often the AI models proposed in the research literature focus on a narrow task and achieve very good performance. However, contextual information is often needed in clinical practice to make a diagnosis (*e.g.*, the patient's familiar history, the patient's clinical history, genetic information or multiple image modalities). Recent studies have demonstrated the usefulness of such approaches for improved diagnosis [59, 166].

Synthetic data

There are frameworks in computer vision to generate synthetic samples with annotations from simulation environments, such as videogames [135, 138]. In medical imaging, some approaches have used atlases or statistical models to generate samples or derive an annotation [42]. Recent progress in generative modelling can make training models with realistic synthetic images stand out over extensive data augmentation on a dataset with limited annotations [39].

Regulation and approval

AI models are different from traditional software tools in that their outcomes are conditioned on the training data or the training algorithm and may evolve when

more data is fed into the model over time. We need a clear understanding of the effects of updating the models and a practical way of monitoring performance evolution. Currently, the Food and Drug Administration in the United States is considering a pre-certification program that focuses on the «culture of quality and organizational excellence» of the developer to speed up the certification of an AI-based product [60]. The European Union is also creating a set of obligations for «high-risk AI systems» before they can be introduced to the market, such as guarantees of a high level of robustness and accuracy or human oversight to reduce the risk [133].

Chapter 6

Contributions

The author of this Thesis also contributed to the following works, presented in chronological order, during the PhD programme:

1. **Campello, V.M.**, Martín-Isla, C., Izquierdo, C., Petersen, S.E., Ballester, M.A.G., Lekadir, K. (2020). "Combining Multi-Sequence and Synthetic Images for Improved Segmentation of Late Gadolinium Enhancement Cardiac MRI." In: Pop, M., et al. *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges. STACOM 2019. Lecture Notes in Computer Science*, vol 12009. Springer, Cham. Doi: 10.1007/978-3-030-39074-7_31.

2. Martín-Isla, C., Asadi-Aghbolaghi, M., Gkontra, P., **Campello, V.M.**, Escalera, S., Lekadir, K. (2020). "Stacked BCDU-Net with Semantic CMR Synthesis: Application to Myocardial Pathology Segmentation Challenge." In: Zhuang, X., Li, L. (eds) *Myocardial Pathology Segmentation Combining Multi-Sequence Cardiac Magnetic Resonance Images. MyoPS 2020. Lecture Notes in Computer Science*, vol 12554. Springer, Cham. Doi: 10.1007/978-3-030-65651-5_1.

3. Martín-Isla, C., **Campello, V. M.**, Izquierdo, C., Raisi-Estabragh, Z., Baeßler, B., Petersen, S. E., Lekadir, K. (2020). "Image-based cardiac diagnosis with machine learning: a review." *Frontiers in cardiovascular medicine*, 7, 1. Doi: 10.3389/fcvm.2020.00001.
JCR IF: 3.6 Q2.

4. Raisi-Estabragh, Z., Izquierdo, C., **Campello, V. M.**, Martín-Isla, C., Jaggi, A., Harvey, N. C., Lekadir, K., Petersen, S. E. (2020). "Cardiac magnetic resonance radiomics: basic principles and clinical perspectives." *European Heart Journal - Cardiovascular Imaging*, Volume 21, Issue 4, April 2020, Pages 349–356. Doi: 10.1093/ehjci/jeaa028.
JCR IF: 6.2 Q1.

5. Liu, Z., Manh, V., Yang, X., Huang, X., Lekadir, K., **Campello V. M.**, Ravikumar, N., Frangi, A. F., Ni, D. (2021). Style Curriculum Learning for Robust Medical Image Segmentation. In: de Bruijne, M., et al. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science*, vol 12901. Springer, Cham. Doi: 10.1007/978-3-030-87193-2_43.

6. Izquierdo, C., Casas, G., Martín-Isla, C., **Campello, V. M.**, Guala, A., Gkontra, P., Rodríguez-Palomares, J. F., Lekadir, K. (2021). "Radiomics-based classification of left ventricular non-compaction, hypertrophic cardiomyopathy, and dilated cardiomyopathy in cardiovascular magnetic resonance." *Frontiers in*

Cardiovascular Medicine, 1502. Doi: 10.3389/fcvm.2021.764312.
JCR IF: 3.6 Q2.

7. Joshi, S., Osuala, R., Martín-Isla, C., **Campello, V. M.**, Sendra-Balcells, C., Lekadir, K., Escalera, S. (2022). "nn-UNet Training on CycleGAN-Translated Images for Cross-modal Domain Adaptation in Biomedical Imaging." In: Crimi, A., Bakas, S. (eds) *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2021. Lecture Notes in Computer Science*, vol 12963. Springer, Cham. Doi: 10.1007/978-3-031-09002-8_47

8. Pina, V., **Campello, V. M.**, Lekadir, K., Seguí, S., García-Santos, J. M., Fuentes, L. J. (2022). "Mathematical Abilities in School-Aged Children: A Structural Magnetic Resonance Imaging Analysis With Radiomics." *Frontiers in neuroscience*, 16, 819069. Doi: 10.3389/fnins.2022.819069.
JCR IF: 4.3 Q2.

9. Kushibar, K., **Campello, V. M.**, Garrucho, L., Linardos, A., Radeva, P., Lekadir, K. (2022). "Layer Ensembles: A Single-Pass Uncertainty Estimation in Deep Learning for Segmentation." In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. MICCAI 2022. Lecture Notes in Computer Science*, vol 13438. Springer, Cham. Doi: 10.1007/978-3-031-16452-1_49.

10. **Campello, V. M.**, Xia, T., Liu, X., Sanchez, P., Martín-Isla, C., Petersen, S. E., Seguí, S., Tsaftaris, S. A., Lekadir, K. (2022). "Cardiac aging synthesis from cross-sectional data with conditional generative adversarial networks." *Frontiers in Cardiovascular Medicine*, 9, 983091. Doi: 10.3389/fcvm.2022.983091.
JCR IF: 3.6 Q2.

11. Zhuang, X., Xu, J., Luo, X., Chen, C., Ouyang, C., Rueckert, D., **Campello, V. M.**, Lekadir, K., Vesal, S., Ravikumar, N., Liu, Y., Luo, G., Chen, J., Li, H., Ly, B., Sermesant, M., Roth, H., Zhu, W., Wang, J., Ding, X., Wang, X., Yang, S., Li, L. (2022). "Cardiac segmentation on late gadolinium enhancement MRI: a benchmark study from multi-sequence cardiac MR segmentation challenge." *Medical Image Analysis*, 81, 102528. Doi: 10.1016/j.media.2022.102528.
JCR IF: 10.9 Q1.

12. Szabo, L., Ruiz-Pujadas, E., McCracken, C., Izquierdo, C., **Campello, V. M.**, Atehortua, A., Petersen, S. E., Lekadir, K., Raisi-Estabragh, Z. (2022). "Cardiac magnetic resonance radiomics for prediction of incident heart failure: a feasibility study in the UK Biobank Imaging cohort." *European Heart Journal*, 43(Supplement_2), ehac544-935. Doi: 10.1093/eurheartj/ehac544.935.
JCR IF: 39.3 Q1.

13. Ruiz-Pujadas, E., Raisi-Estabragh, Z., Szabo, L., Izquierdo, C., **Campello, V. M.**, Martín-Isla, C., Vago, H., Merkely, B., Harvey, N. C., Petersen, S. E., Lekadir, K. (2022). "Atrial fibrillation prediction by combining ECG markers and CMR radiomics." *Scientific Reports*, 12(1), 18876. Doi: 10.1038/s41598-022-21663-w.
JCR IF: 4.6 Q2.

14. Sendra-Balcells, C., **Campello, V.M.**, Torrents-Barrena, J., Ahmed, Y. A., Elattar, M., Ohene-Botwe, B., Nyangulu, P., Stones, W., Ammar, M., Benamer, L. N., Kisembo, H. N., Sereke, S.G., Wanyonyi, S. Z., Temmerman, M., Gratacós, E., Bonet, E., Eixarch, E., Mikolaj, K., Tolsgaard, M. G., Lekadir, K. (2023)

"Generalisability of fetal ultrasound deep learning models to low-resource imaging settings in five African countries." *Scientific Reports*, 13, 2728. Doi: 10.1038/s41598-023-29490-3.
JCR IF: 4.6 Q2.

15. Salih, A. M., Ruiz-Pujadas, E., **Campello, V. M.**, McCracken, C., Harvey, N. C., Neubauer, S., Lekadir, K., Nichols, T. E., Petersen, S. E., Raisi-Estabragh, Z. (2023). "Image-Based Biological Heart Age Estimation Reveals Differential Aging Patterns Across Cardiac Chambers." *Journal of Magnetic Resonance Imaging*. Doi: 10.1002/jmri.28675.
JCR IF: 4.4 Q1.

16. Martín-Isla, C., **Campello, V. M.**, Izquierdo, C., Kushibar, K., Sendra-Balcells, C., Gkontra, P., Sojoudi, A., Fulton, M. J., Arega, T. W., Punithakumar, K., Li, L., Sun, X., Al Khalil, Y., Liu, D., Jabbar, S., Queirós, S., Galati, F., Mazher, M., Gao, Z., Beetz, M., Tautz, L., Galazis, C., Varela, M., Hullebrand, M., Grau, V., Zhuang, X., Puig, D., Zuluaga, M. A., Mohy-ud-Din, H., Metaxas, D., Breeuwer, M., van der Geest, R. J., Noga, M., Bricq, S., Rentschler, M. E., Guala, A., Petersen, S. E., Escalera, S., Rodríguez-Palomares, J. F., Lekadir, K. (2023). "Deep Learning Segmentation of the Right Ventricle in Cardiac MRI: The M&Ms Challenge." *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 7, pp. 3302-3313, July 2023. Doi: 10.1109/JBHI.2023.3267857.
JCR IF: 7.7 Q1.

17. Ruiz-Pujadas, E., Raisi-Estabragh, Z., Szabo, L., McCracken, C., Morcillo, C. I., **Campello, V. M.**, Martín-Isla, C., Atehortua, A. M., Vago, H., Merkely, B., Maurovich-Horvat, P., Harvey, N. C., Neubauer, S., Petersen, S. E., Lekadir, K. (2023). "Prediction of incident cardiovascular events using machine learning and CMR radiomics." *European radiology*, 33(5), 3488-3500. Doi: 10.1007/s00330-022-09323-z.
JCR IF: 5.9 Q1.

Appendix A

Additional results for domain generalization for cardiac diagnosis

A.1 Supplementary material

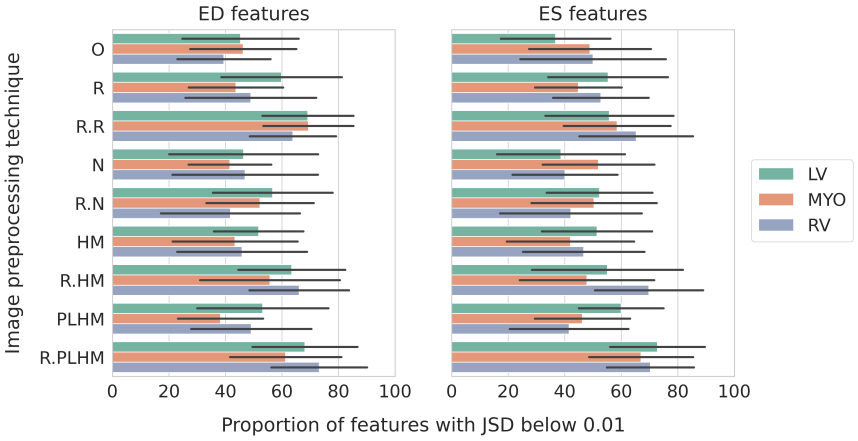


FIGURE A.1: Percentage of texture features below the 0.01 JSD threshold for each ROI for healthy subjects. Results are averaged over feature types and centre pairs and separated in ED and ES frames. Only features with square cross-correlation below 0.9 were considered. The black lines represent the standard deviation. O: original images (without normalisation), R: image intensity rescaling, N: image intensity normalisation, HM: histogram matching and PLHM: piecewise linear histogram matching. An "R." in front of a method means that it is applied at the ROI level.

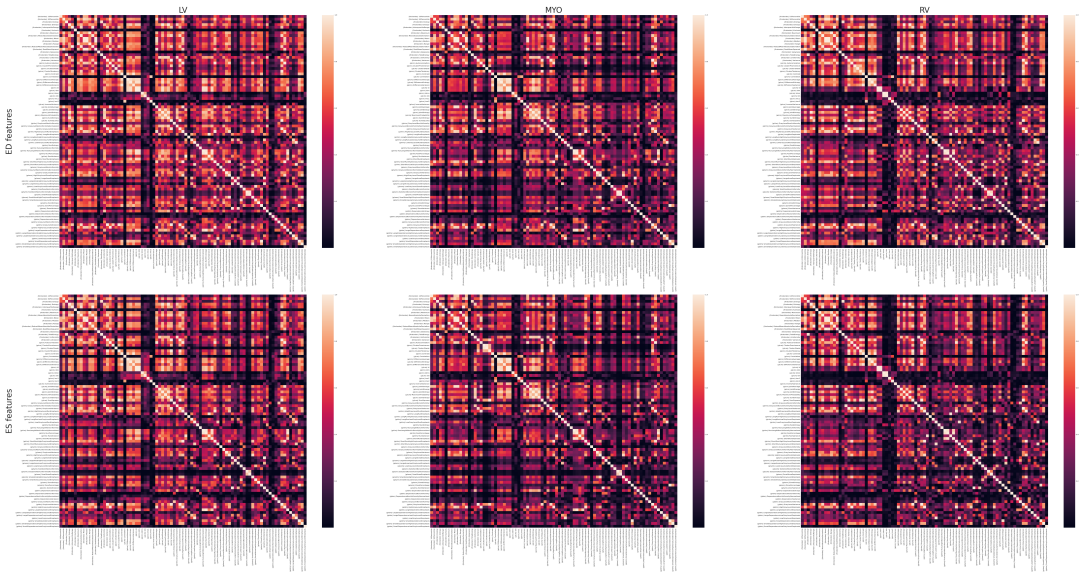


FIGURE A.2: Pairwise square correlation for features extracted from the three different ROIs without the application of any normalisation technique. The correlation between features was very similar for the different preprocessing techniques, showing a negligible standard deviation. Zoom in to see in more detail.

TABLE A.1: List of radiomic features extracted with PyRadiomics. We refer the reader to the library documentation (https://pyradiomics.readthedocs.io/) for the specific definition and interpretation of each feature.

Family	Index	Feature
Shape	1	Elongation
	2	Flatness
	3	LeastAxisLength
	4	MajorAxisLength
	5	Maximum2DDiameterColumn
	6	Maximum2DDiameterRow
	7	Maximum2DDiameterSlice
	8	Maximum3DDiameter
	9	MeshVolume
	10	MinorAxisLength
	11	Sphericity
	12	SurfaceArea
	13	SurfaceVolumeRatio
	14	VoxelVolume
First order	15	10Percentile
	16	90Percentile
	17	Energy
	18	Entropy
	19	InterquartileRange
	20	Kurtosis
	21	Maximum
	22	MeanAbsoluteDeviation
	23	Mean
	24	Median
	25	Minimum
	26	Range
	27	RobustMeanAbsoluteDeviation
	28	RootMeanSquared
	29	Skewness
	30	TotalEnergy
	31	Uniformity
	32	Variance
GLCM	33	Autocorrelation
	34	JointAverage
	35	ClusterProminence
	36	ClusterShade
	37	ClusterTendency
	38	Contrast
	39	Correlation
	40	DifferenceAverage
	41	DifferenceEntropy
	42	DifferenceVariance
	43	JointEnergy
	44	JointEntropy
	45	Imc1
	46	Imc2
	47	Idm
	48	Idmn
	49	Id
	50	Idn
GLCM	51	InverseVariance
	52	MaximumProbability
	53	SumEntropy
	54	SumSquares
GLRLM	55	GrayLevelNonUniformity
	56	GrayLevelNonUniformityNormalized
	57	GrayLevelVariance
	58	HighGrayLevelRunEmphasis
	59	LongRunEmphasis
	60	LongRunHighGrayLevelEmphasis
	61	LongRunLowGrayLevelEmphasis
	62	LowGrayLevelRunEmphasis
	63	RunEntropy
	64	RunLengthNonUniformity
	65	RunLengthNonUniformityNormalized
	66	RunPercentage
	67	RunVariance
	68	ShortRunEmphasis
	69	ShortRunHighGrayLevelEmphasis
	70	ShortRunLowGrayLevelEmphasis
GLSZM	71	GrayLevelNonUniformity
	72	GrayLevelNonUniformityNormalized
	73	GrayLevelVariance
	74	HighGrayLevelZoneEmphasis
	75	LargeAreaEmphasis
	76	LargeAreaHighGrayLevelEmphasis
	77	LargeAreaLowGrayLevelEmphasis
	78	LowGrayLevelZoneEmphasis
	79	SizeZoneNonUniformity
	80	SizeZoneNonUniformityNormalized
	81	SmallAreaEmphasis
	82	SmallAreaHighGrayLevelEmphasis
	83	SmallAreaLowGrayLevelEmphasis
	84	ZoneEntropy
	85	ZonePercentage
	86	ZoneVariance
GLDM	87	DependenceEntropy
	88	DependenceNonUniformity
	89	DependenceNonUniformityNormalized
	90	DependenceVariance
	91	GrayLevelNonUniformity
	92	GrayLevelVariance
	93	HighGrayLevelEmphasis
	94	LargeDependenceEmphasis
	95	LargeDependenceHighGrayLevelEmphasis
	96	LargeDependenceLowGrayLevelEmphasis
	97	LowGrayLevelEmphasis
	98	SmallDependenceEmphasis
	99	SmallDependenceHighGrayLevelEmphasis
	100	SmallDependenceLowGrayLevelEmphasis

GLCM: Gray Level Co-occurrence Matrix, GLRLM: Gray Level Run Length Matrix, GLSZM: Gray Level Size Zone Matrix and GLDM: Gray Level Dependence Matrix.

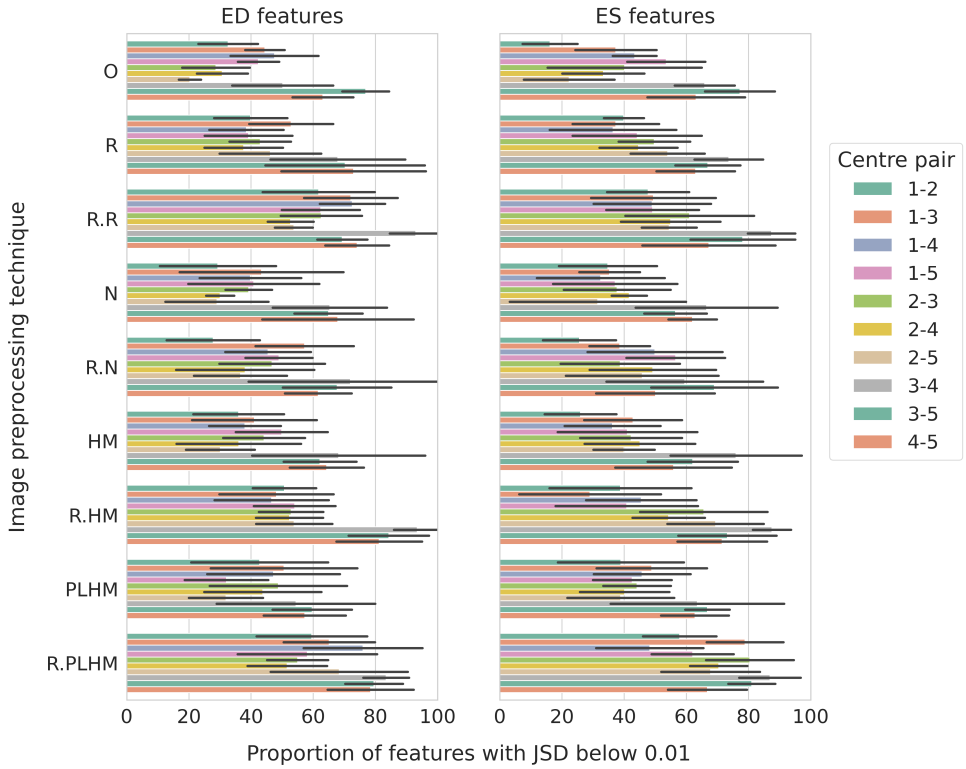


FIGURE A.3: Percentage of texture features below the 0.01 JSD threshold for each centre pair, indexed in alphabetical order as in Table 4.1, for healthy subjects. Results are averaged over feature types and ROIs and separated in ED and ES frames. Only features with square cross-correlation below 0.9 were considered. The black lines represent the standard deviation. O: original images (without normalisation), R: image intensity rescaling, N: image intensity normalisation, HM: histogram matching and PLHM: piecewise linear histogram matching. An "R." in front of a method means that it is applied at the ROI level.

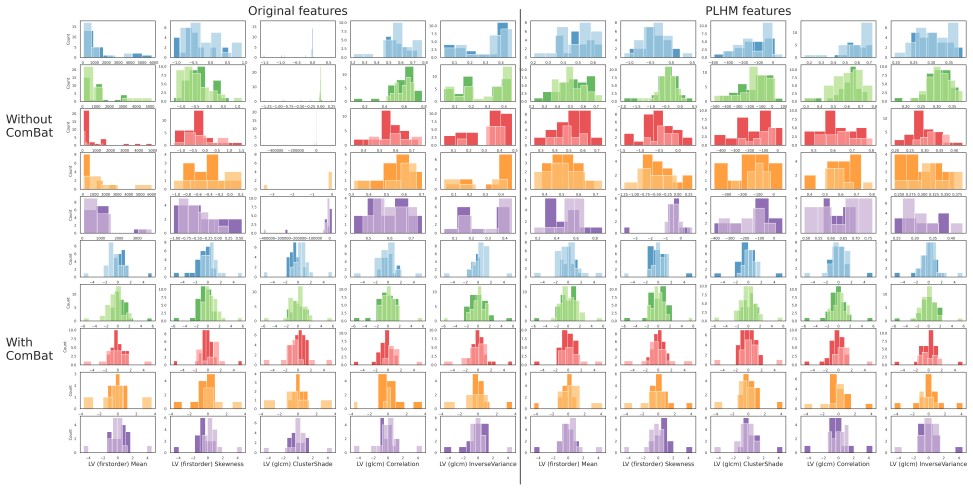


FIGURE A.4: Comparison of histograms of five different radiomic features from the LV across centres and for different normalisation methods. Histograms and separated for healthy (brighter colour) and HCM subjects (lighter colour). Centres are presented in different colours and different rows following the ordering in Table 4.2. The first five rows correspond to methods without ComBat harmonisation, while the last five rows represent the same features with ComBat harmonisation. The first five columns are distributions of features extracted from original images, while the last five are features extracted after PLHM.

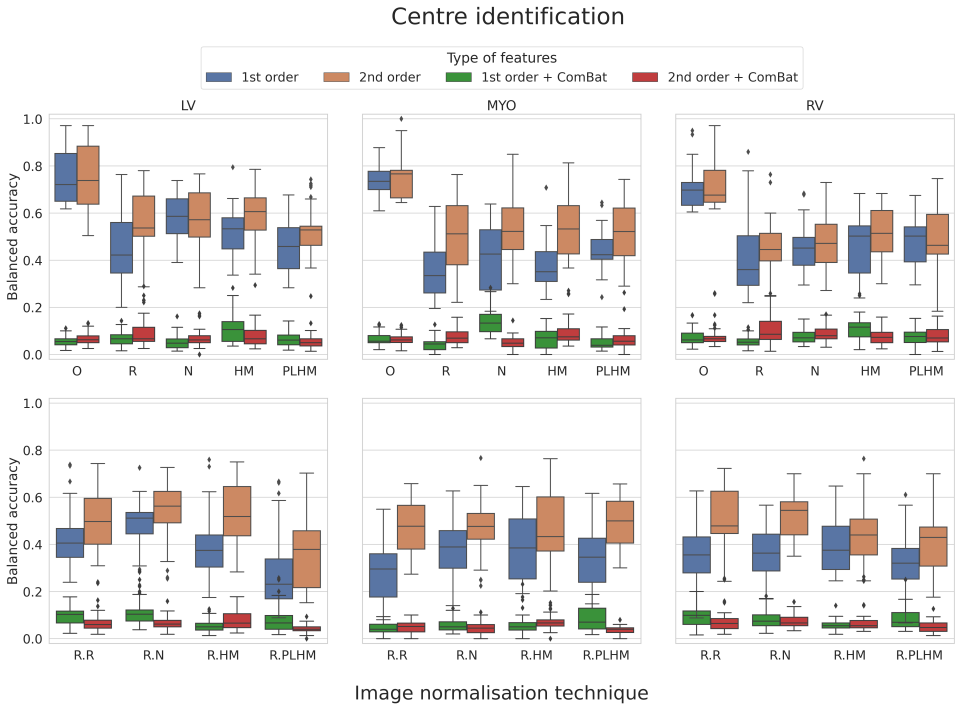


FIGURE A.5: Balanced accuracy of random forest models when predicting the centre of origin of HCM subjects for first and second-order texture features before and after the application of ComBat harmonisation. The row above corresponds to image preprocessing techniques applied at the whole image level, while in the row below they are applied at the ROI level. O: original images (without normalisation), R: image intensity rescaling, N: image intensity normalisation, HM: histogram matching and PLHM: piecewise linear histogram matching. An "R." in front of a method means that it is applied at the ROI level.

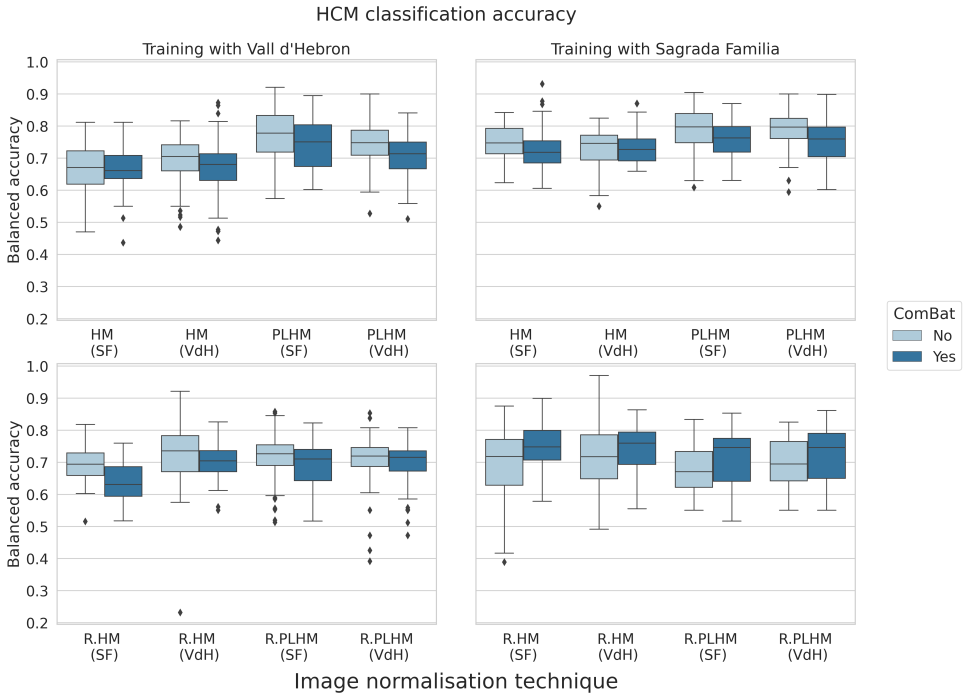


FIGURE A.6: Comparison of balanced accuracy for models trained with two different reference templates from Vall d'Hebron (VdH) and Sagrada Familia (SF) on the HCM classification task. All models were trained with a combination of first and second-order texture features from all ROIs. The first column corresponds to models trained with features extracted from Vall d'Hebron studies, while models in the second column were trained with features from Sagrada Familia studies. The row above corresponds to image preprocessing techniques applied at the whole image level, while in the row below they are applied at the ROI level. HCM: Hypertrophic cardiomyopathy, VdH: Vall d'Hebron, SF: Sagrada Familia, HM: histogram matching and PLHM: piecewise linear histogram matching. An "R." in front of a method means that it is applied at the ROI level.

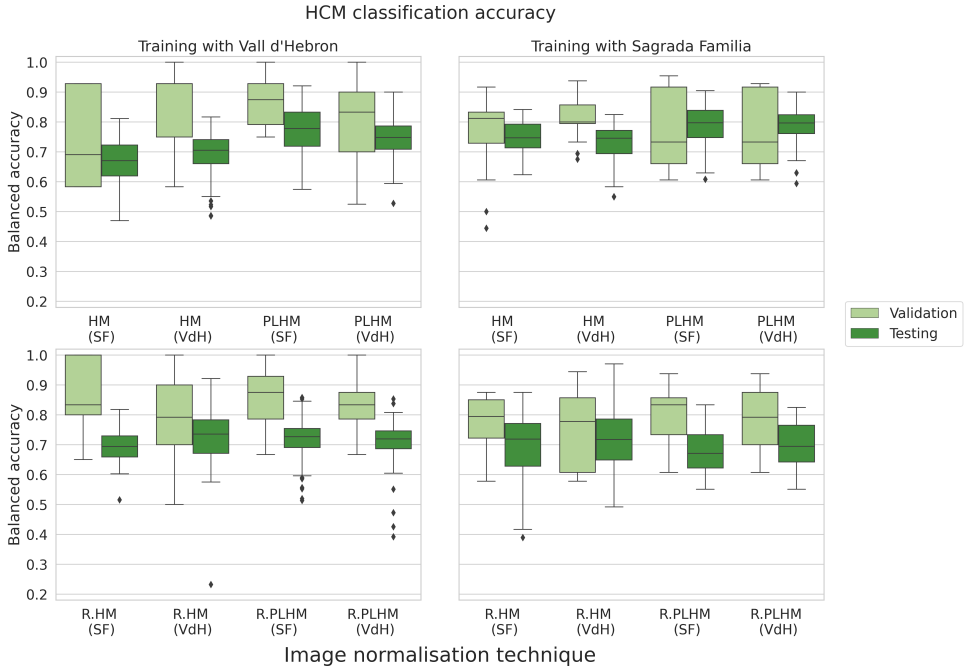


FIGURE A.7: Comparison of validation (same domain) and testing (unseen centres) balanced accuracy for models trained with two different reference templates from Vall d'Hebron (VdH) and Sagrada Familia (SF) on the HCM classification task. Results are presented without ComBat harmonisation. All models were trained with a combination of first and second order texture features from all ROIs. The first column corresponds to models trained with features extracted from Vall d'Hebron studies, while models in the second column were trained with features from Sagrada Familia studies. The row above corresponds to image preprocessing techniques applied at the whole image level, while in the row below they are applied at the ROI level. HCM: Hypertrophic cardiomyopathy, O: original images (without normalisation), R: image intensity recaling, N: image intensity normalisation, HM: histogram matching and PLHM: piecewise linear histogram matching. An "R." in front of a method means that it is applied at ROI level.

Bibliography

[1] Hugo J. W. L. Aerts et al. "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach". In: *Nature Communications* 5.1 (2014). DOI: 10.1038/ncomms5006.

[2] Ehab A. AlBadawy, Ashirbani Saha, and Maciej A. Mazurowski. "Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing". en. In: *Medical Physics* 45.3 (2018), pp. 1150–1158. ISSN: 2473-4209. DOI: 10.1002/mp.12752.

[3] Xènia Albà et al. "Automatic initialization and quality control of large-scale cardiac MRI segmentations". In: *Medical Image Analysis* 43 (Jan. 2018), pp. 129–141. ISSN: 1361-8415. DOI: 10.1016/j.media.2017.10.001.

[4] Léo Andéol et al. "Learning Domain Invariant Representations by Joint Wasserstein Distance Minimization". In: *arXiv preprint arXiv:2106.04923* (June 2021). DOI: 10.48550/arXiv.2106.04923.

[5] Michela Antonelli et al. "The Medical Segmentation Decathlon". en. In: *Nature Communications* 13.1 (July 2022). Number: 1 Publisher: Nature Publishing Group, p. 4128. ISSN: 2041-1723. DOI: 10.1038/s41467-022-30695-9.

[6] Brian Avants, Nicholas J. Tustison, and Gang Song. "Advanced Normalization Tools: V1.0". In: *The Insight Journal* (July 2009). ISSN: 2327-770X. DOI: 10.54294/uvnhin.

[7] Wenjia Bai et al. "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks". In: *Journal of Cardiovascular Magnetic Resonance* 20.1 (Sept. 2018), p. 65. ISSN: 1532-429X. DOI: 10.1186/s12968-018-0471-x.

[8] Wenjia Bai et al. "Multi-atlas segmentation with augmented features for cardiac MR images". In: *Medical Image Analysis* 19.1 (Jan. 2015), pp. 98–109. ISSN: 1361-8415. DOI: 10.1016/j.media.2014.09.005.

[9] Wenjia Bai et al. "Self-Supervised Learning for Cardiac MR Image Segmentation by Anatomical Position Prediction". en. In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019*. Ed. by Dinggang Shen et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 541–549. ISBN: 978-3-030-32245-8. DOI: 10.1007/978-3-030-32245-8_60.

[10] Olivier Bernard et al. "Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?" In: *IEEE Transactions on Medical Imaging* 37.11 (Nov. 2018), pp. 2514–2525. ISSN: 1558-254X. DOI: 10.1109/TMI.2018.2837502.

[11] Konstantinos Bousmalis et al. "Domain Separation Networks". In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc., 2016.

[12] Konstantinos Bousmalis et al. "Unsupervised Pixel-Level Domain Adaptation With Generative Adversarial Networks". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 3722–3731.

[13] L. Breiman. "Random Forests". In: *Machine Learning* 45 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[14] Víctor M. Campello et al. "Combining Multi-Sequence and Synthetic Images for Improved Segmentation of Late Gadolinium Enhancement Cardiac MRI". en. In: *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges*. Ed. by Mihaela Pop et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 290–299. ISBN: 978-3-030-39074-7. DOI: 10.1007/978-3-030-39074-7_31.

[15] Víctor M. Campello et al. "Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge". In: *IEEE Transactions on Medical Imaging* (2021). DOI: 10.1109/TMI.2021.3090082.

[16] DH Carr et al. "Gadolinium-DTPA as a contrast agent in MRI: initial clinical experience in 20 patients". In: *American Journal of Roentgenology* 143.2 (Aug. 1984), pp. 215–224. ISSN: 0361-803X. DOI: 10.2214/ajr.143.2.215.

[17] Alexandre Carré et al. "Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics". In: *Scientific Reports* 10.1 (2020). DOI: 10.1038/s41598-020-69298-z.

[18] Adam Carscadden, Michelle Noga, and Kumaradevan Punithakumar. "A Deep Convolutional Neural Network Approach for the Segmentation of Cardiac Structures from MRI Sequences". en. In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*. Ed. by Esther Puyol Anton et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 250–258. ISBN: 978-3-030-68107-4. DOI: 10.1007/978-3-030-68107-4_25.

[19] Irem Cetin et al. "A Radiomics Approach to Computer-Aided Diagnosis with Cardiac Cine-MRI". en. In: *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*. Ed. by Mihaela Pop et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 82–90. ISBN: 978-3-319-75541-0. DOI: 10.1007/978-3-319-75541-0_9.

[20] Irem Cetin et al. "Radiomics Signatures of Cardiovascular Risk Factors in Cardiac MRI: Results From the UK Biobank". In: *Frontiers in Cardiovascular Medicine* 7 (2020). ISSN: 2297-055X. DOI: 10.3389/fcvm.2020.591368.

[21] Agisilaos Chartsias et al. "Disentangled representation learning in cardiac image analysis". In: *Medical Image Analysis* 58 (Dec. 2019), p. 101535. DOI: 10.1016/j.media.2019.101535.

[22] Avishek Chatterjee et al. "Creating Robust Predictive Radiomic Models for Data From Independent Institutions Using Normalization". In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 3.2 (2019), pp. 210–215. DOI: 10.1109/trpms.2019.2893860.

[23] Anika Cheerla and Olivier Gevaert. "Deep learning with multimodal representation for pancancer prognosis prediction". In: *Bioinformatics* 35.14 (July 2019), pp. i446–i454. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz342.

[24] Chen Chen et al. "Deep Learning for Cardiac Image Segmentation: A Review". In: *Frontiers in Cardiovascular Medicine* 7 (2020). ISSN: 2297-055X. DOI: 10.3389/fcvm.2020.00025.

[25] Chen Chen et al. "Improving the Generalizability of Convolutional Neural Network-Based Segmentation on CMR Images". In: *Frontiers in Cardiovascular Medicine* 7 (2020). ISSN: 2297-055X. DOI: 10.3389/fcvm.2020.00105.

[26] Chen Chen et al. "Unsupervised Multi-modal Style Transfer for Cardiac MR Segmentation". en. In: *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges*. Ed. by Mihaela Pop et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 209–219. ISBN: 978-3-030-39074-7. DOI: 10.1007/978-3-030-39074-7_22.

[27] Cheng Chen et al. "Semantic-Aware Generative Adversarial Nets for Unsupervised Domain Adaptation in Chest X-Ray Segmentation". en. In: *Machine Learning in Medical Imaging*. Ed. by Yinghuan Shi, Heung-Il Suk, and Mingxia Liu. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 143–151. ISBN: 978-3-030-00919-9. DOI: 10.1007/978-3-030-00919-9_17.

[28] Veronika Cheplygina et al. "Transfer Learning for Multicenter Classification of Chronic Obstructive Pulmonary Disease". In: *IEEE Journal of Biomedical and Health Informatics* 22.5 (Sept. 2018), pp. 1486–1496. ISSN: 2168-2208. DOI: 10.1109/JBHI.2017.2769800.

[29] Brian Chu et al. "Best Practices for Fine-Tuning Visual Classifiers to New Domains". en. In: *Computer Vision - ECCV 2016 Workshops*. Ed. by Gang Hua and Hervé Jégou. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 435–442. ISBN: 978-3-319-49409-8. DOI: 10.1007/978-3-319-49409-8_34.

[30] Jorge Corral Acero et al. "A 2-Step Deep Learning Method with Domain Adaptation for Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Magnetic Resonance Segmentation". en. In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*. Ed. by Esther Puyol Anton et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 196–207. ISBN: 978-3-030-68107-4. DOI: 10.1007/978-3-030-68107-4_20.

[31] Gabriela Csurka, ed. *Domain Adaptation in Computer Vision Applications*. en. Advances in Computer Vision and Pattern Recognition. Cham: Springer International Publishing, 2017. ISBN: 978-3-319-58346-4 978-3-319-58347-1. DOI: 10.1007/978-3-319-58347-1.

[32] R Da-Ano, D Visvikis, and M Hatt. "Harmonization strategies for multicenter radiomics investigations". In: *Physics in Medicine & Biology* 65.24 (2020), 24TR02. DOI: 10.1088/1361-6560/aba798.

[33] Wenyuan Dai et al. "Boosting for transfer learning". In: *Proceedings of the 24th international conference on Machine learning*. ICML '07. New York, NY, USA: Association for Computing Machinery, June 2007, pp. 193–200. ISBN: 978-1-59593-793-3. DOI: 10.1145/1273496.1273521.

[34] Adelina Doltra et al. "Emerging concepts for myocardial late gadolinium enhancement MRI". In: *Current cardiology reviews* 9.3 (2013), pp. 185–190.

[35] Jeff Donahue et al. "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition". en. In: *Proceedings of the 31st International Conference on Machine Learning*. PMLR, Jan. 2014, pp. 647–655.

[36] Qi Dou et al. "Domain Generalization via Model-Agnostic Learning of Semantic Features". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.

[37] Qi Dou et al. "Unsupervised Cross-Modality Domain Adaptation of ConvNets for Biomedical Image Segmentations with Adversarial Loss". In: (2018), pp. 691–697. DOI: 10.24963/ijcai.2018/96.

[38] FDA. *Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices*. May 2022. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices (visited on 11/22/2022).

[39] Virginia Fernandez et al. "Can Segmentation Models Be Trained with Fully Synthetically Generated Data?" en. In: *Simulation and Synthesis in Medical Imaging*. Ed. by Can Zhao et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2022, pp. 79–90. ISBN: 978-3-031-16980-9. DOI: 10.1007/978-3-031-16980-9_8.

[40] Jean-Christophe Ferré, Mark S. Shiroishi, and Meng Law. "Advanced Techniques Using Contrast Media in Neuroimaging". English. In: *Magnetic Resonance Imaging Clinics* 20.4 (Nov. 2012), pp. 699–713. ISSN: 1064-9689, 1557-9786. DOI: 10.1016/j.mric.2012.07.007.

[41] Jean-Philippe Fortin et al. "Removing inter-subject technical variability in magnetic resonance imaging studies". In: *NeuroImage* 132 (2016), pp. 198–212. DOI: 10.1016/j.neuroimage.2016.02.036.

[42] A.F. Frangi et al. "Automatic construction of multiple-object three-dimensional statistical shape models: application to cardiac modeling". en. In: *IEEE Transactions on Medical Imaging* 21.9 (Sept. 2002), pp. 1151–1166. ISSN: 0278-0062. DOI: 10.1109/TMI.2002.804426.

[43] Yabo Fu et al. "Deep learning in medical image registration: a review". en. In: *Physics in Medicine & Biology* 65.20 (Oct. 2020), 20TR01. ISSN: 0031-9155. DOI: 10.1088/1361-6560/ab843e.

[44] Peter M. Full et al. "Studying Robustness of Semantic Segmentation Under Domain Shift in Cardiac MRI". en. In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*. Ed. by Esther Puyol Anton et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 238–249. ISBN: 978-3-030-68107-4. DOI: 10.1007/978-3-030-68107-4_24.

[45] Francesco Galati, Sébastien Ourselin, and Maria A. Zuluaga. "From Accuracy to Reliability and Robustness in Cardiac Magnetic Resonance Image Segmentation: A Review". en. In: *Applied Sciences* 12.8 (Jan. 2022). Number: 8 Publisher: Multidisciplinary Digital Publishing Institute, p. 3936. ISSN: 2076-3417. DOI: 10.3390/app12083936.

[46] Yaroslav Ganin et al. "Domain-Adversarial Training of Neural Networks". en. In: *Domain Adaptation in Computer Vision Applications*. Ed. by Gabriela Csurka. Advances in Computer Vision and Pattern Recognition. Cham: Springer International Publishing, 2017, pp. 189–209. ISBN: 978-3-319-58347-1. DOI: 10.1007/978-3-319-58347-1_10.

[47] Priyanka Garg and Trisha Jain. "A Comparative Study on Histogram Equalization and Cumulative Histogram Equalization". en. In: *International Journal of New Technology and Research* 3.9 (Sept. 2017), p. 263242. ISSN: 2454-4116.

[48] Muhammad Ghifary et al. "Deep Reconstruction-Classification Networks for Unsupervised Domain Adaptation". en. In: *Computer Vision - ECCV 2016*. Ed. by Bastian Leibe et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 597–613. ISBN: 978-3-319-46493-0. DOI: 10.1007/978-3-319-46493-0_36.

[49] Muhammad Ghifary et al. "Domain Generalization for Object Recognition with Multi-task Autoencoders". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. Dec. 2015, pp. 2551–2559. DOI: 10.1109/ICCV.2015.293.

[50] Muhammad Ghifary et al. "Scatter Component Analysis: A Unified Framework for Domain Adaptation and Domain Generalization". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.7 (July 2017), pp. 1414–1430. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2016.2599532.

[51] Vahid Ghodrati et al. "Retrospective respiratory motion correction in cardiac cine MRI reconstruction using adversarial autoencoder and unsupervised learning". en. In: *NMR in Biomedicine* 34.2 (2021), e4433. ISSN: 1099-1492. DOI: 10.1002/nbm.4433.

[52] Judy Wawira Gichoya et al. "AI recognition of patient race in medical imaging: a modelling study". English. In: *The Lancet Digital Health* 4.6 (June 2022), e406–e414. ISSN: 2589-7500. DOI: 10.1016/S2589-7500(22)00063-2.

[53] Ian Goodfellow et al. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., 2014, pp. 2672–2680.

[54] Ritu Gothwal, Shailendra Tiwari, and Shivendra Shivani. "Computational Medical Image Reconstruction Techniques: A Comprehensive Review". en. In: *Archives of Computational Methods in Engineering* 29.7 (Nov. 2022), pp. 5635–5662. ISSN: 1886-1784. DOI: 10.1007/s11831-022-09785-w.

[55] Joost J.M. van Griethuysen et al. "Computational Radiomics System to Decode the Radiographic Phenotype". In: *Cancer Research* 77.21 (2017), e104–e107. DOI: 10.1158/0008-5472.can-17-0339.

[56] Hao Guan and Mingxia Liu. "Domain Adaptation for Medical Image Analysis: A Survey". In: *IEEE Transactions on Biomedical Engineering* 69.3 (2022), pp. 1173–1185. DOI: 10.1109/TBME.2021.3117407.

[57] Yanming Guo et al. "Deep learning for visual understanding: A review". en. In: *Neurocomputing*. Recent Developments on Deep Big Vision 187 (Apr. 2016), pp. 27–48. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2015.09.116.

[58] Muhammad Shafiq ul Hassan et al. "Voxel size and gray level normalization of CT radiomic features in lung cancer". In: *Scientific Reports* 8.1 (2018). DOI: 10.1038/s41598-018-28895-9.

[59] Xingxin He et al. "Multi-Modal Retinal Image Classification With Modality-Specific Attention Network". In: *IEEE Transactions on Medical Imaging* 40.6 (June 2021), pp. 1591–1602. ISSN: 1558-254X. DOI: 10.1109/TMI.2021.3059956.

[60] Center for Devices and Radiological Health. *Digital Health Software Precertification (Pre-Cert) Pilot Program*. en. Sept. 2022. URL: https://www.fda.gov/medical-devices/digital-health-center-excellence/digital-health-software-precertification-pre-cert-pilot-program (visited on 02/28/2023).

[61] Judy Hoffman et al. "Simultaneous Deep Transfer Across Domains and Tasks". en. In: *Domain Adaptation in Computer Vision Applications*. Ed. by Gabriela Csurka. Advances in Computer Vision and Pattern Recognition. Cham: Springer International Publishing, 2017, pp. 173–187. ISBN: 978-3-319-58347-1. DOI: 10.1007/978-3-319-58347-1_9.

[62] Shoubo Hu et al. "Domain Generalization via Multidomain Discriminant Analysis". en. In: *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*. PMLR, Aug. 2020, pp. 292–302.

[63] Xiaoqiong Huang et al. "Style-Invariant Cardiac Image Segmentation with Test-Time Augmentation". en. In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*. Ed. by Esther Puyol Anton et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 305–315. ISBN: 978-3-030-68107-4. DOI: 10.1007/978-3-030-68107-4_31.

[64] Xun Huang and Serge Belongie. "Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 1510–1519. DOI: 10.1109/ICCV.2017.167.

[65] Lars J. Isaksson et al. "Effects of MRI image normalization techniques in prostate cancer radiomics". In: *Physica Medica* 71 (2020), pp. 7–13. DOI: 10.1016/j.ejmp.2020.02.007.

[66] Fabian Isensee et al. "Automatic Cardiac Disease Assessment on cine-MRI via Time-Series Segmentation and Domain Specific Features". en. In: *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*. Ed. by Mihaela Pop et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 120–129. ISBN: 978-3-319-75541-0. DOI: 10.1007/978-3-319-75541-0_13.

[67] Fabian Isensee et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature Methods* 18.2 (Dec. 2020), pp. 203–211. ISSN: 1548-7105. DOI: 10.1038/s41592-020-01008-z.

[68] Mehran Javanmardi and Tolga Tasdizen. "Domain adaptation for biomedical image segmentation using adversarial training". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. Apr. 2018, pp. 554–558. DOI: 10.1109/ISBI.2018.8363637.

[69] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1 (2006), pp. 118–127. DOI: 10.1093/biostatistics/kxj037.

[70] Konstantinos Kamnitsas et al. "Unsupervised Domain Adaptation in Brain Lesion Segmentation with Adversarial Networks". en. In: *Information Processing in Medical Imaging*. Ed. by Marc Niethammer et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017, pp. 597–609. ISBN: 978-3-319-59050-9. DOI: 10.1007/978-3-319-59050-9_47.

[71] Justin Ker et al. "Deep Learning Applications in Medical Image Analysis". In: *IEEE Access* 6 (2018), pp. 9375–9389. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2017.2788044.

[72] Firas Khader et al. "Adaptive Preprocessing for Generalization in Cardiac MR Image Segmentation". en. In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*. Ed. by Esther Puyol Anton et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 269–276. ISBN: 978-3-030-68107-4. DOI: 10.1007/978-3-030-68107-4_27.

[73] Mahendra Khened, Varghese Alex Kollerathu, and Ganapathy Krishnamurthi. "Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers". In: *Medical Image Analysis* 51 (Jan. 2019), pp. 21–45. ISSN: 1361-8415. DOI: 10.1016/j.media.2018.10.004.

[74] Prannay Khosla et al. "Supervised Contrastive Learning". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 18661–18673.

[75] Stefan Klein et al. "elastix: A Toolbox for Intensity-Based Medical Image Registration". In: *IEEE Transactions on Medical Imaging* 29.1 (Jan. 2010), pp. 196–205. ISSN: 1558-254X. DOI: 10.1109/TMI.2009.2035616.

[76] Fanwei Kong and Shawn C. Shadden. "A Generalizable Deep-Learning Approach for Cardiac Magnetic Resonance Image Segmentation Using Image Augmentation and Attention U-Net". en. In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*. Ed. by Esther Puyol Anton et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 287–296. ISBN: 978-3-030-68107-4. DOI: 10.1007/978-3-030-68107-4_29.

[77] Edward Korot et al. "Predicting sex from retinal fundus photographs using automated deep learning". en. In: *Scientific Reports* 11.1 (May 2021). Number: 1 Publisher: Nature Publishing Group, p. 10286. ISSN: 2045-2322. DOI: 10.1038/s41598-021-89743-x.

[78] Tanja Kurzendorfer et al. "Left ventricle segmentation in LGE-MRI using multiclass learning". In: *Medical Imaging 2019: Image Processing*. Vol. 10949. SPIE, Mar. 2019, pp. 585–590. DOI: 10.1117/12.2511610.

[79] Kaisar Kushibar et al. "Supervised Domain Adaptation for Automatic Subcortical Brain Structure Segmentation with Minimal User Interaction". en. In: *Scientific Reports* 9.1 (May 2019). Number: 1 Publisher: Nature Publishing Group, p. 6742. ISSN: 2045-2322. DOI: 10.1038/s41598-019-43299-z.

[80] Thomas Küstner et al. "CINENet: deep learning-based 3D cardiac CINE MRI reconstruction with multi-coil complex-valued 4D spatio-temporal convolutions". en. In: *Scientific Reports* 10.1 (Aug. 2020). Number: 1 Publisher: Nature Publishing Group, p. 13710. ISSN: 2045-2322. DOI: 10.1038/s41598-020-70551-8.

[81] Maxime W. Lafarge et al. "Domain-Adversarial Neural Networks to Address the Appearance Variability of Histopathology Images". en. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Ed. by M. Jorge Cardoso et al. Lecture Notes in Computer Science. Cham:

Springer International Publishing, 2017, pp. 83–91. ISBN: 978-3-319-67558-9. DOI: 10.1007/978-3-319-67558-9_10.

[82] Alain Lalande et al. "Emidec: A Database Usable for the Automatic Evaluation of Myocardial Infarction from Delayed-Enhancement Cardiac MRI". en. In: *Data* 5.4 (Dec. 2020). Number: 4 Publisher: Multidisciplinary Digital Publishing Institute, p. 89. ISSN: 2306-5729. DOI: 10.3390/data5040089.

[83] Joonsang Lee et al. "Radiomics feature robustness as measured using an MRI phantom". In: *Scientific Reports* 11.1 (2021). DOI: 10.1038/s41598-021-83593-3.

[84] Constance D. Lehman et al. "Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection". In: *JAMA Internal Medicine* 175.11 (Nov. 2015), pp. 1828–1837. ISSN: 2168-6106. DOI: 10.1001/jamainternmed.2015.5231.

[85] Karim Lekadir et al. "FUTURE-AI: Guiding Principles and Consensus Recommendations for Trustworthy Artificial Intelligence in Medical Imaging". In: *arXiv preprint arXiv:2109.09658* (Sept. 2021). DOI: 10.48550/arXiv.2109.09658.

[86] Chenxin Li et al. "Domain generalization on medical imaging classification using episodic training with task augmentation". In: *Computers in Biology and Medicine* 141 (Feb. 2022), p. 105144. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2021.105144.

[87] Haoliang Li et al. "Domain Generalization with Adversarial Feature Learning". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. June 2018, pp. 5400–5409. DOI: 10.1109/CVPR.2018.00566.

[88] Hongwei Li, Jianguo Zhang, and Bjoern Menze. "Generalisable Cardiac Structure Segmentation via Attentional and Stacked Image Adaptation". en. In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*. Ed. by Esther Puyol Anton et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 297–304. ISBN: 978-3-030-68107-4. DOI: 10.1007/978-3-030-68107-4_30.

[89] Hongwei Li, Andrii Zhygallo, and Bjoern Menze. "Automatic Brain Structures Segmentation Using Deep Residual Dilated U-Net". en. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Ed. by Alessandro Crimi et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 385–393. ISBN: 978-3-030-11723-8. DOI: 10.1007/978-3-030-11723-8_39.

[90] Lei Li et al. "Random Style Transfer Based Domain Generalization Networks Integrating Shape and Spatial Information". en. In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*. Ed. by Esther Puyol Anton et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 208–218. ISBN: 978-3-030-68107-4. DOI: 10.1007/978-3-030-68107-4_21.

[91] Ya Li et al. "Deep Domain Generalization via Conditional Invariant Adversarial Networks". en. In: *Computer Vision - ECCV 2018*. Ed. by Vittorio Ferrari et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 647–663. ISBN: 978-3-030-01267-0. DOI: 10.1007/978-3-030-01267-0_38.

[92] Ya Li et al. "Domain Generalization via Conditional Invariant Representations". en. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (Apr. 2018). ISSN: 2374-3468. DOI: 10.1609/aaai.v32i1.11682.

[93] Ming-Yu Liu and Oncel Tuzel. "Coupled Generative Adversarial Networks". In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc., 2016.

[94] Xiao Liu et al. "Disentangled Representations for Domain-Generalized Cardiac Segmentation". en. In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*. Ed. by Esther Puyol Anton et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 187–195. ISBN: 978-3-030-68107-4. DOI: 10.1007/978-3-030-68107-4_19.

[95] Xiao Liu et al. "Semi-supervised Meta-learning with Disentanglement for Domain-Generalised Medical Image Segmentation". en. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Ed. by Marleen de Bruijne et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 307–317. ISBN: 978-3-030-87196-3. DOI: 10.1007/978-3-030-87196-3_29.

[96] Xiao Liu et al. "vMFNet: Compositionality Meets Domain-Generalised Segmentation". en. In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022*. Ed. by Linwei Wang et al. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022, pp. 704–714. ISBN: 978-3-031-16449-1. DOI: 10.1007/978-3-031-16449-1_67.

[97] Zhendong Liu et al. "Remove Appearance Shift for Ultrasound Image Segmentation via Fast and Universal Style Transfer". In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. Apr. 2020, pp. 1824–1828. DOI: 10.1109/ISBI45749.2020.9098457.

[98] Mingsheng Long et al. "Deep Transfer Learning with Joint Adaptation Networks". en. In: *Proceedings of the 34th International Conference on Machine Learning*. PMLR, July 2017, pp. 2208–2217.

[99] Mingsheng Long et al. "Learning Transferable Features with Deep Adaptation Networks". en. In: *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, June 2015, pp. 97–105.

[100] Mingsheng Long et al. "Unsupervised Domain Adaptation with Residual Transfer Networks". In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc., 2016.

[101] Pablo Arratia López et al. "WarpPINN: Cine-MR image registration with physics-informed neural networks". In: *arXiv preprint arXiv:2211.12549* (Nov. 2022). DOI: 10.48550/arXiv.2211.12549.

[102] Chunwei Ma, Zhanghexuan Ji, and Mingchen Gao. "Neural Style Transfer Improves 3D Cardiovascular MR Image Segmentation on Inconsistent Data". en. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by Dinggang Shen et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 128–136. ISBN: 978-3-030-32245-8. DOI: 10.1007/978-3-030-32245-8_15.

[103] Jun Ma. "Histogram Matching Augmentation for Domain Adaptation with Application to Multi-centre, Multi-vendor and Multi-disease Cardiac Image Segmentation". en. In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*. Ed. by Esther Puyol Anton et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 177–186. ISBN: 978-3-030-68107-4. DOI: 10.1007/978-3-030-68107-4_18.

[104] Ali Madani et al. "Semi-supervised learning with generative adversarial networks for chest X-ray classification with ability of data domain adaptation". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. Apr. 2018, pp. 1038–1042. DOI: 10.1109/ISBI.2018.8363749.

[105] Faisal Mahmood, Richard Chen, and Nicholas J. Durr. "Unsupervised Reverse Domain Adaptation for Synthetic Medical Images via Adversarial Training". In: *IEEE Transactions on Medical Imaging* 37.12 (Dec. 2018), pp. 2572–2581. ISSN: 1558-254X. DOI: 10.1109/TMI.2018.2842767.

[106] Carlos Martin-Isla et al. "Image-Based Cardiac Diagnosis With Machine Learning: A Review". In: *Frontiers in Cardiovascular Medicine* 7 (Jan. 2020), p. 1. ISSN: 2297-055X. DOI: 10.3389/fcvm.2020.00001.

[107] Ninareh Mehrabi et al. "A Survey on Bias and Fairness in Machine Learning". In: *ACM Computing Surveys* 54.6 (July 2021), 115:1–115:35. ISSN: 0360-0300. DOI: 10.1145/3457607.

[108] Xueyan Mei et al. "RadImageNet: An Open Radiologic Deep Learning Research Dataset for Effective Transfer Learning". In: *Radiology: Artificial Intelligence* 4.5 (Sept. 2022), e210315. DOI: 10.1148/ryai.210315.

[109] Rosa-María Menchón-Lara et al. "Efficient convolution-based pairwise elastic image registration on three multimodal similarity metrics". en. In: *Signal Processing* 202 (Jan. 2023), p. 108771. ISSN: 0165-1684. DOI: 10.1016/j.sigpro.2022.108771.

[110] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. "Domain Generalization via Invariant Feature Representation". en. In: *Proceedings of the 30th International Conference on Machine Learning*. PMLR, Feb. 2013, pp. 10–18.

[111] László G. Nyúl and Jayaram K. Udupa. "On standardizing the MR image intensity scale". In: *Magnetic Resonance in Medicine* 42.6 (1999), pp. 1072–1081. DOI: 10.1002/(sici)1522-2594(199912)42:6<1072::aid-mrm11>3.0.co;2-m.

[112] Ziad Obermeyer and Ezekiel J. Emanuel. "Predicting the Future — Big Data, Machine Learning, and Clinical Medicine". In: *New England Journal of Medicine* 375.13 (Sept. 2016), pp. 1216–1219. ISSN: 0028-4793. DOI: 10.1056/NEJMp1606181.

[113] Ozan Oktay et al. "Attention U-Net: Learning Where to Look for the Pancreas". In: *arXiv preprint arXiv:1804.03999* (May 2018). DOI: 10.48550/arXiv.1804.03999.

[114] Ozan Oktay et al. "Multi-input Cardiac Image Super-Resolution Using Convolutional Neural Networks". en. In: *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016*. Ed. by Sebastien Ourselin et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 246–254. ISBN: 978-3-319-46726-9. DOI: 10.1007/978-3-319-46726-9_29.

[115] Natsuko Onishi et al. "Ultrafast dynamic contrast-enhanced breast MRI may generate prognostic imaging markers of breast cancer". en. In: *Breast Cancer Research* 22.1 (May 2020), p. 58. ISSN: 1465-542X. DOI: 10.1186/s13058-020-01292-9.

[116] Maxime Oquab et al. "Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. June 2014, pp. 1717–1724. DOI: 10.1109/CVPR.2014.222.

[117] Fanny Orlhac et al. "How can we combat multicenter variability in MR radiomics? Validation of a correction procedure". In: *European Radiology* (2020). DOI: 10.1007/s00330-020-07284-9.

[118] Catherine M. Otto. *The Practice of Clinical Echocardiography*. ClinicalKey 2012. Elsevier/Saunders, 2012. ISBN: 9781437727654.

[119] Sinno Jialin Pan et al. "Domain Adaptation via Transfer Component Analysis". In: *IEEE Transactions on Neural Networks* 22.2 (Feb. 2011), pp. 199–210. ISSN: 1941-0093. DOI: 10.1109/TNN.2010.2091281.

[120] Mario Parreño, Roberto Paredes, and Alberto Albiol. "Deidentifying MRI Data Domain by Iterative Backpropagation". en. In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*. Ed. by Esther Puyol Anton et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 277–286. ISBN: 978-3-030-68107-4. DOI: 10.1007/978-3-030-68107-4_28.

[121] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[122] Antonio Pepe et al. "Detection, segmentation, simulation and visualization of aortic dissections: A review". In: *Medical Image Analysis* 65 (2020), p. 101773. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2020.101773.

[123] Christian S. Perone et al. "Unsupervised domain adaptation for medical imaging segmentation with self-ensembling". en. In: *NeuroImage* 194 (July 2019), pp. 1–11. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2019.03.026.

[124] Caroline Petitjean et al. "Right ventricle segmentation from cardiac MRI: A collation study". In: *Medical Image Analysis* 19.1 (2015), pp. 187 –202. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2014.10.004.

[125] Chi-Hieu Pham et al. "Multiscale brain MRI super-resolution using deep 3D convolutional networks". en. In: *Computerized Medical Imaging and Graphics* 77 (Oct. 2019), p. 101647. ISSN: 0895-6111. DOI: 10.1016/j.compmedimag.2019.101647.

[126] Guillem Pons-Lladó, ed. *Protocols for Cardiac MR and CT*. en. Cham: Springer International Publishing, 2016. ISBN: 978-3-319-30830-2 978-3-319-30831-9. DOI: 10.1007/978-3-319-30831-9.

[127] Ferran Prados et al. "Spinal cord grey matter segmentation challenge". en. In: *NeuroImage* 152 (May 2017), pp. 312–329. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2017.03.010.

[128] Esmeralda Ruiz Pujadas et al. "Prediction of incident cardiovascular events using machine learning and CMR radiomics". en. In: *European Radiology* 33.5 (May 2023), pp. 3488–3500. ISSN: 1432-1084. DOI: 10.1007/s00330-022-09323-z.

[129] Esther Puyol-Antón et al. "Fairness in Cardiac Magnetic Resonance Imaging: Assessing Sex and Racial Bias in Deep Learning-Based Segmentation". In: *Frontiers in Cardiovascular Medicine* 9 (2022). ISSN: 2297-055X. DOI: 10.3389/fcvm.2022.859310.

[130] Esther Puyol-Antón et al. "Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation". en. In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021*. Ed. by Marleen de Bruijne et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 413–423. ISBN: 978-3-030-87199-4. DOI: 10.1007/978-3-030-87199-4_39.

[131] Perry Radau et al. "Evaluation Framework for Algorithms Segmenting Short Axis Cardiac MRI." In: *The MIDAS Journal* (July 2009). DOI: 10.54294/g80ruo.

[132] Zahra Raisi-Estabragh et al. "Repeatability of Cardiac Magnetic Resonance Radiomics: A Multi-Centre Multi-Vendor Test-Retest Study". In: *Frontiers in Cardiovascular Medicine* 7 (2020). DOI: 10.3389/fcvm.2020.586236.

[133] *Regulatory framework proposal on artificial intelligence | Shaping Europe's digital future*. en. Feb. 2023. URL: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (visited on 02/28/2023).

[134] Jacob C. Reinhold et al. "Evaluating the impact of intensity normalization on MR image synthesis". In: *Medical Imaging 2019: Image Processing*. Ed. by Elsa D. Angelini and Bennett A. Landman. SPIE, 2019. DOI: 10.1117/12.2513089.

[135] Stephan R. Richter et al. "Playing for Data: Ground Truth from Computer Games". en. In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 102–118. ISBN: 978-3-319-46475-6. DOI: 10.1007/978-3-319-46475-6_7.

[136] Stephen J. Riederer, Eric G. Stinson, and Paul T. Weavers. "Technical Aspects of Contrast-enhanced MR Angiography: Current Status and New Applications". In: *Magnetic Resonance in Medical Sciences* 17.1 (2018), pp. 3–12. DOI: 10.2463/mrms.rev.2017-0053.

[137] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Lecture Notes in Computer Science*. Springer International Publishing, 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

[138] German Ros et al. "The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[139] Mina Saber, Dina Abdelrauof, and Mustafa Elattar. "Multi-center, Multi-vendor, and Multi-disease Cardiac Image Segmentation Using Scale-Independent Multi-gate UNET". en. In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*. Ed. by Esther Puyol Anton et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 259–268. ISBN: 978-3-030-68107-4. DOI: 10.1007/978-3-030-68107-4_26.

[140] Veit Sandfort et al. "Reliable segmentation of 2D cardiac magnetic resonance perfusion image sequences using time as the 3rd dimension". en. In: *European Radiology* 31.6 (June 2021), pp. 3941–3950. ISSN: 1432-1084. DOI: 10.1007/s00330-020-07474-5.

[141] Cian M. Scannell, Amedeo Chiribiri, and Mitko Veta. "Domain-Adversarial Learning for Multi-Centre, Multi-Vendor, and Multi-Disease Cardiac MR Image Segmentation". en. In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*. Ed. by Esther Puyol Anton et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 228–237. ISBN: 978-3-030-68107-4. DOI: 10.1007/978-3-030-68107-4_23.

[142] Thomas Schaffter et al. "Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms". In: *JAMA Network Open* 3.3 (Mar. 2020), e200265. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2020.0265.

[143] Maxime Sermesant et al. "Applications of artificial intelligence in cardiovascular imaging". en. In: *Nature Reviews Cardiology* 18.8 (Aug. 2021). Number: 8 Publisher: Nature Publishing Group, pp. 600–609. ISSN: 1759-5010. DOI: 10.1038/s41569-021-00527-2.

[144] Divya Shanmugam et al. "Better Aggregation in Test-Time Augmentation". en. In: *Proceedings of the IEEE international conference on computer vision*. 2021, pp. 1214–1223.

[145] Avan Suinesiaputra et al. "A collaborative resource to build consensus for automated left ventricular segmentation of cardiac MR images". In: *Medical Image Analysis* 18.1 (Jan. 2014), pp. 50–62. ISSN: 1361-8415. DOI: 10.1016/j.media.2013.09.001.

[146] Baochen Sun and Kate Saenko. "Deep CORAL: Correlation Alignment for Deep Domain Adaptation". en. In: *Computer Vision - ECCV 2016 Workshops*. Ed. by Gang Hua and Hervé Jégou. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 443–450. ISBN: 978-3-319-49409-8. DOI: 10.1007/978-3-319-49409-8_35.

[147] Huaiqiang Sun et al. "Psychoradiologic Utility of MR Imaging for Diagnosis of Attention Deficit Hyperactivity Disorder: A Radiomics Analysis". In: *Radiology* 287.2 (2018), pp. 620–630. DOI: 10.1148/radiol.2017170226.

[148] Qian Tao et al. "Deep Learning-based Method for Fully Automatic Quantification of Left Ventricle Function from Cine MR Images: A Multivendor, Multicenter Study". In: *Radiology* 290.1 (2019), pp. 81–88. DOI: 10.1148/radiol.2018180513.

[149] Daniel Shu Wei Ting et al. "Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes". In: *JAMA* 318.22 (Dec. 2017), pp. 2211–2223. ISSN: 0098-7484. DOI: 10.1001/jama.2017.18152.

[150] Antonio Torralba and Alexei A. Efros. "Unbiased look at dataset bias". In: *CVPR 2011*. June 2011, pp. 1521–1528. DOI: 10.1109/CVPR.2011.5995347.

[151] Phi Vu Tran. "A Fully Convolutional Neural Network for Cardiac Segmentation in Short-Axis MRI". In: *arXiv preprint arXiv:1604.00494* (Apr. 2017). DOI: 10.48550/arXiv.1604.00494.

[152] Alberto Traverso et al. "Repeatability and Reproducibility of Radiomic Features: A Systematic Review". In: *International Journal of Radiation Oncology * Biology * Physics* 102.4 (2018), pp. 1143–1158. ISSN: 0360-3016. DOI: 10.1016/j.ijrobp.2018.05.053.

[153] Eric Tzeng et al. "Adversarial Discriminative Domain Adaptation". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017, pp. 2962–2971. DOI: 10.1109/CVPR.2017.316.

[154] Eric Tzeng et al. "Deep Domain Confusion: Maximizing for Domain Invariance". In: *arXiv preprint arXiv:1412.3474* (Dec. 2014). DOI: 10.48550/arXiv.1412.3474.

[155] Johannes Uhlig et al. "Gadolinium-based Contrast Agents for Cardiac MRI: Use of Linear and Macrocyclic Agents with Associated Safety Profile from 154 779 European Patients". In: *Radiology: Cardiothoracic Imaging* 2.5 (Oct. 2020), e200102. DOI: 10.1148/ryct.2020200102.

[156] Hyemin Um et al. "Impact of image preprocessing on the scanner dependence of multi-parametric MRI radiomic features and covariate shift in multi-institutional glioblastoma datasets". In: *Physics in Medicine & Biology* 64.16 (2019), p. 165011. DOI: 10.1088/1361-6560/ab2f44.

[157] Roshan Reddy Upendra, Richard Simon, and Cristian A Linte. "A Deep Learning Framework for Image Super-Resolution for Late Gadolinium Enhanced Cardiac MRI". In: *2021 Computing in Cardiology (CinC)*. Vol. 48. Sept. 2021, pp. 1–4. DOI: 10.23919/CinC53138.2021.9662790.

[158] Roshan Reddy Upendra, Richard Simon, and Cristian A. Linte. "Joint deep learning framework for image registration and segmentation of late gadolinium enhanced MRI and cine cardiac MRI". In: *Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling*. Vol. 11598. SPIE, Feb. 2021, pp. 96–103. DOI: 10.1117/12.2581386.

[159] M. Usman et al. "Free breathing whole-heart 3D CINE MRI with self-gated Cartesian trajectory". en. In: *Magnetic Resonance Imaging* 38 (May 2017), pp. 129–137. ISSN: 0730-725X. DOI: 10.1016/j.mri.2016.12.021.

[160] Stéfan van der Walt et al. "scikit-image: image processing in Python". In: *PeerJ* 2 (2014), e453. DOI: 10.7717/peerj.453.

[161] Christopher L. Welle, Flavius F. Guglielmo, and Sudhakar K. Venkatesh. "MRI of the liver: choosing the right contrast agent". en. In: *Abdominal Radiology* 45.2 (Feb. 2020), pp. 384–392. ISSN: 2366-0058. DOI: 10.1007/s00261-019-02162-5.

[162] Chenchu Xu et al. "Contrast agent-free synthesis and segmentation of ischemic heart disease images using progressive sequential causal GANs". en. In: *Medical Image Analysis* 62 (May 2020), p. 101668. ISSN: 1361-8415. DOI: 10.1016/j.media.2020.101668.

[163] Zheng Xu et al. "Exploiting Low-Rank Structure from Latent Domains for Domain Generalization". en. In: *Computer Vision - ECCV 2014*. Ed. by David Fleet et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014, pp. 628–643. ISBN: 978-3-319-10578-9. DOI: 10.1007/978-3-319-10578-9_41.

[164] Jun Yang, Rong Yan, and Alexander G. Hauptmann. "Cross-domain video concept detection using adaptive svms". In: *Proceedings of the 15th ACM international conference on Multimedia*. MM '07. New York, NY, USA: Association for Computing Machinery, Sept. 2007, pp. 188–197. ISBN: 978-1-59593-702-5. DOI: 10.1145/1291233.1291276.

[165] Li Yao et al. "A Strong Baseline for Domain Adaptation and Generalization in Medical Imaging". In: *arXiv preprint arXiv:1904.01638* (Apr. 2019). DOI: 10.48550/arXiv.1904.01638.

[166] Qi Ying et al. "Multi-Modal Data Analysis for Alzheimer's Disease Diagnosis: An Ensemble Model Using Imagery and Genetic Features". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. Nov. 2021, pp. 3586–3591. DOI: 10.1109/EMBC46164.2021.9630174.

[167] Stephen S F Yip and Hugo J W L Aerts. "Applications and limitations of radiomics". In: *Physics in Medicine and Biology* 61.13 (2016), R150–R166. DOI: 10.1088/0031-9155/61/13/r150.

[168] Tom Young et al. "Recent Trends in Deep Learning Based Natural Language Processing [Review Article]". In: *IEEE Computational Intelligence Magazine* 13.3 (Aug. 2018), pp. 55–75. ISSN: 1556-6048. DOI: 10.1109/MCI.2018.2840738.

[169] Kun-Hsing Yu, Andrew L. Beam, and Isaac S. Kohane. "Artificial intelligence in healthcare". en. In: *Nature Biomedical Engineering* 2.10 (Oct. 2018). Number: 10 Publisher: Nature Publishing Group, pp. 719–731. ISSN: 2157-846X. DOI: 10.1038/s41551-018-0305-z.

[170] Kun-Hsing Yu and Isaac S. Kohane. "Framing the challenges of artificial intelligence in medicine". en. In: *BMJ Quality & Safety* 28.3 (Mar. 2019). Publisher: BMJ Publishing Group Ltd Section: Viewpoint, pp. 238–241. ISSN: 2044-5415, 2044-5423. DOI: 10.1136/bmjqs-2018-008551.

[171] Qian Yue et al. "Cardiac Segmentation from LGE MRI Using Deep Neural Network Incorporating Shape and Spatial Priors". en. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by Dinggang Shen et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 559–567. ISBN: 978-3-030-32245-8. DOI: 10.1007/978-3-030-32245-8_62.

[172] Fatemeh Zabihollahy et al. "Fully automated segmentation of left ventricular scar from 3D late gadolinium enhancement magnetic resonance imaging using a cascaded multi-planar U-Net (CMPU-Net)". en. In: *Medical Physics* 47.4 (2020), pp. 1645–1655. ISSN: 2473-4209. DOI: 10.1002/mp.14022.

[173] John R. Zech et al. "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study". en. In: *PLOS Medicine* 15.11 (Nov. 2018), e1002683. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1002683.

[174] Jing Zhang et al. "Recent Advances in Transfer Learning for Cross-Dataset Visual Recognition: A Problem-Oriented Perspective". In: *ACM Computing Surveys* 52.1 (Feb. 2019), 7:1–7:38. ISSN: 0360-0300. DOI: 10.1145/3291124.

[175] Yao Zhang et al. "Semi-supervised Cardiac Image Segmentation via Label Propagation and Style Transfer". en. In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*. Ed. by Esther Puyol Anton et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 219–227. ISBN: 978-3-030-68107-4. DOI: 10.1007/978-3-030-68107-4_22.

[176] Yue Zhang et al. "Task Driven Generative Modeling for Unsupervised Domain Adaptation: Application to X-ray Image Segmentation". en. In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018*. Ed. by Alejandro F. Frangi et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 599–607. ISBN: 978-3-030-00934-2. DOI: 10.1007/978-3-030-00934-2_67.

[177] Fan Zhou et al. "Domain generalization via optimal transport with metric similarity learning". en. In: *Neurocomputing* 456 (Oct. 2021), pp. 469–480. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2020.09.091.

[178] Jun-Yan Zhu et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232.

[179] Xiahai Zhuang. "Multivariate Mixture Model for Cardiac Segmentation from Multi-Sequence MRI". en. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*. Ed. by Sebastien Ourselin et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 581–588. ISBN: 978-3-319-46723-8. DOI: 10.1007/978-3-319-46723-8_67.

[180] Xiahai Zhuang. "Multivariate Mixture Model for Myocardial Segmentation Combining Multi-Source Images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.12 (Dec. 2019), pp. 2933–2946. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2018.2869576.

[181] Xiahai Zhuang et al. "Cardiac segmentation on late gadolinium enhancement MRI: A benchmark study from multi-sequence cardiac MR segmentation challenge". en. In: *Medical Image Analysis* 81 (Oct. 2022), p. 102528. ISSN: 1361-8415. DOI: 10.1016/j.media.2022.102528.

[182] Alex Zwanenburg et al. "The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping". In: *Radiology* 295.2 (2020), pp. 328–338. ISSN: 1527-1315. DOI: 10.1148/radiol.2020191145.

Generalizability in multi-centre cardiac image analysis with machine learning

In recent years we are seeing how Artificial Intelligence (AI) is changing the world with powerful tools capable of solving complex tasks. However, to incorporate these tools into our daily life, we need to understand them better and identify potential failures.

In this Thesis, we provide a self-contained overview of current challenges and applications of AI in cardiovascular imaging research and present a series of results targeted at solving the domain shift problem, that causes models to be biased on new domains.

If securely implemented, AI has the potential to democratize complex tools and algorithms that are usually in a few hands.