University of London

Imperial College of Science, Technology and Medicine

Department of Computing

Machine Learning Approaches to Model Cardiac Shape in Large-Scale Imaging Studies

Carlo Biffi

Abstract

Recent improvements in non-invasive imaging, together with the introduction of fully-automated segmentation algorithms and big data analytics, has paved the way for large-scale population-based imaging studies. These studies promise to increase our understanding of a large number of medical conditions, including cardiovascular diseases. However, analysis of cardiac shape in such studies is often limited to simple morphometric indices, ignoring large part of the information available in medical images. Discovery of new biomarkers by machine learning has recently gained traction, but often lacks interpretability. The research presented in this thesis aimed at developing novel explainable machine learning and computational methods capable of better summarizing shape variability, to better inform association and predictive clinical models in large-scale imaging studies.

A powerful and flexible framework to model the relationship between three-dimensional (3D) cardiac atlases, encoding multiple phenotypic traits, and genetic variables is first presented. The proposed approach enables the detection of regional phenotype-genotype associations that would be otherwise neglected by conventional association analysis. Three learning-based systems based on deep generative models are then proposed. In the first model, I propose a classifier of cardiac shapes which exploits task-specific generative shape features, and it is designed to enable the visualisation of the anatomical effect these features encode in 3D, making the classification task transparent. The second approach models a database of anatomical shapes via a hierarchy of conditional latent variables and it is capable of detecting, quantifying and visualising onto a template shape the most discriminative anatomical features that characterize distinct clinical conditions. Finally, a preliminary analysis of a deep learning system capable of reconstructing 3D high-resolution cardiac segmentations from a sparse set of 2D views segmentations is reported. This thesis demonstrates that machine learning approaches can facilitate high-throughput analysis of normal and pathological anatomy and of its determinants without losing clinical interpretability.

Ai miei genitori

Declaration of Originality

I, Carlo Biffi, declare that the work presented in this thesis is my own, unless specifically acknowledged.

Acknowledgements

I would to express my gratitude to my supervisors, Daniel Rueckert and Declan O'Regan, for the amazing opportunity of carrying out my PhD research alongside them. The combination of their exceptional knowledge of the field, invaluable advice and support has greatly inspired and immensely helped me throughout this journey.

There is a long list of friends and colleagues at the the BioMedIA group and the Robert Steiner MR Unit who extensively helped and taught me in these years, and with whom I share a lot of happy and fun memories. I am extremely thankful to everyone. A special mention goes to Giacomo Tarroni, Wenjia Bai, Antonio de Marvao, Ozan Oktay and Juan Cerrolaza whose precious contribution and feedback has significantly boosted the quality of the research work presented in this thesis. Many thanks to Maxime Sermesant and David Firmin for their feedback during the examination process. I would also like to thank the British Heart Foundation for funding my research.

Finally, last but not least, I would like to deeply thank my parents for their unconditional support and encouragement throughout all my studies. And to Marta, who always believed in me and embarked with me in this adventure almost five years ago, no words can express my gratitude and happiness of having her be a part of my life.

Acronyms

2D two-dimensional

3D three-dimensional

AD Alzheimer's Disease

AE Autoencoder

ASM Active Shape Model

AHA American Heart Association

bSSFP balanced Steady-State Free Precession

BSA Body Surface Area

CMR Cardiovascular Magnetic Resonance

CNN Convolutional Neural Network

CVAE Conditional Variational Autoencoder

ED End-Diastole

ES End-Systole

FDR False Discovery Rate

GLM General Linear Model

GPU Graphics Processing Unit

GWAS Genome-wide Association Study

HC Healthy Control

HCM Hypertrophic Cardiomyopathy

HVol Healthy Volunteer

LAX Long Axis

LV Left Ventricle

LVM Left Ventricular Mass

LVCV Left Ventricular Cavity Volume

LVV Left Ventricular Volume

MLP Multi-layered Perceptron

MR Magnetic Resonance

MRI Magnetic Resonance Imaging

NN Neural Network

PCA Principal Components Analysis

TFCE Threshold-Free Cluster Enhancement

SAX Short Axis

SBP Systolic Blood Pressure

SDM Statistical Deformation Model

SNP Single Nucleotide Polymorphism

SPM Statistical Parametric Map

SSM Statistical Shape Model

LVAE Ladder Variational Autoencoder

VAE Variational Autoencoder

WT Wall Thickness

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This chapter introduces the motivation for the research undertaken in this work and aims at providing the reader with all the necessary clinical background. In Section 1.1, notions of human cardiovascular anatomy and pathology are introduced. The main class of cardiac diseases studied in this work, cardiomyopathies, are also reviewed. This is followed, in Section 1.2, by an explanation of the clinical role of cardiac imaging and of imaging-derived clinical indices, with a special focus on cardiovascular magnetic resonance (CMR) imaging. Section 1.3 summarises the history and motivation for large-scale population-based cardiac imaging studies, while Section 1.5 introduces the challenges related with modeling cardiac shape in such studies. Finally, this chapter is concluded with a summary of the objectives and contributions of this research work (Section 1.6), and with an overview of thesis content (Section 1.7).

1.1 The Human Heart

The human heart is a hollow, muscular organ located in the thoracic cavity and consisting of four chambers: the left and right atria and the left and right ventricle. A diagram of the human heart anatomy is shown in Figure 1.1. The heart plays a key role in the
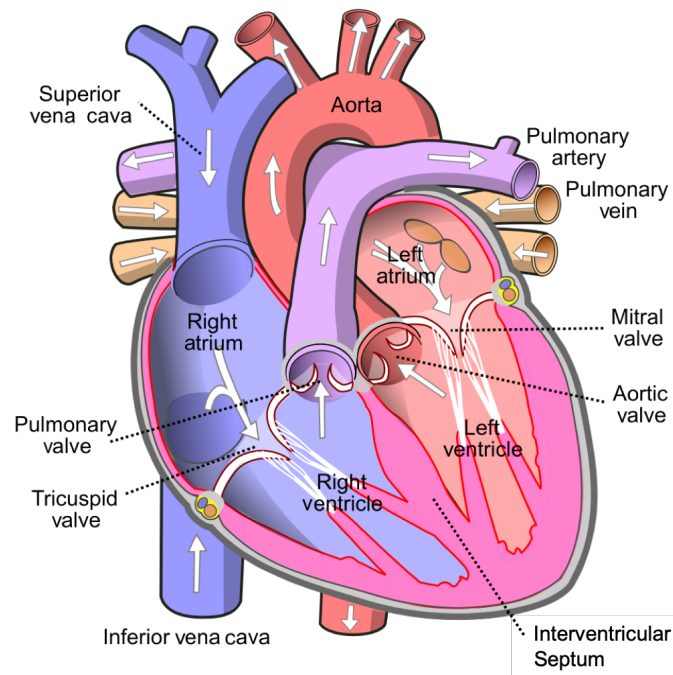
Figure 1.1: Anterior cross-section of the human heart.[1]

cardiovascular system by propelling blood through the human body. This is carried out via two functionally distinct pathways: the systemic circuit and the pulmonary circuit. The pulmonary circuit carries deoxygenated blood from the right ventricle to the lungs, via the pulmonary artery; and returns oxygenated blood to the left atrium, via the pulmonary veins. The oxygen-rich blood is then pumped, via the mitral valve, from the left atrium to the left ventricle (LV), where the systemic circuit starts. The systemic circuit is responsible for delivering, through the aorta and the body arteries, oxygen and nutrients to the tissues of the body from the LV. The systemic circuit is completed when deoxygenated blood returns to the right atria through the body veins and is again provided to pulmonary circuit.

The pumping motion of the heart is triggered by electrical stimuli in pacemaker cells of the right atrium, which directly control the heart rate. During each heart beat, a contraction phase of the atria and ventricles to pump blood out to the pulmonary arteries and to the aorta, named systole, is followed by a relaxation phase, named diastole. In normal

---

[1]Figure adapted from Wikipedia: https://en.wikipedia.org/wiki/Heart

conditions, the heart beats between 65 to 75 times per minute. In case it stops pumping blood, the cells of the body will not be able to survive without the necessary nutrients, resulting in cell death.

1.1.1 The Left Ventricle of the Human Heart

This work focusses on the analysis of the left ventricle. The LV is a conical cavity, often represented schematically as a prolate ellipsoid, located in the bottom left portion of the heart, below the left atrium and with the right ventricle hugging it [1]. As shown in Figure 1.1, separating the right and left ventricle (RV and LV) is the interventricular septum, which is concave in shape and bulges into the right ventricle. A smooth inlet portion that contains the mitral valve apparatus and a smooth outlet portion leading to the aortic valve are located at the top of the LV.

In imaging protocols, the line that passes through the center of the mitral valve orifice and the apex of the LV is commonly referred to as the long axis of the heart, while the axis perpendicular to it is named the short axis of the heart [2, 3]. To facilitate clinical studies and diagnosis, the American Heart Association (AHA) 17-segment model is conventionally used to divide the LV into 17 distinct regional segments (Figure 1.2). In particular, based on autopsy data, the AHA 17-segment model divides the LV into equal thirds perpendicular to its long axis to create three circular sections named basal, mid-cavity, and apical. Then, the model further divides the basal and mid-cavity sections into six segments, the apical section into four, and the final, seventeenth segment represents the apical cap, or apex. The LV apex contains the true muscle at the extreme tip of the ventricle where there is no longer cavity present.

The LV, as the rest of cardiac muscle, consists of three layers: the epicardium (forming the outermost wall), the myocardium (the middle layer) and the endocardium (the innermost layer, which has contact with the blood-pool). The myocardium is the thickest of the
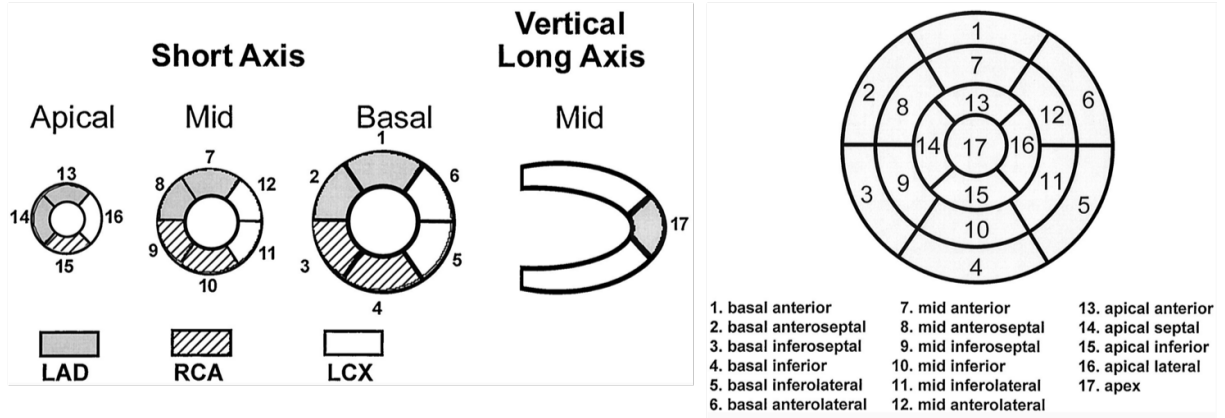
Figure 1.2: The American Heart Association's 17-Segment Model (AHA-17). Figure adapted from [4].

three layers, mainly composed of cardiomyocytes. The LV is responsible for pumping oxygenated blood all over the body, developing a much higher pressure than the other heart chambers, and for this reason is the thickest of the heart's chambers. The healthy LV myocardial wall reaches the maximum thickening at its base, while thinning to only 1-2 mm at its apex. Septal LV wall thickness (WT), which is of particular importance in the diagnosis of many cardiovascular diseases, is on average 7-12mm in healthy adults [5, 3].

1.1.2 Cardiovascular Diseases and Cardiomyopathies

Cardiovascular diseases are the leading cause of morbidity and mortality globally, representing the 31% of all deaths worldwide (17.9 million) in 2016, according to the World Health Organization (WHO) [6]. Heart failure, a clinical condition in which the heart is unable to pump blood to meet the body's needs, is among the main contributors to this public health issue, with an estimated prevalence of 23 million subjects worldwide [7]. Heart failure is a final common stage of several cardiovascular diseases including cardiomyopathies, hypertension, and coronary or valvular heart disease, which adversely modify the structure or the function of the heart, provoking this impairment. Due to the advances in the understanding of the biological mechanisms driving it, slowing or revers-

ing heart failure has now become an important therapy goal. This makes the quantitative assessment of the heart changes over time associated to this condition crucial.

Cardiomyopathies are a common global set of cardiovascular diseases of the myocardium which manifest with heterogeneous structural and functional phenotypes. These diseases are characterized by heterogeneous complex etiologies and a significant percentage of their number is of genetic origin, origin which has been increasingly uncovered over the past three decades. Cardiomyopathies are typically characterized by abnormal thickening, weakening or stiffening of the myocardial wall causing significant global systolic alteration in the absence of congenital heart disease, hypertension, coronary or valvular heart disease [8]. The classification of primary cardiomyopathies, *i.e.* cardiomyopathies not originating from a different pre- or co- existing condition, has been repeatedly updated over the past decades. The most common classification divides them into five categories: dilated, hypertrophic, restrictive and arrhythmogenic right ventricular cardiomyopathy, and left ventricular noncompaction [9].

Hypertrophic cardiomyopathy (HCM)

This thesis focuses on hypertrophic cardiomyopathy (HCM). This cardiomyopathy has a heterogeneous clinical spectrum and it is defined by unexplained myocardial hypertrophy in the absence of other cardiac or systemic diseases [10, 3]. The reported prevalence of HCM is 0.02 – 0.2% (1:5000 – 1:500) in the general population. Sudden cardiac death is the HCM most devastating expression, and represents the most frequent cause of sudden death in the young with an annualized rate presumed to be 1% per year [11, 10, 3].

Because of marked heterogeneity in clinical manifestations, phenotype and prognosis, HCM often represents a dilemma to cardiologists with regards to diagnostic criteria, clinical course, and management [10]. The clinical literature of the last 60 years has largely investigated HCM, and reported hypertrophy patterns which vary widely in extent and location, while the LV cavity remains non-dilated with normal or hyperdynamic global
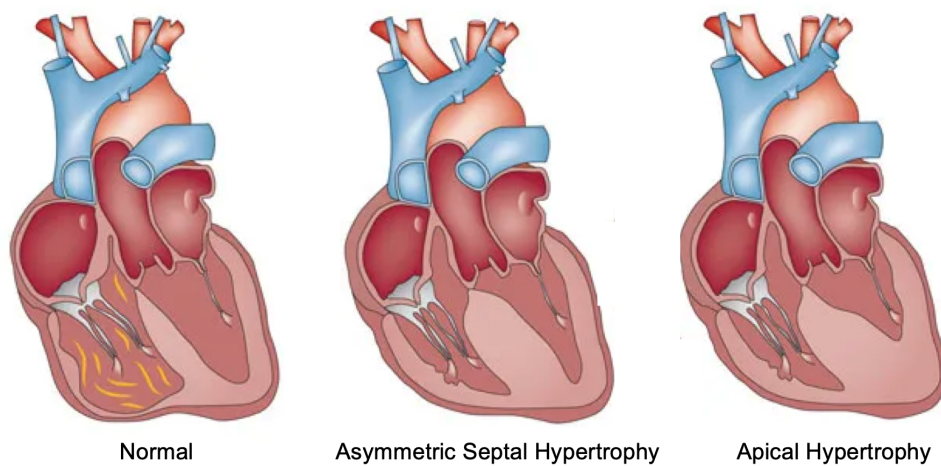
Figure 1.3: The two most common phenotypes in HCM: asymmetric septal hypertrophy and apical hypertrophy. Figure adapted from [12].

systolic function. Among the heterogeneous spectrum of hypertrophy patters that affect the LV, the two most common reported patterns affect the interventricular septum and the apex; these patterns are reported in Figure 1.3.

The genetic bases of HCM have been investigated for over 30 years [13, 14]. Pathogenic variants in genes encoding sarcomere proteins are identified approximately in 60-70% and in 10-50% of the HCM patients with or without a HCM family history [15]. In the American College of Cardiology Foundation and American Heart Association Task Force guidelines from 2011 [16], it is emphasized that the diagnosis of HCM should be confirmed in the presence of LV hypertrophy and an identified sarcomere mutation, or in the presence of LV hypertrophy $\geq 15mm$ without extra genetic findings. However, a substantial proportion of HCM patients lack familial history, and considerable uncertainty remains about the evidence of causality on genetic mutations found in HCM patients and on the interpretation of genetic testing results. This has resulted in clinical guidelines conservative in assigning pathogenicity to genetic variants. As an example, the guidelines of the European Society of Cardiology [17] from 2014 use the purely morphological disease definition of a LV hypertrophy $\geq 15mm$ that is not solely explained by abnormal loading conditions, making the diagnosis of HCM depending only on imaging criteria.

1.2 Cardiac Imaging

Cardiac imaging techniques are a set of non-invasive tests used to diagnose and monitor the structural and functional behaviour of the heart, with the aim of avoiding unnecessary invasive procedures. Common modalities for cardiac imaging include Computer Tomography (CT), Positron Emission Tomography (PET), Single-Photon Emission Computed Tomography (SPECT), Ultrasound (US) and Magnetic Resonance (MR) [3]. CT imaging produces tomographic images from X-ray transmission by rotating an X-ray source and a digital detector system around the subject under examination. The obtained acquisitions have high spatial and temporal resolution, but employ a significant amount of ionizing radiation and for this reason are mainly used for diagnosis and follow up of coronary artery disease. PET and SPECT are functional imaging techniques that measure positron and gamma emitting radioligands injected intravenously to study metabolic processes. SPECT is employed to study myocardial perfusion and coronary heart disease in conjunction with CT or MR. PET is more expensive than SPECT and for this reason, even if it provides better contrast and spatial resolution, it is mainly used for research purposes. Cardiac US or Echocardiogram uses ultrasonic waves to produce a real-time visualisation of the heart chambers and, by Doppler US, of the blood flow. The low costs and absence of ionising radiation associated with cardiac US makes it the most common and cheapest investigation for a first assessment of the heart status. Magnetic resonance (MR) imaging exploits magnetic fields and radio-frequency waves to acquire tomographic images with unmatched soft tissue contrast. As for US, MRI does not use ionising radiation, making it relatively safe and enabling repeatable acquisitions. In contrast with US, MRI has higher soft tissue contrast and larger field of view and smaller intra-observer and interobserver variability. These characteristics make MR an accurate, highly reproducible modality well-suited for populations studies [18]. On the downsides, MRI is an expensive modality and causes discomfort when compared with US, as patients are required to remain still inside a narrow tube for at least 30 minutes, while also wearing headphones for

reducing the loud noise caused by the gradient coils in the scanner. MR imaging is also impacted by artifacts due to heart and respiratory motion due to its long imaging times. Among the many different modalities of cardiac imaging, this thesis deals with data from MR imaging exclusively, and thus the following of this section will mainly focus on MR, which represents the current gold standard for the assessment of cardiac structure and function.

1.2.1   Cardiac Magnetic Resonance Imaging

Cardiac MR (CMR) allows for evaluation of infarct segments, characterisation of perfusion, motion and wall abnormalities, making it an excellent modality for myocardial assessment [19, 20, 3]. A challenging aspect of imaging the heart with CMR is presented by the complex and rapid heart motion in addition to two other types of motion: the respiratory motion and the patient motion due to relatively long image acquisition times. Heart motion can be addressed by means of fast acquisitions and of electrocardiography (ECG) gating, which enables the identification of the phase of the cardiac cycle to which each acquired image belongs. In healthy individuals, respiratory motion can be alleviated by means of breath holding acquisitions, single-shot fast imaging techniques or respiratory-triggered acquisitions by placing a respiratory tracking device on the subject's abdomen [21, 3].

The preferred MR sequence for myocardial assessment is cine balanced steady-state free precession (bSSFP) [22]. This acquisition provides good temporal resolution and high blood vs myocardium contrast, and very high signal-to-noise ratio. Cine bSSFP MR acquisitions are acquired in stacks of thick 2D slices from different imaging planes and for multiple cardiac phases so that to include the whole cardiac cycle [23, 24]. Ventricular volumetric measurements are routinely obtained from these SAX acquisitions [25]. A typical acquisition starts from a fast single-shot sequence which aims at identifying the long-axis (LAX) plane of the heart. From the LAX plane, two planes orthogonal to it can
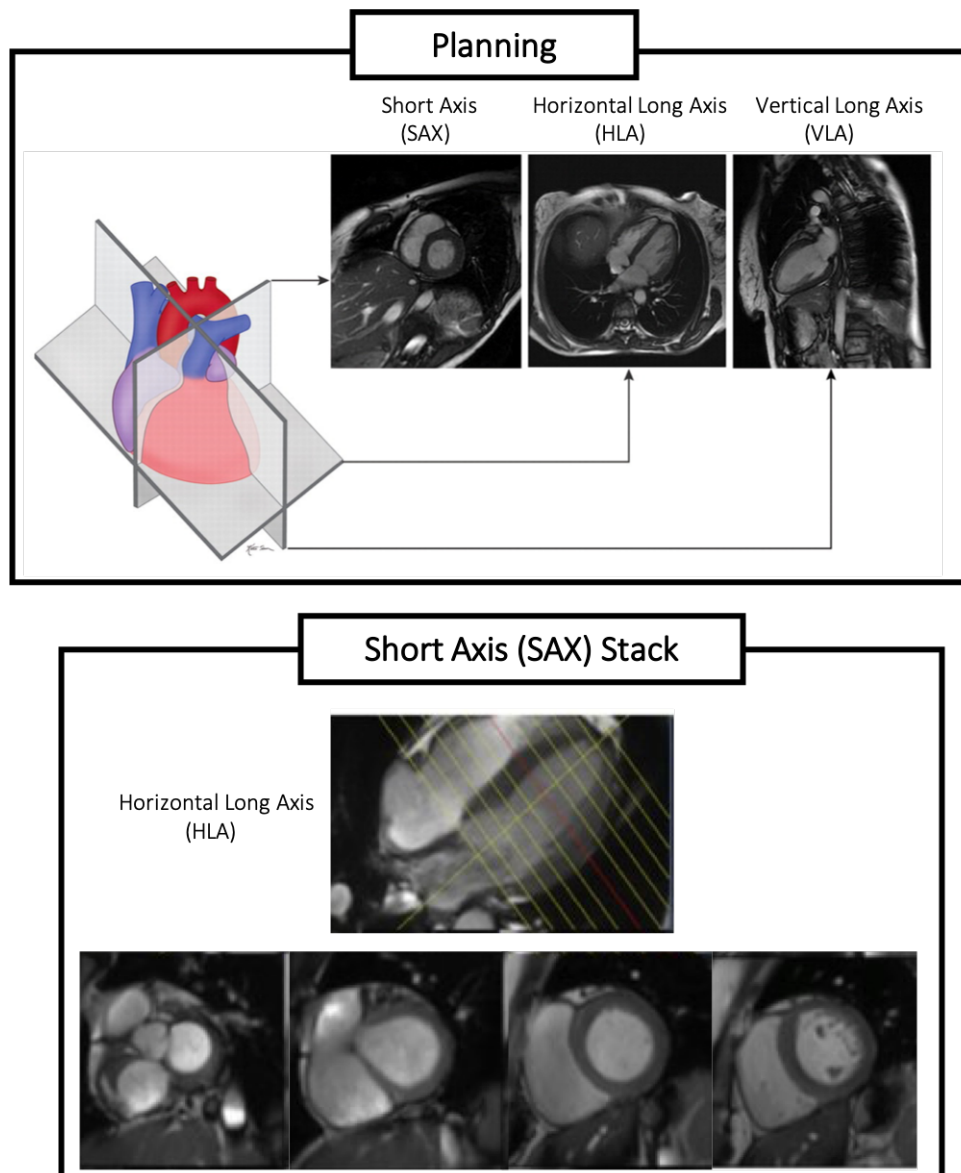
Figure 1.4: Top: An example of the planning used for the short-axis bSSFP cine stack. Bottom: Examples of short-axis slices. Figure adapted from [24] and [27].

be identified: the short-axis (SAX) plane and the vertical long-axis (VLA) plane. Then, a stack of 2D images is acquired along the SAX plane, creating a 3D representation of the LV. However, only one or two of these images are acquired during a breath-hold and, also because of cost factors, these acquisitions result in a in-plane resolution (1-2.5mm) much higher than the through plane resolution (8-10mm) [26]. The location of these planes with respect to the cardiac anatomy and an example of a 2D cine bSSFP cardiac MR acquisition are depicted in Fig. 1.4.

(a) 2D cine MR   (b) 3D cine MR
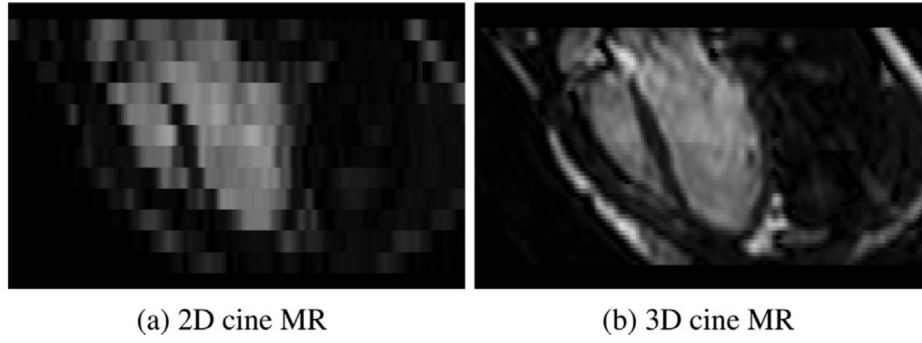
Figure 1.5: Resolution comparison of 2D vs 3D bSSFP cine sequences. Image courtesy of [31].

Misalignment between the acquired slices of the SAX stack is common in 2D cine acquisitions. This is caused by slightly different subjects' positions during breath-holding acquisitions or poor compliance with breath holding instructions, even in the healthy population [28, 29]. Another limitation of 2D cine acquisitions is the need to find a trade-off between number of required breath-holds and the slice gaps between each acquired image of the 2D stack, which can cause localised or regional effects to be missed. These challenges can be overcome by using 3D bSSFP sequences which enable isotropic high-resolution acquisitions and to image the full volume in a single breath hold [30]. However, 3D bSSFP sequences have worse contrast between blood and the myocardium and long breath holds, and therefore are not commonly used in clinical practice. An example of 2D vs 3D bSSFP cine sequences are reported in Fig. 1.5.

Several other CMR imaging acquisitions are also routinely employed in the clinical routine [2, 3]. Myocardial tagging has been introduced to evaluate myocardial wall motion at a regional level during the whole cardiac cycle [32, 2]. This is achieved by altering the magnetization of the myocardial tissue, such as in the spatial modulation of magnetization (SPAMM) technique that produces a series of images with hypointense stripes (or tags). Other popular CMR tagging sequences include delay alternating with nutations for tailored excitation (DANTE), which provides high-density pattern of thin stripes, and complementary SPAMM (CSPAMM), which was introduced to improve the SPAMM

contrast, providing better defined stripes [32]. Late gadolinium enhancement CMR (LGE-CMR) is instead used to identify areas of myocardial fibrosis thanks to a gadolinium-based contrast agent [33], while CMR imaging perfusion also employs a gadolinium-based contrast agent to derive blood flow (perfusion) maps of the heart, both under stress and at rest, for the assessment og known or suspected coronary artery disease [2, 3].

1.2.2 Indices of Cardiac Function

The quantitative assessment of cardiac structure and function is of key importance in managing patients with cardiovascular disease. The importance of accurate quantitative analysis has been increasingly recognized for accurate diagnosis, prognostication and therapeutic planning. Indices of cardiac function can be grouped into two classes: 1) global indices, which assess the overall ability of the heart to supply blood to the body; 2) local indices, which are used to assess regional contraction by quantifying myocardial deformation [2].

For the LV assessment, the heart chamber of interest in this work, global indices comprise:

- Left ventricular mass (LVM) which represents the mass of the LV myocardium. It is obtained by the product of the LV myocardial volume and an estimated myocardial density.

- Left ventricular volume (LVV) which measures the volume of the blood enclosed by the LV. The LVV curve during a cardiac circle is usually derived to assess LV perfomance.

- Stroke volume (SV). SV quantifies the volume of blood pumped by the LV within an heart beat, and it is thus defined by the difference between the LV end-diastolic volume (LVEDV) and the end-systolic volume (LVESV).

- Left ventricular ejection fraction (LVEF) which coincides with the ratio of stroke volume to LV end-diastolic volume (LVEDV). LVEF represents the volume of blood ejected by the LV during a contraction divided by the volume volume of the blood enclosed by the LV at end-diastole (ED).

- Cardiac output (CO) which is the amount of blood ejected from the LV per minute and is computed by multiplying the stroke volume by the heart rate (HR).

- Cardiac index (CI) which is instead computed by dividing the cardiac output to body surface area (BSA), hence accounting for the size of the individual.

A widely recognised regional analysis technique for the LV myocardium is to use the AHA 17-segment model in combination with local indices to measure regional LV abnormality. The most adopted local indices are regional wall thickening, regional wall motion and regional myocardial strain. Regional wall thickening (RWT) measures the change in myocardial wall thickness (WT) between two phases, usually end-systole (ED) and end-systole (ES), where WT is defined as the distance between the endocardial and epicardial surfaces. Regional wall motion (RWM) measures the wall motion, perpendicular to the endocardial wall, of a myocardial segment between two cardiac phases. Regional myocardial strain (RMSt) is the fractional change of the myocardial length between two cardiac phases in the radial, circumferential and longitudinal directions.

1.3   Population-based Cardiac Imaging Studies

The structure and function of the heart continuously adapts (remodels) either because of physiological factors (*e.g.* physical exercise), or as a consequence of disease. Pathological cardiac remodelling is a clinical term that refers to any change in the size and shape of the heart in response to cardiac disease [34]. Early quantitative assessment of cardiac remodeling through imaging-derived indices is of crucial importance for the detection and

effective treatment of almost all cardiovascular conditions. At present, despite substantial advances in their treatment, diagnosis of heart pathologies is still too often made at late symptomatic stages, leading to late, costly and partially effective therapies. As a consequence, characterizing the early stages of cardiovascular diseases has now become one the main focuses of contemporary cardiovascular research, with the final goal of developing accurate, effective preventive healthcare [35, 36].

Large part of our current understanding of human diseases relies on population-based studies. These studies aim at testing a clinical hypotheses over a defined population, *i.e.* a group of individuals who share a common characteristic, such as age range, ethnicity, or health condition. From an historical perspective, in the cardiac field the most relevant studies include the Framingham Heart Study [37], the INTERHEART study [38], and the MONICA project [39], from which the current knowledge of epidemiological risk factors of cardiovascular disease arised.

Due to the improvements of imaging modalities and reduced costs, non-invasive imaging has been increasingly incorporated into population-based studies in the last decades [40]. This integration aims at investigating the association between morphological and functional alterations of the human anatomy, as measured by imaging-derived indices, and clinical factors. This new scenario provides a unique opportunity for deepening our understanding of the mechanisms and characteristics of cardiovascular disease [41]. However, the unprecedented volume and speed of accumulation of such databases makes their analysis impractical without automated, data-driven aiding tools. In addition, deciding how to summarize the high-resolution phenotypical information provided by such images is of pivotal importance. Large part of the current literature summarises this information using global and regional indices of heart function, to then employ them in association studies or predictive clinical models [42, 43, 35, 44, 45, 46].

Estimation of the majority of the clinical indices used in population-based imaging studies requires delineation of the endocardial and epicardial boundaries of the myocardium, a

process called image segmentation. In order to do so, manual or semi-automated segmentation by experts has represented for many years standard clinical practice. However, this constitutes a tedious, time-consuming task, prone to intra- and inter- observer variability [47]. For this reason, the development of computational tools aimed at achieving fully automated image segmentation have been widely pursued. This resulted in the introduction of many fully automated image segmentation pipelines [31, 48, 49, 50], many of them now enabling the derivation of imaging-derived indices with drastically reduced physician supervision, mainly limited to a quality control step of the obtained segmentation before clinical report finalisation [51].

The introduction of fully-automated learning-based segmentation pipelines of anatomical structures has consequently paved the way for *large-scale* population-based imaging initiatives. In such studies, clinically relevant imaging indices are automatically derived for all the subjects' images in a cohort under study. Later, these indices are employed as variables in statistical models to investigate their association with epidemiological and clinical data, leaving to scientists only the tasks of formulating clinical hypotheses and of choosing the right computational and mathematical methods to test them. Examples of large-scale population-based initiatives carrying out cardiovascular imaging include Multi-Ethnic Study of Atherosclerosis (MESA) [52] - the first large-scale population study to employ cardiac MRI, UK Biobank [53], the Jackson Heart Study (JHS) [54], the Cardiac Atlas Project [55] and the UK Digital Heart Project [56].

Currently adopted classifications of cardiovascular diseases are grounded in the results provided by population-based cardiac imaging studies. In these studies, indices of mass and volume obtained from echocardiography and CMR were adopted to represent the cardiac phenotype, hence assuming anatomic uniformity of effect [57]. However, this has been proved to be sub-optimal for early discovery of alterations due to impending disease, patient stratification and management [58, 59, 60]. Relevant to this work, this also affected the characterisation of the diseases of the myocardium, often resulting in

crude diagnostic criteria [58, 61].

1.3.1   Cardiac Imaging-Genetics

The development of many cardiac conditions, including cardiomyopathies, depends on a complex interaction between environmental and genetic factors [62]. The past decades have witnessed a tremendous progress in the development of high-throughput DNA sequencing and genotyping technologies. These technologies now enable cost-effective and fast sequencing of large groups of individuals and made genetic testing increasingly employed in the diagnosis of inherited cardiac conditions. Moreover, these improvements together with the hypothesis that imaging phenotypes bear a close relationship to the underlying biology, has made the integration of genetics data within population-based imaging studies more common [63, 64, 65]. The derived class of studies, named imaging-genetics studies, aims at discovering novel genotype-phenotype associations between genetic variants and imaging phenotypes. The UK Digital Heart Project [56] and UK Biobank [53] are examples of such initiatives in the cardiac domain.

Genetics studies already helped shed new light on the genetic factors governing extreme ventricular phenotypes as imaged by non-invasive imaging, and proved how common genetic variants are implicated in a wide range of cardiovascular diseases [66, 67]. However, a common factor limiting current research is the ability to employ detailed, high-dimensional representation of human phenotypes in such studies and to successfully combine them with the already high-dimensional genotypic data [68, 69, 70]. So far, this has resulted in the adoption of one dimensional variables, such as the global or regional indices of cardiac function, as means to represent the heart phenotype [71, 72]. However, this crude approach lacks of the necessary resolution for the assessment of asymmetric, regional genotypic effects [73, 69], and often produced underpowered studies requiring large, prohibitive sample sizes to discover or replicate previously discovered phenotype-genotype associations [70, 63].

1.4   Medical Image Analysis with Machine Learning

The introduction of machine learning approaches in medical image analysis shows promise to fully automate many image processing tasks. Machine learning methods can effectively leverage large amounts of annotated imaging data to make accurate, automated predictions from them. However, it is only until the introduction of deep learning models, a subclass of machine learning methods which exploit neural networks, that large-scale fully automated medical image analysis was made possible [74, 75, 76]. The reason behind the success of deep learning models relies on their extraordinary ability at automatically learning complex task-specific models from large datasets using a general-purpose learning procedure [77]. These models can efficiently exploit complex image features learned through stacked layers of non-linear processing constituting a neural network, bypassing feature hand-crafting steps usually present in traditional machine learning pipeline designs. This new learning paradigm, named end-to-end learning, has been made possible only in the last decade by the technological progress in computer hardware (e.g. graphical processing units (GPUs) and tensor processing units (TPUs)), big data infrastructures to collect vast amounts of imaging data and efficient stochastic optimization methods [77]. Due to these improvements, deep learning models managed to outperform previous traditional machine learning methods in a variety of medical imaging tasks, which often relied on hand-crafted features and had significantly slower inference time [75, 76, 78]. In the context of medical image segmentation, these algorithms can learn the necessary image features to reproduce the delineation of the anatomical contours of interest on unseen images with high accuracy, either in healthy and pathological populations [31, 48, 49, 50]. Prior to them, although some relatively efficacious methods were already commercially available, lengthy and variable manual post-processing was required [51, 75, 79]. Similar improvements have been obtained in automatic organ or landmarks localization, often an important pre-processing step for image segmentation, medical image synthesis, reconstruction and enhancement, tumor or lesion detection and classification and quality

control of medical images [74, 80, 79].

1.5 Modeling Cardiac Shape in Large-Scale Imaging Studies

The successful introduction of machine learning methods into medical image processing pipelines, together with improvements and reduced costs of non-invasive imaging and data management infrastructures, has paved the way for large-scale population-based imaging studies. Due to these improvements, large databases of cardiac images can be fully- or semi- automatically annotated with minimal human expert effort and costs, showing promise to deepen our understanding of cardiovascular disease. Traditional approaches in clinical research employ linear regression modelling of clinical variables and crude volumetric imaging-derived metrics. However, these metrics mostly ignore shape information available in modern cardiac imaging examinations and are largely insensitive to early pathological cardiac remodeling [81, 82, 83]. In order to address this limitation, computational and statistical modelling of cardiac shape, motion and physiology which leverage large annotated imaging datasets have received substantial interests [84, 55]. The aim of this class of approaches is to discover novel imaging biomarkers, which afford precise and regional quantification of disease-relevant cardiac variability, and to employ them in association and predictive models to uncover their relationship with common environmental and genetic risk factors [59, 85].

Three-dimensional shape data obtained medical image segmentation algorithms on large datasets is typically high-dimensional, presenting its own set of challenges for drawing inferences. This problem is further exacerbated when integration with large-scale genetic and epidemiological data is aimed or a compact set of parameters is needed to inform clinical prediction models [63, 59]. To date, two approaches attempting at going beyond crude metrics of shape variation while addressing the high-dimensional nature of shape

data have been proposed [85]. The first approach transforms shape data into a lower-dimensional, latent representation that summarises the shape variability by means of an average shape and its main modes of deformation. Principal component analysis (PCA) or its modifications have mainly been employed for this task. This aims at obtaining a low-dimensional latent representation, considered as a new set of shape descriptors, to be used in association studies or in classification and clinical prediction models [86, 85]. However, the shape variation encoded by these new parameters does not guarantee their usefulness in subsequent analysis as the captured variation may have small correlation with the clinical hypothesis under exam, largely impeding the full exploitation of the acquired data [87, 85]. The second class of approaches consists in maintaining the high-resolution, high-dimensional phenotyping provided by state-of-the-art imaging techniques to derive 3D structural cardiac phenotypes at each and every vertex of a computational model of an anatomy of interest. Then, the association between variations of these 3D phenotypes and clinical variables of interest is investigated by computing univariate statistics at each and every shape vertex, independently. This approach is named mass univariate analysis [88], and it originates from neuroimaging research, and had been loosely investigated in the cardiac domain [89, 56].

Machine learning approaches showed great promise in addressing big data challenges in cardiovascular research [51, 75], including the analysis of anatomical shape variability in large populations. One of the strengths of these approaches relies in their ability to discover unknown patterns in the data without requiring any a priori assumption, potentially enabling the discovery of unknown disease-relevant shape biomarkers of cardiovascular disease, which could more accurately predict clinical outcomes and better correlate with genetic and environmental covariates. On the other hand, machine learning approaches, and especially deep learning approaches, often lack interpretability in the feature extraction and decision process, and seldom result in new metrics and biomarkers that can be readily understood by a clinician. This poses the need for the development of approaches and software that can fully exploit such classes of model while, at the same time, being

interpretable by an human [76].

1.6 Thesis Objectives and Contributions

The ultimate objective of population-studies of the heart is to extract novel, useful clinical insights from large-scale imaging databases. The research presented in this thesis contributes to improve such studies by modeling heart shape variation using machine learning and advanced statistical analysis. The work in this thesis aims at providing automated, data-driven computational methods capable of better summarizing shape variability than conventional volumetric imaging metrics or traditional dimensionality reduction techniques, and which can be used to inform association, classification and predictive clinical models. At the same time, this thesis aims at developing explainable models, enabling the visualisation of the captured anatomical variation in three-dimension. The main contributions of this work can be summarized as follows:

A framework for 3D imaging-genetics. A mass univariate framework able to derive associations between genetic variations and 3D high-dimensional cardiac phenotypes obtained from MR imaging is developed. Extensive experiments on genetic and synthetic data showing how the proposed framework enables to derive computationally-efficient are reported, statistically powerful inferences on imaging-genetics datasets, even in the presence of small genetic effect sizes not detectable with conventional imaging parameters. Moreover, a general purpose R package to perform this class of studies is also introduced.

Learning generative *and* discriminative deep features of cardiac shapes. Two 3D deep learning models for automatic classification of cardiac shapes from patients with cardiac diseases associated with structural remodeling are proposed. By exploiting state-of-the-art deep generative models, the proposed approaches are specifically designed to

learn both generative and discriminative anatomical shape cardiac features. More importantly, it enables the visualisation of the anatomical effect in three-dimension. This provides an interpretable classifier of anatomical shapes which achieves high accuracy in the categorisation of healthy and remodelled left ventricles when tested on unseen data.

Modeling shape variation in large-scale imaging databases using deep hierarchical generative models. A deep learning approach, based on deep hierarchical generative models, to model a large database of anatomical shapes through a hierarchy of conditional latent variables is introduced. Due to the generative properties of the model, the encoded three-dimensional anatomical effect can be visualised on a template anatomical shape. Furthermore, the model enables automated, data-driven extraction and visualisation of the shape differences distinguishing different populations.

Learning to generate 3D high-resolution cardiac shape from 2D views. Finally, this work introduces a deep conditional generative model architecture able to reconstruct 3D high-resolution LV segmentations from a sparse set 2D LV segmentations of routinely acquired MR images. This is evaluated on unseen healthy volunteers, outperforming competing architectures and showing promise for future investigations.

1.7   Thesis Overview

This PhD thesis is structured as follows. Following this chapter, the state-of-the-art computational anatomy and machine learning methods of interest for this work are reviewed (Chapter 2). Chapter 3 introduces the mass univariate framework for 3D imaging-genetics and reports extensive experiments on real genetic and synthetic data, showing its effectiveness in capturing regional genotype-phenotype interactions. In Chapter 4, a 3D deep learning classification model able to learn discriminative and generative features of diseases

associated with structural remodeling is presented. This approach yields high accuracy in the categorisation of healthy and remodelled left ventricles and enables the visualisation in 3D of the anatomical variability encoded by the adopted task-specific features. Chapter 5 presents instead a deep learning system able to model an imaging-derived database of anatomical shapes via a hierarchy of conditional latent variables and capable of detecting and visualising on a template shape the most discriminative anatomical features that characterize distinct clinical conditions under exam. In Chapter 6, a preliminary study showing the feasibility of the reconstruction of 3D high-resolution LV segmentations from a sparse set of 2D views LV segmentations using generative models is reported. Finally, in Chapter 7, the work presented in this thesis, its limitations and future work are summarised.

Chapter 2

Background

This chapter is composed of three sections providing all the methodological background for this PhD project. In Section 2.1 the field of Computational Anatomy and its main achievements in the cardiac domain are reviewed. This comprises an introduction to cardiac atlases and their construction process (2.1.1), to statistical shape models (2.1.2) together with a survey of their applications for cardiac shape analysis (2.1.3), and to statistical deformation models (2.1.4). In Section 2.2, a survey of the imaging-genetics methods is presented (2.2.1)followed by a detailed introduction to mass univariate analysis (2.2.3). Section 2.3 provides an introduction to deep learning (2.3) and its application to shape analysis (2.3.2), with a special focus on the autoencoder (2.3.3) and variational autoencoder (2.3.4)) architectures, which are of particular interest for this work.

2.1 Computational Anatomy

Shape, as defined by Kendall [90], *"is all the geometrical information that remains when location, scale and rotational effects are filtered out from an object"*. Notably, whilst in mathematical and computer vision applications changes induced by scale transformation are removed in the analysis of shape, in medical application changes in size are considered

as part of the biological shape variation to investigate. Computational anatomy is an interdisciplinary research field that aims at modelling the normal and pathological shape variability of the human anatomy [91]. Identifying statistical shape differences that characterize healthy individuals and patients affected by a disease is one of the main interests in medical image analysis, showing promise to improve diagnosis, disease staging and treatment.

2.1.1 Computational Atlases

Shape and size of anatomical organs can show significant variation in a population of medical images. Computational atlases have been introduced to provide a common reference system to rigorously compare and analyse subjects. This provides many advantages, including accurately quantifying and modeling the organs shape and size differences among group of subjects, studying their changes over time, and encoding structural, functional, mechanical and electrophysiological information, often from different imaging modalities, into a single framework [85, 92, 93]. Computational atlases are defined by the combination of a template anatomy and a deformation (registration) technique to map the information of other subjects to this template [91, 94, 95, 84, 96]. The generic approach to construct a template anatomy starts from a first averaging of all the anatomies under exam after spatial normalisation. The purpose of this step is to filter out location and rotational effects and is typically achieved by registering all the individual anatomies to an initial template. Such initial template can be obtained from another dataset, or be an anatomy selected from the available dataset. Then, an iterative procedure alternating registration steps to align the imaged anatomical structures to the template, and template update via averaging of the registered structures, is performed. The process normally stops when the difference between two successive templates is smaller than a threshold [97, 98, 99]. The two most common types of computational atlases are volumetric and surface atlases, which differ with respect to the computational means employed
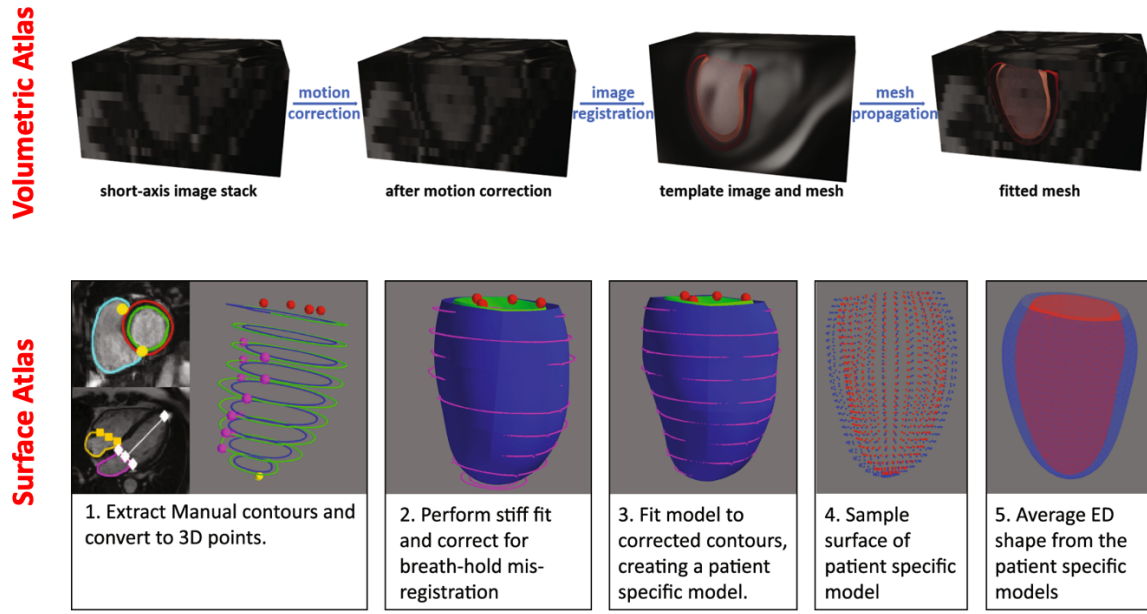
Figure 2.1: Volumetric and surface atlases construction schemes. Image adapted from [81].

to represent the anatomy [81, 59]. Figure 2.1 reports a scheme summarising the pipeline for their construction.

The template of a volumetric atlas consists of an image and its associated segmentation, hence encoding both intensity and geometric information. In order to derive such atlas, inter-subject spatial normalisation is achieved through volumetric image registration, which exploits image intensity features to compute the desired transformations. After correction for inter-slice motion, each image is registered via a global rigid transformation to the common reference system, and can be refined by a coarse non-rigid B-spline transformation [96]. The computed transformation is also used to warp the respective segmentation. This procedure enables to derive an appearance and a shape atlas. The simplest example of an appearance atlas template is the mean intensity image of the dataset. A random image or the image closest to the population mean intensity image have also been investigated as appearance templates [100]. However, these templates may not be truly representative of the population, especially in the presence of large inter-subject variation. For this reason, more complex models involving the estimation of the

(a) Template at ED.   (b) Label map at ED.   (c) Probability map at ED.   (d) Surface mesh at ED.
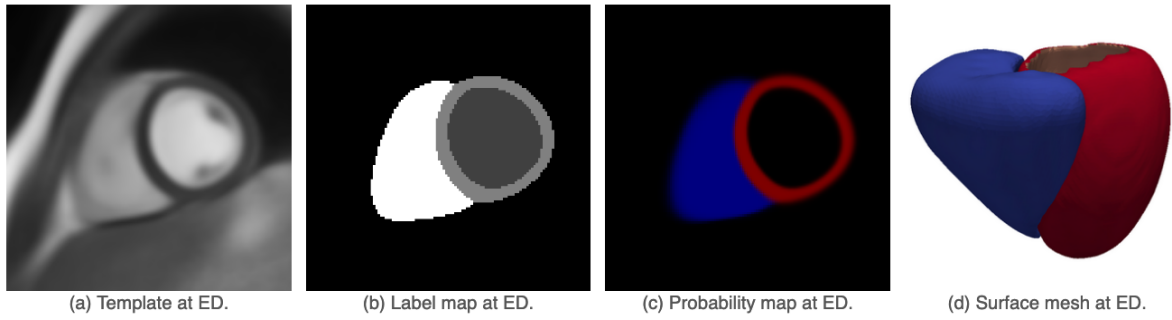
Figure 2.2: Components of the volumetric atlas of the UK Digital Heart project. Image courtesy of [96]. The first elements show the 2D mid-ventricular template image, its label and probability maps at end-diastole (ED), while at the far right the corresponding 3D surface mesh is also reported.

true population mean intensity have been proposed [98, 95, 101]. The shape atlas can be similarly obtained by fusing the subjects' segmentations after their registration to the common reference space. Alternatively, individual segmentations can be used to model the probability that a particular spatial location has a certain label [102, 96], generating a so-called probabilistic template. Probabilistic templates providing a tissue probability map or other local image statistics have been also later proposed to account for different resolutions and contrasts, registration distortions and the uncertainty related to underlying tissue type [103, 104]. Figure 2.2 showcases the described components for the UK Digital Heart project volumetric atlas [96]. In the scenario in which the initial template is a higher-resolution atlas, the inverse of the computed transformation can be applied to the higher-resolution shape atlas to derive a higher-resolution shape representation for each subject under investigation. Being a result of the warping of the same template to the same anatomical positions in each subject's image, the obtained higher-resolution shape representations will be consisting of co-registered shape vertices.

The template of a surface atlas is a surface of one or more anatomies of interest, describing their average shape in a population [84, 105, 106, 107, 108]. In the first step of its construction as illustrated in the second row of Figure 2.1, the segmentation or the landmarks associated with each individual image in the database under exam are used to construct a subject's specific surface mesh. This is obtained by fitting a 3D finite

element model to the obtained segmentation or landmarks, which also enables correction for inter-slice motion by shifting the surface contours in-plane. Moreover, it guarantees shape co-registration [84]. The spatial normalisation in this case is obtained by rigid registration of the meshes directly (by matching mesh vertices or landmarks), while the template shape is obtained by averaging the mesh coordinates.

Once a shape representation for each subject has been obtained, several phenotypes can be derived from the each individual shape [83, 96]. These include LV mass (LVM) by computing the volume of the voxels labelled as left ventricle myocardium in the corresponding segmentation, and LV end-diastolic volume (LVEDV) and the end-systolic volume (LVESV) by computing the volume of the voxels labelled as left ventricle cavity at ED and ES. Moreover, vertex-wise phenotypes can be derived at each co-registered vertices. These can include wall thickness (WT), measured as the length of segment perpendicular to a midwall plane equidistant to the endocardial and epicardial surfaces, fractional wall thickening (FWT), computed as the difference in WT at ED and ES divided by WT at ED, and many other 3D cardiac phenotypes including vertex-wise wall stress [83, 109].

2.1.2   Statistical Shape Models

Anatomical shape information provided by computational atlases is the building block of statistical shape models (SSMs) [110]. SSMs are a class of statistical methods that aim to study the distribution of global and regional shape variations across a population without making any *a priori* assumption on it [111, 112]. This is typically achieved by capturing a small set of parameters, named modes of variation, that encode the main variation around a mean shape representative of the population under study. These modes can be then used for subsequent association or predictive analysis as well as prior information for a wide range of applications [85].

SSMs were firstly introduced by the work of *Cootes et al.*, who proposed the currently most

common approach, called Active Shape Model (ASM) [113, 114]. In this work, landmark points, a set of points distributed across the surface of each anatomy, were used as shape representation. Each shape within the dataset was parameterised with the same number of co-registered points, *i.e.* each point was placed in the same anatomical location in all the shapes under exam. Then, the different anatomies were normalised with respect to rotation and translation, as in the computational atlas construction process, and gathered into a common space. Once aligned, 3D shapes are still typically high dimensional and thus require dimensionality reduction techniques to study their variabilty. In the ASM work, principal component analysis (PCA, [115]) was used for this purpose and it is still, by far, the most common method deployed to extract a mean shape and its modes of variation from a population of anatomies under study.

PCA is a popular dimensionality reduction technique which linearly transforms high-dimensional input data into a new coordinate system in which a new set of uncorrelated, mutually orthogonal basis vectors is employed to describe the input data variability [115]. Moreover, these new variables, named principal components, are designed so that each of them encodes the highest variance possible. In this way, the input data can be approximated using a linear combination of a mean value and a subset of principal components representing large part of the input data variance. In the study of shape variation, this provides an efficient and compact parameterisation of anatomical variability [116], which can also be used to inform common medical image processing [112]. As an example, in Figure 2.3, the first three PCA modes of shape variation for the healthy LV in UK Digital Heart project are showcased.

2.1.3   Cardiac Shape Analysis

The decomposition of a cardiac structure, which has been imaged in a population of interest, into a mean shape and its principal modes of deformation opens the way for two distinct types of subsequent analysis in the cardiac domain [117, 118]. The first
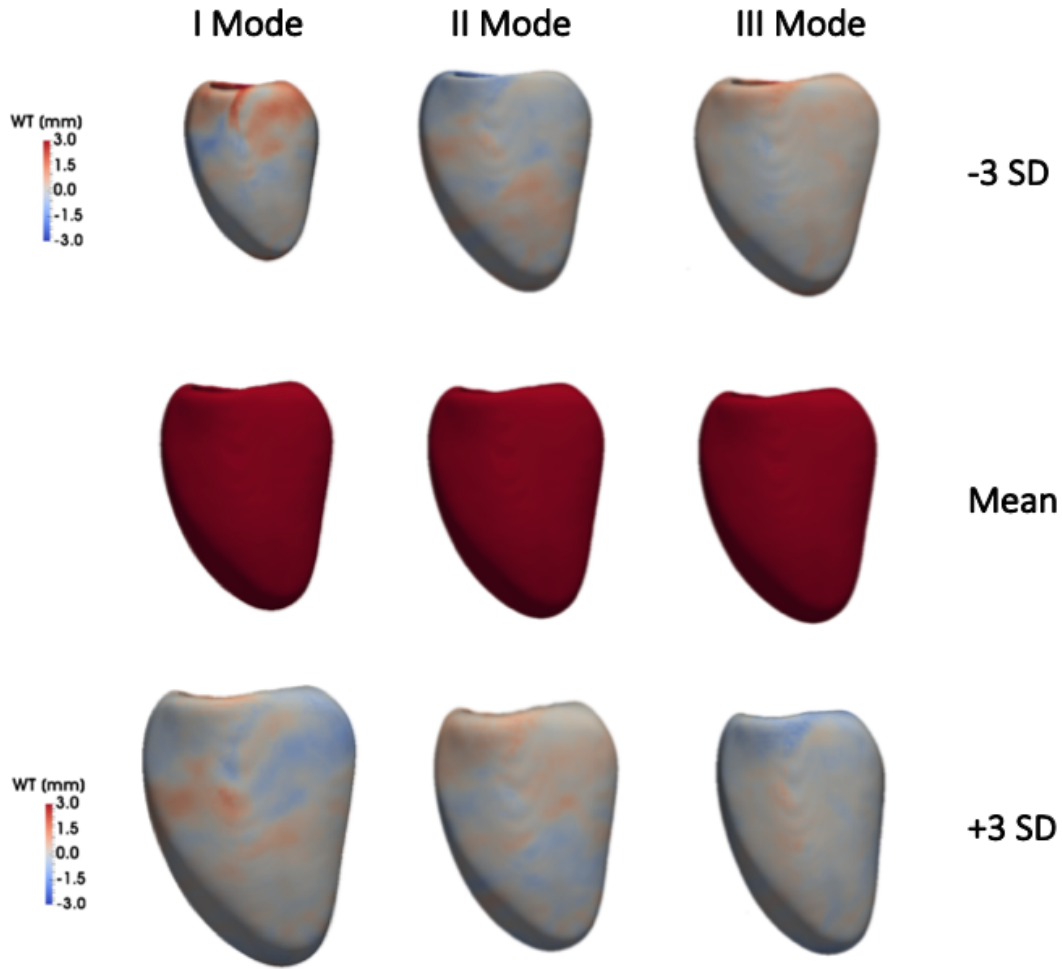
Figure 2.3: The first three modes of shape variation for the healthy LV in UK Digital Heart project, septal wall is showcased. Vertex-wise difference in wall thickness (WT) between each derived shape and the mean surface shape is plotted at each vertex. It can be noticed that mainly changes in size and sphericity are captured by PCA, while patterns in WT are neglected.

type, named descriptive analysis, aims at identifying the most common patterns and characteristics of a population of interest and to compare them against the ones of other populations. The second type, named predictive analysis, aims at deriving the relationships between shape and clinical variables by applying regression or classification models. The underling hypothesis of both these two classes of approaches is that automated shape characterization methods perform better than crude, one-dimensional mass and volume indices currently used in the clinic.
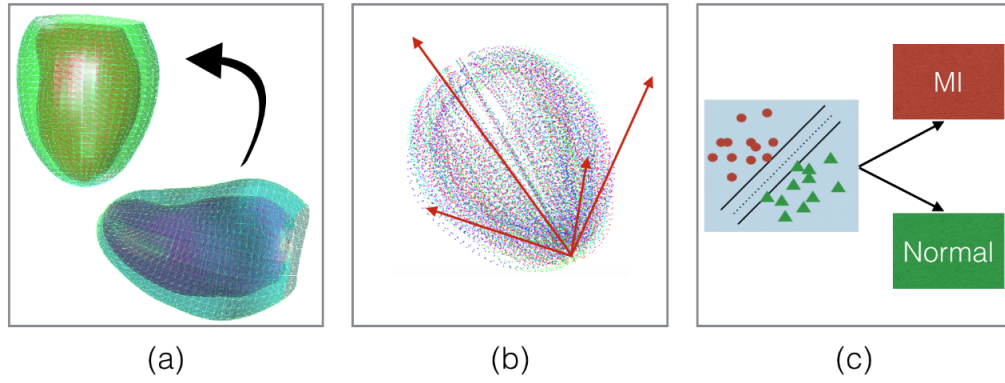
Figure 2.4: Example of a predictive shape analysis approach - Figure from [119]. (a) Anatomical shapes are aligned to a common reference system. (b) PCA shape components are extracted for each shape, modeling shape variation in a new reference system. In a second step, shape components are then used as features in a subsequent clinical prediction task (c). In this application, the model aims at learning the best decision boundary in the shapes feature space for the classification of healthy subjects and myocardial infarction (MI) patients.

The first descriptive analysis study of cardiac data aimed at describing the variability of 2D left ventricle contours in echocardiograms [120]. Later, statistical significant differences between the modes of deformation of healthy controls and diabetic subjects [121], women with higher cardiovascular risk profiles [122], pre-term born subjects in adult life [105] or congenital heart disease patients [123] were also investigated in subsequent shape analysis studies. In the same fashion, Ardekani *et al.* [124] showed how PCA-derived modes of deformation can be used to quantify LV shape differences between ischemic cardiomyopathy and global nonischemic cardiomyopathy patients. More recently, Bai *et al.* applied PCA analysis on a database of 1000+ LV shapes from the UK Digital Heart project to visualise their main modes of deformation [96]. Gilbert *et al.* demonstrated how volumetric and surface atlases provide similar LV principal component on 5000 LV shapes from the UK BioBank and tested their association with common cardiovascular risk factors [81].

An early work in predictive analysis is the work of Perperidis *et al.* [125], in which normal patients are differentiated from hypertrophic cardiomyopathy patients by means of LV PCA-derived modes of variation using a k-weighted nearest-neighbour classifier. Varela

et al. derived left atria shapes of 144 subjects from MRI images and then applied PCA to extract features of left atria shape to predict post-ablation atrial fibrillation recurrence [86]. In the 2015 challenge at the Statistical Atlases and Computational Models of the Heart MICCAI workshop [119], 11 methods for feature extraction and of classification of LV after myocardial infarction and healthy controls were proposed, including methods exploiting PCA-derived modes of deformation [106]. An illustration of one of these approaches is reported in Figure 2.4. However, a strong limitation of deploying PCA shape components as shape features is that they do not necessarily encode the shape information needed to differentiate between disease classes. For this purpose, approaches that search for new axes of variation that are disease-meaningful have been later proposed. Lekadir *et al.* [126] and Zhang *et al.* [127] investigated the of use partial least squares to decompose LV shape variability into modes of deformation that are optimally, independently associated with distinct remodeling indices. Dawes *et al.* proposed to employ supervised PCA to predict survival from 3D right ventricular systolic motion in a cohort of pulmonary hypertension patients [128]. However, such methods are still confined to modelling shape variations only in a linear fashion, are mainly able to capture global anatomical changes and rely on already known clinical indices, which could hide the discovery of unknown factors of variability.

2.1.4   Statistical Deformation Models

Closely related to SSMs are statistical deformation models (SDMs) [110]. The key idea behind this model is to perform statistical analysis directly on the dense deformation fields which describe the correspondence between an image template and subjects' images [110, 129]. SDMs first applications dates back to the introduction of deformation-based morphometry [130] in which highly non-linear registration algorithms were used to register subjects' shapes to a surface template. Then, this class of approaches employs statistical analysis on the parameters describing the computed deformation fields, aiming

at localising global shape differences between the different populations under study [131]. Similar approaches employing large deformation diffeomorphic metric mapping (LDDMM) algorithms to derive high-dimensional velocity fields, subsequently reduced in dimensionality, were also proposed [132, 82, 133]. Examples in the cardiac domain include the work of Mansi *et al.* in which a right ventricular shape representation for patients with repaired tetralogy of Fallot is derived [82], and the work of Jia *et al.* in which left atrial shape features related to post-ablation outcome in atrial fibrillation are computed [133]. More recently, Zhang *et al.* develops a novel probabilistic model for factor analysis in the space of diffeomorphisms which exploits principal geodesic analysis to extract a compact representation of image deformation [134]. Advantages of SDMs over SSMs include the possibility of both describing the intra- and inter-structure variability across a population to perform shape analysis without the need for medical image segmentation [135], which has been a challenging task until the development of machine learning approaches for medical image segmentation. Drawbacks of SDMs include their high computational cost and mathematical complexity together with their slow optimisation process, requiring many iterations to reach satisfactory convergence [136, 137, 131].

2.2 Imaging genetics

Imaging genetics is a research field that aims to evaluate the effect of genetic variation on functional, physiological, and anatomical phenotypes imaged by medical imaging techniques [63]. The genetic variation is often expressed in terms of single-nucleotide polymorphisms (SNPs) or candidate genes. SNPs are sites in the DNA sequence at which more than one nucleotide (adenine, thymine, cytosine or guanine) is found with a minimum frequency of 0.5% in a population. They are the most common type of genetic variation among humans and while most of them have no or minimal effect on health or development, it is believed that the coordinated effect of some of them regulates disease susceptibility and therapy response, making their discovery crucial for personalised

	Candidate Phenotype	Whole-brain
Candidate SNP/Gene		
Genome-wide		
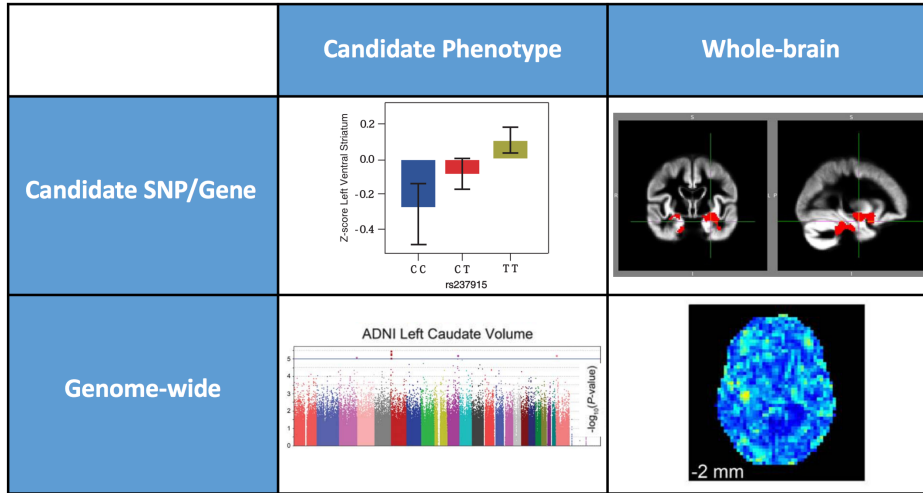
Figure 2.5: The four classes of imaging genetics methodologies. Columns: dependent/outcome variable, either a candidate phenotype (such as an imaging metric) or the whole structure. Rows: independent variable, either a candidate SNP/gene (such as an imaging metric) or the whole structure. Image adapted from [141, 142, 143, 144]

medicine [63, 138, 69, 139]. Historically, imaging genetics studies have been introduced and mainly applied in the neuroimaging domain, and only recently research efforts have been conducted in the cardiovascular field. From a methodological point of view [140], imaging genetic studies can be divided into four categories as summarized in Fig. 2.5.

2.2.1 Methodological Approaches

Candidate-phenotype candidate-gene analyses. These studies investigate the degree of association between a quantitative imaging-derived metric (also named candidate phenotype) and SNPs or pre-specified genes. This class of approaches provided the first evidence of the feasibility of imaging genetics research [145]. Despite their simplicity, they were able to demonstrate how quantitative biomarkers extracted from imaging modalities, already used as indicators of biological processes, disease status or treatment response, are also heritable and under genetic control [146], hence enabling inferences between brain status and genetics [147]. As an example, Loth *et al.* [141] tested the association between SNPs of the oxytocin receptor gene (OXTR, candidate SNP) and the functional MRI

bold activity of the ventral striatum of adolescent human brain (candidate phenotype). This latter was already recognized as a biomarker of social-affective problems. The study found that adolescents with homozygotic minor alleles had significantly lower activity (Figure 2.5, top left), therefore suggesting that OXTR-regulated brain mechanisms are associated with social-affective problems [141].

Candidate-phenotype genome-wide analyses. This approach again employs a quantitative imaging measurement as outcome variable, however in this case the association of SNPs of the whole-genome is tested against it, *i.e.* making no prior genetic assumption on the set of genetic mutations to test as in the case of the previous class of studies. In other words, a genome-wide association study (GWAS) using an imaging metric as outcome variable is performed (further information about these methods is reported in Section 2.2.2). The bottom left image of Figure 2.5 shows an example of outcome of one these studies in which several genetic variants of dopamine-related genes where found associated with the volume of the caudate subcortical brain structure assessed through structural MR imaging (candidate phenotype) [142].

Whole-image candidate-gene analyses. In order to precisely map genetic influences over the entire medical image, these studies compute a pixel- or voxel- wise map expressing the association between one or more genetic variants and the value at each pixel or voxel over the image. This type of investigation is usually referred to as mass univariate analysis [88] and was introduced in order to overcome the lack of phenotyping power of quantitative imaging-derived biomarkers, which are often insensitive to regional or asymmetric effects [73, 69, 148], leading to reduced statistical power and missing associations. As an example, Filippini *et al.* reported a positive association between the APOE $\epsilon 4$ allele (APOE4) - a genetic risk factor for sporadic Alzheimer's disease - and regional brain cortical atrophy as assessed by T1-weighted MR images [143]. The results are reported in a Manhattan plot, a plot reporting SNPs along the X-axis and the negative logarithm

of their association p-value on the Y-axis, at the top right of Figure 2.5.

Whole-image genome-wide analyses.   This approach aims to scan the entire genome and entire brain at the same time for non-random associations, and it is the most challenging one from a statistical perspective. It is motivated by the fact that, despite the interpretability and computational tractability of the above studies involving candidate phenotypes and candidate genes/SNPs, a priori knowledge on how to select these variables is not always available. Discarding part of the biological information contained in the genome or in an image could cause to miss potential associations, especially when the underlying mechanisms and the genetic factors of the disease process are largely unknown. However, merging the extremely high-dimensional genetic and imaging data imposes many challenges, making it still an open research topic. The first attempt in this direction was made by Stein *et al.* [144] who studied, using parallel computing, the univariate relation between 448,293 SNPs and 31,622 voxels in whole-brain maps of local brain volumetric change in elderly subjects with Alzheimer's disease. In the study, several biologically interesting SNPs were identified and the association maps of the one scoring the lowest p-value was plotted on a template image (Fig.   2.5, bottom right). However, no SNP had whole-brain whole-genome significance due to the massive multiple testing correction required by the proposed approach [144]. This could be due to the fact that univariate-imaging univariate-genetic association tests completely ignore information from neighboring voxels or nearby genetic variants, which still makes the statistical power of these techniques low. To overcome this, joint modeling of the imaging and genetic data looks a promising area of research. Recent works employed multivariate models regularized with sparsity or rank constraints to select subsets of putative genetic markers and affected voxels, reducing the burden of multiple testing correction problem [149, 150, 151]. However, this class of techniques largely employs computationally expensive iterative optimization procedures and permutation schemes, are prone to overfitting problems, and have higher mathematical complexity than the other three categories of studies, which

makes the evaluation of the significance of the obtained findings difficult.

Additional details about GWAS and mass univariate analysis will be provided in the next two sections (2.2.2 and 2.2.3) due to their relevance to this work.

2.2.2 Genome-wide Association Studies

A genome-wide association study (GWAS) scans hundreds of thousands to millions of SNPs of an individual to test their statistical association with categorical outcomes (such as case versus control) or continuous outcomes (such as an imaging metric for candidate-phenotype genome-wide analyses) [138]. These studies are based on the concept of linkage disequilibrium, which expresses the degree of correlation between pairs of SNPs in a continuous genomic region. Due to linkage disequilibrium, a large group of SNPs can be represented through only one marker - a tag SNP - in GWAS, due to their high correlation. This reduces the number of SNPs to be studied, making these analyses feasible. GWAS are supported by the "common disease - common variant" hypothesis stating that complex diseases are caused by several common alleles (loci) that taken individually have a negligible effect on the phenotype, unlike Mendelian diseases which are caused by a mutation of a single gene and are studied in candidate genes studies ("rare disease - rare variant" hypotheses). As mentioned above, the main advantage of these studies is that they are unbiased by prior knowledge on the relationship between certain genes and complex phenotypes (as opposed to candidate genes studies). Their main limitation is the enormous number of statistical tests being performed that could potentially cause a large number of false positive results.

In 2006, the first GWAS were extremely successful and led to important results as the discovery of coding variants involved in age-related macular degeneration [152, 153] and of the IL23R Crohn's disease gene [154]. However, in the following years several GWAS studies suffered from the lack of a standardised protocol, leading to many associations

that were not validated in external cohorts and no more reached the modern definition of genome-wide significance ($p < 5 \times 10^{-8}$), such as in the case of the association between the gene INSIG2 and childhood obesity [155]. In the cardiovascular field, the normal distribution of ventricular mass, ejection fraction, volume assessments in populations suggests that this phenotype is a complex trait influenced by multiple genes. Several loci have been discovered to be associated with LVM [156, 157, 158] and other structural or functional ventricular phenotypes [71, 159]. However, a very small number of the discovered loci has been replicated in subsequent studies [160]. It is believed that these discoveries have only accounted for a small proportion of phenotypic variance and that a strong contribution to the missing heritability might be due to the low-resolution adopted for heart phenotyping, *i.e* summarizing the complex heart phenotype with a handful set of global metrics [69, 139].

2.2.3   Mass Univariate Analysis

A commonly used tool for whole-image candidate-gene studies is mass univariate analysis [88, 161]. Mass univariate analysis constitutes the building block of voxel-based morphometry (VBM), arguably the most successful method in the Computational Anatomy field, which was popularised by the methodological work of Ashburner and Friston and the statistical parametric mapping software package [162, 163]. A mass univariate study analyses data involving a massive number of univariate hypothesis tests (e.g., t-tests) at every pixel or voxel of an image, and produces a statistic image, or image of statistics, named statistical parametric map. This resulting statistic image is then assessed for statistical significance and the regions in which statistical significance is found are highlighted on the original image [164]. The voxel- or pixel-wise test can express the local difference in the tissue maps between two studied groups as in VBM, or the association between a candidate gene and the intensity values at that spatial location as in whole-image candidate-gene studies. Both approaches start from a computational atlas which is

obtained through tissue segmentation, spatial normalisation, Jacobian modulation, and spatial smoothing of the images of a population [162]. Tissue segmentation is usually needed as the calculations are performed at each voxel within the segmentation mask so as to reduce the multiple comparison problem always generating from this approach.

Multiple Comparison Problem

How to mitigate the problem of multiple comparisons that arises in mass univariate studies has received a lot of research interest [165, 166]. As the name suggests, the issue arises when many statistical hypothesis tests are performed at the same time. Indeed, if the significance level of single statistical test is set to $\alpha = 0.05$, the accepted false positive rate for a test, i.e. the probability of falsely rejecting the null hypothesis for that test, is 5%. However, if the number of comparisons is increased, the probability of falsely rejecting the null hypothesis at least once, named the family-wise error rate, increases as well [166]. There exists a number of solutions to the problem of multiple comparisons that can generally be divided into two types. One approach uses probability theory, while the second uses permutation testing approaches to estimate the null distribution for the univariate statistic.

The simplest and firstly introduced probability theory method to correct for multiple comparisons is the Bonferroni correction, which dates back to 1936 [167]. Under the assumption that the N mass univariate tests performed are independent, the method adjusts the significance threshold to $\hat{\alpha} = \frac{\alpha}{N}$, typically resulting in a overly conservative procedure. The conservative nature of these approaches relies in its test independence assumption: imaging data typically present high spatial correlation between adjacent locations, hence the significance threshold should account for this and be set to a higher value. Moreover, instead of controlling for the family-wise error rate, imaging genetics studies usually can afford to a small proportion of false positive results in exchange of greater statistical power. For this reason, methods to control the proportion of significant

results that are false positive, named false discovery rate (FDR) corrections have been introduced [168]. One the firstly introduced procedures for this class of correction is the original work of Benjamini and Hochberg [169], which also assumes that the tests are either independent or positively correlated. Hence, Benjamini and Yekutieli [170] introduced a formula to control the false discovery rate regardless of dependence, but which is also more conservative. Later, Benjamini *et al.* [171] developed a new procedure intended to be less conservative when a large proportion of tests are testing a null effect.

Approaches using permutation testing can be divided into two classes [172]. The first approach estimates the distribution of the maximum effect across the whole image assuming the null hypothesis is true at every location, and the statistical significance of the tests at every pixel or voxel is compared against it. This distribution is obtained by computing the maximum value of the statistic under exam for every permutation or bootstrap of the data, and it takes into account their spatial correlation. The second approach makes use of cluster-based statistics and it is the most popular [173]. In particular, it finds clusters of voxels or pixels with effects surpassing some threshold (often the t-value or F-value that would be significant with no correction) or according to some other condition [166], and sums all the statistic values to form a cluster mass statistic. A permutation testing approach is then used to estimate the null distribution for this statistic by finding the largest cluster across a large number of random permutations or bootstrapped samples [174].

The General Linear Model

In mass univariate approaches, a general linear model (GLM) is typically fitted at each pixel or voxel location to model relationship between the intensity value there and, for instance, a candidate gene or SNP of interest [88, 161]. In a GLM, the relationship between a random variable $\boldsymbol{y} = (y_i) \in R^n$ representing the intensity value at a pixel or voxel in the n subjects under study is modelled using a linear combination of predictors $\boldsymbol{\beta x}$, where $\mathbf{x}$ is the vector of predictors and $\boldsymbol{\beta}$ the magnitude of their effect to be estimated. The

predictors $\mathbf{x}$ can also include set of covariates (such as age, sex, ethnicity) to account for known clinical variability. Moreover, a term accounting for the model error and which assumes it as an independent identically distributed normal error $\boldsymbol{\epsilon}$ is also added, resulting in a final model of the form $\mathbf{y} = \boldsymbol{\beta}\mathbf{x} + \boldsymbol{\epsilon}$. The values of the coefficients $\boldsymbol{\beta}$ are usually computed through maximum likelihood estimation, $\boldsymbol{\beta} = (\mathbf{X^T X})^{-1}\mathbf{X^T y}$ [88]. To test significance their statistical significance, the t-statistic is calculated by computing the difference between the estimated values of $\boldsymbol{\beta}$ and the value of $\hat{\boldsymbol{\beta}}$ under the null hypothesis (which is expressed as $\hat{\boldsymbol{\beta}} = 0$ in case of no significant association), and by dividing it for the standard error (SE) of $\hat{\boldsymbol{\beta}}$, i.e. $t = \frac{\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_0}{SE(\boldsymbol{\beta})}$.

Mass Univariate Analysis of the Heart

Mass univariate analysis has only recently been applied to 3D cardiac imaging, and although the approach is feasible in population-based studies, it may provide overly-conservative inferences without considering spatial dependencies in the underlying data [83, 175, 89]. The first application in the cardiac domain is the work of de Creane *et al.* [89], in which mass univariate analysis is employed to quantify abnormal myocardial motion from velocity fields. Previous work of our group employing mass univariate analysis demonstrated how adverse regional remodelling is associated with increased systolic blood pressure in healthy volunteers [83]. However, these mass univariate analyses did not derive any significant vertices after multiple testing correction, even though large and plausible regions of the ventricle demonstrated similar association coefficients and reached significance before multiple testing correction.

2.3   Deep Learning for Shape Analysis

2.3.1   Deep Learning

Deep learning has revolutionised many computer vision and medical imaging applications due to its ability to learn hierarchies of task-specific features directly from data, largely outperforming traditional approaches employing handcrafted features [74, 77, 176]. Examples include image classification [177], semantic segmentation and object detection [178, 179, 78], natural language processing [180] and speech recognition [181]. Given input data $\boldsymbol{x}$, and a desired output $\boldsymbol{y} = f(\boldsymbol{x})$, deep learning models approximate f by an artificial neural network composed of a deep hierarchy of processing layers. These layers consist of a set nodes (or neurons) which apply non-linear and/or linear transformations to their input.

The two most common and basic layers adopted in deep learning models are the fully-connected and convolutional layers. Given a layer composed of a set of S nodes and P input nodes, the output of each node o_i in a fully-connected layer is obtained as

$$o_i = \sum_{p \in P} \omega_{ip} \ i_p + b_i \tag{2.1}$$

where ω_{ip} is the weight of the connection between an input node p (which can be the network input or the output of a previous layer) and the node i, while b_i is a bias term. Fully-connected layers are usually followed by a non-linear function named activation function. Stacking fully-connected layers on top of each other creates one of the most simple neural networks architectures: the multi-layered perceptron (MLP). Despite its simplicity, an MLP is able to compute virtually any non-linear complex function f by exploiting matrix multiplication. However, despite this architeture and its principles were introduced in the early sixties [182], efficient training of deep neural networks was only made possible in the last decade by the combination of fast graphics processing units
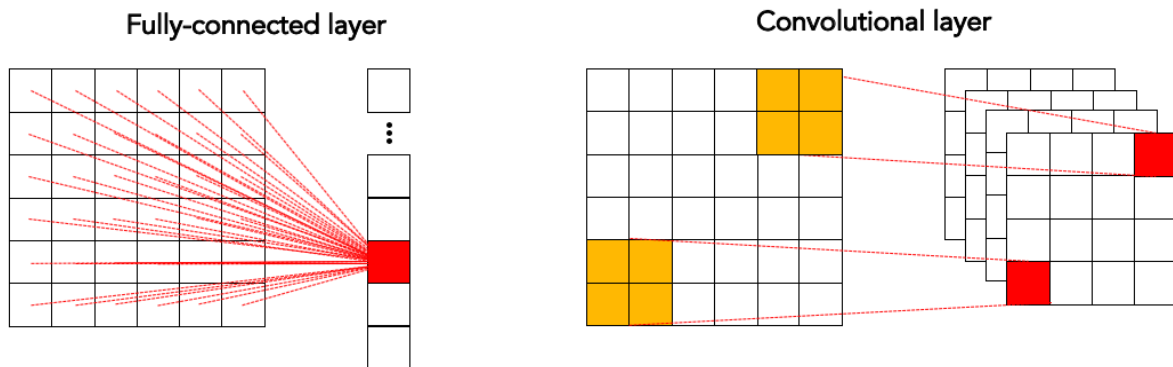
Figure 2.6: Spatial connectivity comparison of fully-connected vs convolutional layers.

(GPUs) and stochastic optimization algorithms.

Deep networks parameters $\boldsymbol{\theta}$, which consists of all the weights ω and biases b_i used by a neural network to model $f \simeq f_{\boldsymbol{\theta}}$, are learned from data via the optimization of an objective function $g(\boldsymbol{\theta})$. Being the optimization process non-convex, stochastic gradient-based optimization on small subsets of the training data (called mini-batches) is mostly employed for fast and effective network training [77]. This consist of iteratively updating the network parameters by an update rule on mini-batches of the training data. In its simplest formulation, the update rule involves modifying each model parameter $\theta_t \in \boldsymbol{\theta}_t$, at an iteration t, by a term proportional to the derivative of the objective function with respect to the parameter $\theta_t = \theta_{t-1} - \alpha \nabla_\theta g_t(\theta_{t-1} - 1)$. The derivative $\nabla_\theta g_t(\theta_{t-1} - 1)$ can be efficiently computed via the back-propagation algorithm and chain rule, while α is called the learning rate [183]. In the most recently adopted procedures, improved update rules exploiting the mean and variance of the gradient to update the learning rate over time, including the currently most frequently used Adam optimizer [184], have been proposed to improve network training.

Convolutional layers employ the convolution operation instead of general matrix multiplication as in fully-connected layers (a schematic representation is reported in Figure 2.6). Thus such layers only need to learn the values of a small convolution kernel $\boldsymbol{K}$, instead of the all weights connecting input-output nodes pairs, a property called local connectivity.
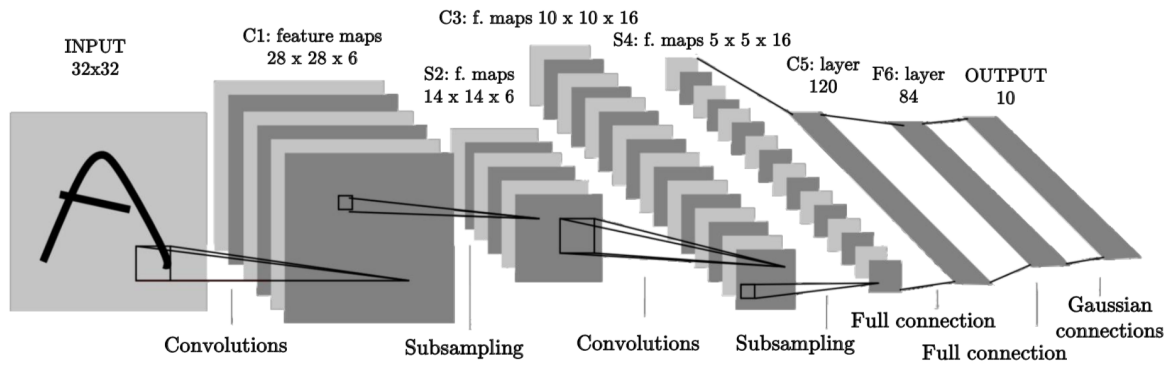
Figure 2.7: LeNet-5 architecture. The input image is fed to two convolution and pooling (subsampling), and then to two fully connected layer to predict handwritten digits.

This effectively reduces the number of parameters to be learnt and increases efficiency. Moreover, convolutional layers better exploit the statistical correlation between neighbouring nodes on grid-structured (euclidean) data, such as 2D images. In particular, a convolutional kernel can learn specific corners and edges and respond to them similarly at any location in the image, a property called translation invariance, in contrast to fully-connected layers which can recognize them only if they are in the same position in the training set. In a convolution layer the convolutional operation is usually followed by an activation function, as in MLPs. The number of pixels the filter moves each time (striding), the kernel dimension and the amount of padding adopted in a convolution layer all influence the resulting output, also called feature map. At each layer, many filters can be applied resulting in as many feature maps.

Fully-connected and convolutional layers together with pooling operations have been the backbone of many of the earliest deep learning architectures. Pooling operations are adopted to down-sample the input shape by replacing a small input region with a single value, for example its maximum value in *max* pooling. In Figure 2.7, as an example, one of the first CNNs architectures which have been applied commercially (for handwritten numbers recognition on cheques), named *LeNet-5*, is reported. *LeNet-5*, as all CNN architectures, learns low-level feature maps in the first layer, while more complex high-level features will be learnt at deeper layers before its output. In order to classify

different digits, a final fully-connected layer with a number of nodes K equal to possible digits constitutes the output of the network, to which the softmax function is applied activation function. The softmax function normalises the output nodes values so that each item is in the range (0, 1) and their sum is 1:

$$\sigma(o)_i = \frac{e^{o_i}}{\sum_{k=1}^{K} e^{o_k}} \quad i = 1, \ldots K \tag{2.2}$$

An example of objective function is categorical cross-entropy

$$g = H(p, q) = -\sum_{i=1}^{K} p(i) \log q(i) \tag{2.3}$$

where $p(i)$ is equal to 1 or the correct category, 0 otherwise, and $q(i) = \sigma(o)_i$. The minimisation of g via a stochastic gradient-based optimization procedure enables the learning of discriminative digits features directly from the training data and different features will be derived for different tasks. Thus, no hand-engineered image features and ad-hoc preprocessing methods as in traditional machine learning are needed. This training methodology is also referred to as *end-to-end* learning and also enables the mutual training and optimization of multiple networks and objectives simultaneously [77].

Deep learning models, as all machine learning models, can both suffer of overfitting and underfitting problems. Overfitting arises when the network parameters model the training data too well, including details and noise of the training data, which is a common issue in deep networks due to their overparametrization [77]. This prevents the model to learn the overall data pattern, causing poor generalization over unseen datasets. On the contrary, underfitting happens when the learned model is unable to learn appropriate data patterns, typically due to insufficient number of parameters or inadequate training. Disadvantages typically associated with deep learning in comparison with standard machine learning models include the need of large amount of labeled data for these algorithms in order to

perform better than traditional methods, their complexity and lack of standard theory and interpretability.

2.3.2   Shape Representation Learning

The shape analysis field has not been left untouched by deep learning breakthroughs, aiming at learning better shape representations. What makes a good shape representation? In broad terms, a good shape representation is *"one that makes a subsequent learning task easier"* [77]. Representation learning is an active research topic in computer vision and graphics, and shape representation learning is a subtopic which aims at enabling complex 3D reasoning, aiding visual recognition and partial or whole shape synthesis tasks by learning robust, complex shape features from data [185, 186]. In the same fashion, medical shape analysis is interested in learning good shape representations, generally for disease prediction, shape synthesis and unsupervised clustering [119, 187, 188, 189, 190]. In addition, the constraint of interpretability is further added in medical shape analysis, as the motivation behind a clinical prediction or a shape synthesis process is often as important as the accuracy with which the task is performed [191, 60].

A key decision to make when studying 3D shapes is about their parametrization, which is typically selected depending on the application [192, 193, 194]. Deep learning approaches showed great performance analysing data organized on euclidean grid-like structures, such as 2D/3D images, on which translation invariance, local connectivity, and compositionality properties of CNNs can be exploited [195]. For this reason, voxelization, the process of converting a shape into a 3D grid that encodes the geometric information, has characterized the first applications and has been widely used to enable deep learning of 3D shape representations [196, 197, 198, 199, 193, 200]. Alternatives including projecting 3D data into multiple views [201], or using surfaces meshes [202], point clouds [186] or skeletons [203] have been also successfully explored.

As in traditional approaches, large part of the deep learning applications for shape analysis have resorted to learn a low-dimensional parametrization to model shape variability, also called shape descriptors [194]. These approaches have often been seen as a replacement of PCA-based approaches, providing excellent results in 3D object synthesis, matching, recognition and retrieval [204, 205, 206, 201, 185, 199, 207, 202, 200]. Their superior performance over PCA components or hand-crafted representations lies in their ability to learn both local and global shape descriptors, and to be learnable end-to-end and task-specific from data, thus requiring minimal expert knowledge [194]. At the core of such approaches are the autoencoder and variational autoencoder approaches, which will be introduced in detail in the next two sections. This section is then concluded with a discussion on intrepretable deep learning, of crucial importance for medical image and anatomical shape analysis.

2.3.3 Autoencoder

An autoencoder (AE) is a non-linear dimensionality reduction method which employs a deep neural network to learn a compact feature representation $\boldsymbol{z} \in \mathcal{Z} \subset R^d$ of input data $\boldsymbol{x} \in \mathcal{X} \subset R^D$, with typically $d << D$ [183, 208]. It consists of two networks: an encoder network $f_\phi : \mathcal{X} \to \mathcal{Z}$ and a decoder network $g_\theta : \mathcal{Z} \to \mathcal{X}$, *i.e* the encoder network is a function f that maps the input $\boldsymbol{x}$ to $\boldsymbol{z}$, while the decoder is a function g that aims at reconstructing the input $\boldsymbol{x}$ from the latent code $\boldsymbol{z}$, $\boldsymbol{x'} = g(\boldsymbol{z})$. A schematic representation of the AE architecture is reported in Fig. 2.8. The two networks are learned by minimizing a loss function $\phi, \theta = argmin_{\phi,\theta} \mathcal{L}(\boldsymbol{x'}, \boldsymbol{x})$ which measures the quality of the reconstructions, making the $\boldsymbol{x'}$ as close as possible to the original input $\boldsymbol{x}$. If f_ϕ and g_θ are parameterized with one linear layer and the loss function $\mathcal{L}$ is computed through the mean squared error, an AE can be proved equivalent to PCA [176].

AEs have been traditionally categorized as unsupervised learning algorithms [77]. In this setting, the work of Zhu *et al.* [201] computes a 3D shape descriptor by projection of
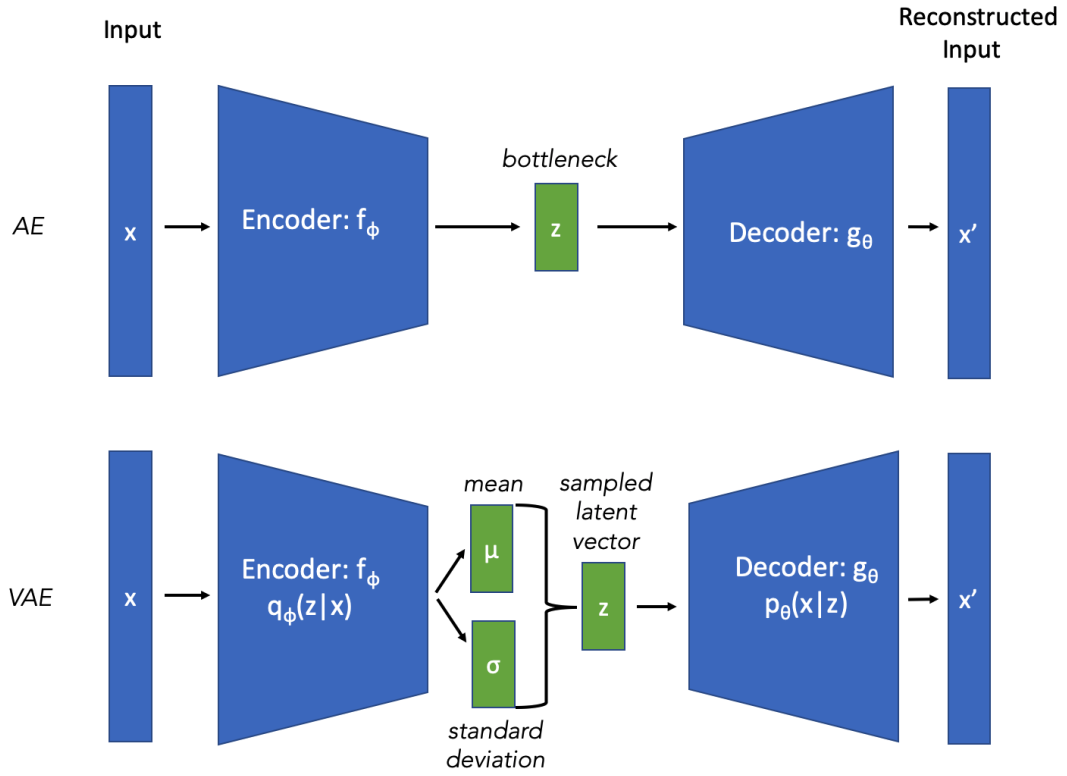
Figure 2.8: Scheme of the Autoencoder (AE) and Variational Autoencoder (VAE) architectures. Both architectures aim at reconstructing the input $\boldsymbol{x}$ by encoding and decoding it into a latent space $\boldsymbol{z}$. However, the VAE architecture learns a distribution over it - a multivariate Gaussian with a diagonal covariance structure in the reported figure.

the 3D object into 2D images and by aggregating their AE-derived latent representations into a unique shape descriptor. However, AEs are often framed as discriminative models when used for deep learning of shape descriptors [194]. A discriminative model is a model that aims at learning the decision boundary between classes. To achieve this, the popular works of Fang *et al.* [204], Dai *et al.* [209] and Xie *et al.* [205] employ many-to-one AEs, which decode shapes of the same category to a unique target value in the decoding step. This forces the latent representation of same class shapes to be similar, while distinct class shape will have representations as different as possible, hence making the shape descriptor sensitive to inter-group structural variations.

Autoencoder-based models have also been used to learn compact representations of medical images, aiding many image processing tasks [74]. These tasks include deep neural network pre-training [210, 211], image denoising [212], real-time image reconstruction [213],

incorporating shape priors for semantic segmentation [214, 215, 216], low-dimensional representations of diffeomorphism shape matching [217], content-based retrieval of medical images [218]. Moreover, Oktay *et al.* [215] showed how autoencoder-derived features of LV segmentations outperform PCA features in the classification of healthy subjects versus dilated cardiomyopathy and HCM patients on the MICCAI ACDC dataset [60].

2.3.4 Variational Autoencoders

While discriminative models learn the decision boundary between classes, generative models explicitly model the class distributions. This can enable the synthesis of new, unseen shapes, which is not possible, for example, with standard AE models. Formally, given input data $\boldsymbol{x}$ and their class label $\boldsymbol{y}$, a discriminative model learns the conditional probability $p(\boldsymbol{y}|\boldsymbol{x})$ of a label given the input from the train data. Conversely, a generative model learns the joint probability $p(\boldsymbol{x}, \boldsymbol{y})$ of the inputs $\boldsymbol{x}$ and labels $\boldsymbol{y}$ [219]. Within the class of generative models, the variational autoencoder (VAE) [220] has been successful at learning generative models of deforming 3D shapes for a variety of applications, including shape space embedding and generation, outperforming state-of-the-art methods [185], [202].

The idea behind VAEs is to learn a latent variable model $p_\theta(\boldsymbol{z}, \boldsymbol{x}) = p_\theta(\boldsymbol{z})p_\theta(\boldsymbol{x}|\boldsymbol{z})$ of learnable parameters θ, latent variables $\boldsymbol{z} \in \mathcal{R}^d$ and observed variable $\boldsymbol{x} \in \mathcal{R}^D$, $d \ll D$. The observed variable $\boldsymbol{x}$ is a shape parametrisation in the context of shape analysis, for example a 3D segmentation or point cloud. $p_\theta(z)$ is the prior distribution over the latent variables z, $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ is the generative (decoding) function. By marginalising over the latent variables z, the VAE model evidence can be written as

$$p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} p_\theta(\boldsymbol{x}, \boldsymbol{z})d\boldsymbol{z} = \int_{\boldsymbol{z}} p_\theta(\boldsymbol{x}|\boldsymbol{z})p_\theta(\boldsymbol{z})d\boldsymbol{z} \tag{2.4}$$

Given a training set of N training examples $X = \{\boldsymbol{x}_j, \ j = 1,...N\}$, the direct solution

of $\log(p_\theta(\boldsymbol{x}))$ for these N example is computationally infeasible, as it requires to compute the integral in Eq. 2.4 over all the z values. In the case of VAEs, variational inference is employed to solve this problem, by casting it as an optimization problem. In particular, a variational distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ to approximate the posterior distribution $p_\theta(\boldsymbol{z}|\boldsymbol{x})$ of the latent variables $\boldsymbol{z}$ is introduced. This function $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is also called encoder network, and its parameters ϕ are called variational parameters. These parameters are optimized so that $q_\phi(\boldsymbol{z}|\boldsymbol{x}) \approx p_\theta(\boldsymbol{z}|\boldsymbol{x})$. The optimization objective of the VAE is the evidence lower bound (ELBO), which can be obtained by manipulating the log-likelihood $\log\ p_\theta(\boldsymbol{x})$ of the training data $\boldsymbol{x}$ as follows.

$$\log\ p_\theta(\boldsymbol{x}) = \mathop{\mathbb{E}}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x})] = \tag{2.5}$$

$$= \mathop{\mathbb{E}}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log\left[\frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})}\right]\right] = \mathop{\mathbb{E}}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log\left[\frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\frac{q_\phi(\boldsymbol{z}|\boldsymbol{x})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})}\right]\right] = \tag{2.6}$$

$$= \mathop{\mathbb{E}}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log\left[\frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right]\right] + \mathop{\mathbb{E}}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log\left[\frac{q_\phi(\boldsymbol{z}|\boldsymbol{x})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})}\right]\right] \tag{2.7}$$

The second term in Eq. 2.7 is known as Kullback-Leibler (KL) divergence between $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and $p_\theta(\boldsymbol{z}|\boldsymbol{x})$, $\mathcal{D}_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x}))$. The closer the two distributions are, the smaller the KL value is, which is by definition non-negative. Given this, the first term in Eq. 2.7 is a lower bound of $\log\ p_\theta(\boldsymbol{x})$, named evidence lower bound or variational lower bound. Thus, the maximization of the ELBO is equivalent to compute and minimize the KL divergence $\mathcal{D}_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x}))$, which would be computationally intractable. Given input data $X = \{\boldsymbol{x}_j,\ j = 1,...N\}$, the ELBO value for each data-point can also be rewritten as:

$$\mathcal{L}_{\theta,\phi}(\boldsymbol{x}_j) = \mathop{\mathbb{E}}_{q_\phi(\boldsymbol{z}|\boldsymbol{x}_j)}\left[\log\left[\frac{p_\theta(\boldsymbol{x}_j, \boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x}_j)}\right]\right] = \mathop{\mathbb{E}}_{q_\phi(\boldsymbol{z}|\boldsymbol{x}_j)}\left[\log\left[\frac{p_\theta(\boldsymbol{x}_j|\boldsymbol{z})p_\theta(\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x}_j)}\right]\right] = \tag{2.8}$$

$$= \underset{q_\phi(\boldsymbol{z}|\boldsymbol{x}_j)}{\mathbb{E}} \left[\log\left[p_\theta(\boldsymbol{x}_j|\boldsymbol{z})\right] - \underset{q_\phi(\boldsymbol{z}|\boldsymbol{x}_j)}{\mathbb{E}} \left[\log\left[\frac{q_\phi(\boldsymbol{z}|\boldsymbol{x}_j))}{p_\theta(\boldsymbol{z})} \right] \right] \right] = \qquad (2.9)$$

$$= \underset{q_\phi(\boldsymbol{z}|\boldsymbol{x}_j)}{\mathbb{E}} \left[\log\left[p_\theta(\boldsymbol{x}_j|\boldsymbol{z})\right] - \mathcal{D}_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}_j)||p_\theta(\boldsymbol{z})) \qquad (2.10)$$

which can be optimized via stochastic gradient descent by minimizing its negative value. The first term in the lower bound represents a reconstruction loss, i.e. how accurate is the generative model $p_\theta(\boldsymbol{x}_j)$ in the reconstruction of the input data $\boldsymbol{x}_j$ from the latent space values $\boldsymbol{z}$ using the generative (or decoder) network $p_\theta(\boldsymbol{x}_j|z)$. In a deep learning model, the functions $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and $p_\theta(\boldsymbol{x}_j|\boldsymbol{z})$ are parametrised by two neural networks f_ϕ and g_θ. The second term is a regularization term that makes $q_\phi(\boldsymbol{z}|\boldsymbol{x}_j)$ match with its prior distribution $p_\theta(\boldsymbol{x}_j)$ on the latent variables $\boldsymbol{z}$. A common choice of the form of $q_\phi(\boldsymbol{z}|\boldsymbol{x}_j)$ is a multivariate Gaussian with a diagonal covariance structure, *i.e.* $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2\boldsymbol{I})$. Figure 2.8 illustrates a scheme of the VAE architecture in this setting.

Due to their generative properties, VAEs have been used in an extensive number of computer vision applications, including high-resolution random image generation [221], small graph and molecule generation [222], semi-supervised text classification [223], prediction of scene evolution [224]. Moreover, VAEs have been successfully used for 3D object shape synthesis and completion [185, 202] and for single image voxelized 3D shape retrieval [225, 226]. In the medical imaging domain, VAEs have been exploited to approximate the distribution and likelihood of previously unseen MR images [227], to learn a low-dimensional manifold of 3D fetal skull segmentations [228] and hippocampal meshes [229], to learn a low-dimensional probabilistic deformation model for cardiac image registration [230] and to learn a domain-invariant latent space of segmentations for cross-modality image segmentation [231]. In the same domain, 3D VAEs alone or coupled with another convolutional neural newtwork outperformed prediction accuracy of several state-of-the-art predictors of Alzheimer disease when trained to capture anatomical shape variations

in structural MRI scans of the brain [229, 232].

2.3.5   Interpreting Deep Learning Models

In contrast to traditional machine learning models, deep learning methods are not transparent by design, *i.e.* they are not by themselves understandable by a human [233, 234, 191]. This is an issue of crucial importance in medical image analysis as computer-based systems need to prove reliable and explainable to both clinicians and patients. In particular, such methods are required to provide the same performance in the face of distributional shifts, such as the ones arising from imaging protocol changes, different pre-processing or patient demographics, and changes in scanner design and construction. This proved to be extremely challenging with deep learning models due to their tendency to overfit [235, 236]. Moreover, an increasing body of legislation, such as the new European Union's General Data Protection Regulation, allows patients to ask for an explanation of an algorithmic decision that has been made about them, the so named right to explanation, making even more challenging the application of deep learning models in the clinical practice [237].

An increasing number of works have been focusing on understanding deep learning networks, especially CNNs based models, trying to reduce their often criticised *black-box-ness* nature. This is often attempted by using *post-hoc* explainability techniques which aim at investigating how a trained deep model makes its decisions [233]. An interesting direction pursed for *post-hoc* analysis is based on uncertainty analysis. Two types of uncertainty measures are typically defined in these works: epistemic, which measures how small perturbations of training data modify model parameters, and aleatoric uncertainty, which measures how input parameter changes would affect the prediction for one particular example [238, 239]. Notably, measures of uncertainty have been employed in medical image analysis to quantify the uncertainty over the network output, which can be used to improve diagnostic performance [240, 241] and to assess segmentation accuracy [242, 243].

The most popular track of works however consists in the analysis of the input or intermediate feature maps obtained by either exploiting backpropagation algorithms and their modification [234, 244], or by sampling input patches that maximize activation of hidden units [245, 246]. In the medical image analysis domain these approaches have been used to indicate whether shape and edge information of the structure of interest are used in the network prediction [247, 248]. The obtained visualisations are able to show insights of the low and high level image patterns learned, providing a qualitative yet subjective means for model interpretation. Moreover, such approaches typically do not scale well when many layers are employed, and typically require an observer to manually assess many images and their saliency maps to draw general conclusions [249, 250].

Chapter 3

Three-dimensional Cardiovascular Imaging-Genetics: A Mass Univariate Framework

This chapter is based on: Biffi, Carlo, *et al.* "Three-dimensional cardiovascular imaging-genetics: a mass univariate framework." *Bioinformatics* 34, no. 1 (2018): 97-103.

3.1 Introduction

One of the most complex unanswered questions in cardiovascular biology is how genetic and environmental factors influence the structure and function of the heart as a 3D organ [62]. Genetic testing is increasingly used to confirm or rule out diagnosis of inherited cardiac conditions and to assess relatives' disease risk, and it is increasingly included in large-scale population-based studies [64]. However, our understanding of the penetrance and expressivity of variants associated with inherited cardiac conditions as well as the bi-
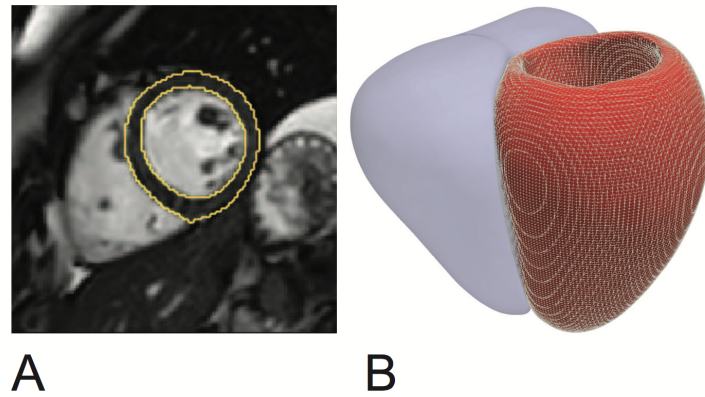
Figure 3.1: UK Digital Heart 3D high-resolution phenotypes derivation. (A) Short axis cardiac magnetic resonance image demonstrating automated segmentation of the endocardial and epicardial boundaries of the left ventricle. (B) The segmentation is used to construct a three dimensional mesh of the cardiac surfaces (left ventricle shown in red, right ventricle in blue) that is co-registered to a standard coordinate space. Phenotypic parameters, such as wall thickness, are then derived for each vertex in the model.

ology of heart development and within-population variation remains limited [251, 69, 146]. One of the motivations behind this is partially attributed to the lack of precision and resolution in the assessment and modeling of the cardiac phenotype [73, 69]. CMR is the gold-standard for quantitative structural imaging of the heart [18, 3], providing a rich source of high-resolution anatomic and motion-based data. However, conventional phenotyping relies on manual analysis, often reducing the detailed imaging information to crude global indices of volume and mass, only relatively sensitive to regional or asymmetric morphological and functional changes [58, 82]. Coupling CMR images with fully automated segmentation algorithms enables high-throughput 3D phenotyping of the cardiac phenotype within a population of interest [63, 85]. Moreover, it facilitates the construction of high-resolution computational atlases (Figure 3.1), which provide a means to study 3D phenotypic variation and its determinants at thousands of vertices in a standardized coordinate space (more details reported in Section 2.1.1). Therefore, the development of methods enabling sensitive and reproducible mapping of genotype-phenotype associations, *i.e.* associations between genetic variation and 3D high-resolution phenotypes, constitute an unprecedented occasion to help refining our knowledge of the pathogenicity of genetic variants [65, 63].

3.1.1   Related Work

One common approach to uncover genotype-phenotype associations is to transform the spatially-correlated phenotypic data of the imaged 3D high-dimensional phenotype into a smaller set of latent variables [41, 117], for example by exploiting PCA to derive a set of uncorrelated principal components or similar dimensionality reduction approaches [115]. However, this new set of principal variables would not provide an explicit and easily interpretable model relating genotype to phenotype, with reduced applicability to clinical studies. Furthermore, the obtained low-dimensional description might not encode the anatomical variation of interest. An approach widely used in functional neuroimaging [88], named mass univariate analysis, offers an interesting alternative (Section 2.2.3 provides a general introduction to the topic). This class of approaches derives an univariate statistic expressing evidence of a given genotype-phenotype effect at each vertex of the 3D atlas template, independently, using a general linear model. This produces a so-called statistical parametric map which, after multiple testing correction, enables the direct appreciation of both regional and global genotype-phenotype associations to a human investigator, and which also accounts for additional clinical covariates. In the cardiac domain, preliminary work from de Marvao *et al.* [83] showed how the increase of systolic blood pressure (SBP) is associated with regional patterns of LV hypertrophy and increased LV wall stress in hypertensive subjects, despite not achieving statistical significance after the required multiple testing correction. This was a consequence of the overly-conservative nature of this correction when spatial dependencies in the underlying data are not considered. In my MRes project report [252], I translated a statistical approach from neuroimaging research, named threshold-free cluster-enhancement (TFCE) [166], for the sensitive detection of spatially coherent association signals which guarantees, at the same time, a robust control for multiple testing. The proposed approach resulted in a mass univariate framework which achieved statistical significant results in all the experimental settings investigated by de Marvao *et al.* [83].

3.1.2 Contribution

In this chapter, the previously proposed mass univariate approach I proposed during my MRes thesis project is extended and validated to the cardiovascular imaging-genetics domain through experiments on real genetic and synthetic data, with a special focus on GWAS results validation. In order to achieve this, a fast and memory-efficient R package is also introduced and made publicly available providing all the necessary tools to perform this class of studies[1].

3.2 Study Population

3.2.1 Imaging and Clinical Data

The proposed method is tested on 1,124 Caucasian healthy volunteers from the UK Digital Heart Project at Imperial College London [56]. Healthy volunteer refers to an individual who did not have any known cardiovascular or metabolic disease, or that was not taking prescription medicines at the time of the scan. Cardiac MR (CMR) was performed on a 1.5-T Philips Achieva system (Best, the Netherlands). To capture the whole-heart phenotype, a high-spatial resolution 3D bSSFP cine sequence was used that assessed the LV in a single breath-hold (60 sections, repetition time 3.0 ms, echo time 1.5 ms, flip angle 50°, field of view $320 \times 320 \times 112$ mm, matrix 160×95, reconstructed voxel size $1.2 \times 1.2 \times 2$ mm, 20 cardiac phases, temporal resolution 100 ms, typical breath-hold 20 s). Images were stored on an open-source database (MRIdb, Imperial College London, UK) [253]. Conventional volumetric analysis of the cine images was performed using CMRtools (Cardiovascular Imaging Solutions, London, UK) following a standard protocol [254]. Blood pressure was acquired in accordance with the guidelines of the European Society of Hypertension [255] using a calibrated oscillometric device (Omron

---

[1]https://github.com/UK-Digital-Heart-Project/mutools3D

	Full Cohort ($N = 1,124$)	**Males** ($N = 511$)	**Females**($N = 613$)
Age [*years*]	43.4 ± 13.3 (19-77)	43.2 ± 13.0 (19-77)	43.5 ± 13.2 (20-77)
BSA [m^2]	1.84 ± 0.2	1.98 ± 0.16	1.72 ± 0.14
SBP [$mmHg$]	119.3 ± 14	125.0 ± 12.7	114.65 ± 13.2

Table 3.1: A summary of the 1,124 Caucasian subjects of UK Digital Heart Project at Imperial College cohort employed in this work.

M7, Omron Corporation, Kyoto, Japan). Five measures were taken and last two were averaged to obtain the final value. Body surface area (BSA) was calculated by Mosteller formula $BSA(m^2) = \sqrt{\frac{Ht(cm)*Wt(kg)}{3600}}$, after measurement of height (Ht) and weight (Wt).

3.2.2 Genetic Data

Genotyping of common variants in the 1,124 Caucasian subjects was carried out using an Illumina HumanOmniExpress-12v1-1 single nucleotide polymorphism (SNP) array (Sanger Institute, Cambridge). Clustering, calling and scoring of SNPs was performed using Illumina GenCall software [256]. Samples were pre-phased with SHAPEIT [257] and imputation was performed using IMPUTE2 [258] with the UK10K dataset as a reference (www.uk10k.org). Quality of the genotypes was evaluated both on a per-individual and per-marker level using in-house Perl scripts and PLINK software [259]. SNPs were removed if they had a Impute Information (INFO) score < 0.4, missing call rate in more than 1% of samples, minor allele frequency of less than 1% or deviated significantly from Hardy-Weinberg equilibrium ($p > 0.001$). Only non-related individuals with CEU ethnicity were retained. The total genotyping rate in these individuals was 0.997 and the total number of variants available was 9.4 million.

3.2.3 Atlas-based segmentation and co-registration

All image processing was performed with Matlab (MathWorks, Natick, Mass, USA). A previously validated cardiac segmentation framework exploiting previous knowledge of

heart anatomy from a set of manually annotated atlases was used. Details are reported in [96, 83]. After a rigid registration step in which all the obtained segmentations were rigidly aligned to an higher-resolution template segmentation, its corresponding high-resolution average template surface mesh was warped to each subject space by employing the inverse of the registration field from subject segmentation to template segmentation. This enabled the derivation of high-resolution subjects' meshes composed of 49,876 epicardial co-registered vertices in a standard coordinate space, and consequently a cardiac atlas encoding the phenotypic variation for the study population (Figure 3.1) [96]. In particular, for each mesh, wall thickness (WT) was measured by computing the distance between respective vertices on the endocardial and epicardial surfaces at end-diastole.

3.3 Method

In the following paragraphs the framework originally introduced in [252] is summarized and how it can be adapted to test associations between genetic variables and a 3D cardiac phenotype defined at each vertex of a computational atlas is explained in detail. The framework is outlined in Figure 3.2.

3.3.1 Mass Univariate Analysis

The degree of association between a 3D cardiac phenotype, such as LV wall thickness (WT), and genetic variables can be inferred via a general linear model. Given N subjects in a population under study, the general linear model has the form of $\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\epsilon}$, where

- $\mathbf{Y}$ is a $[1, N]$ vector containing, for example, the WT values of all the N subjects at one vertex of the LV.

- $\mathbf{X}$ is a $[V+1, N]$ design matrix used to model the hypothesis under investigation. $\mathbf{X}$ contains in its columns the subject's values of V clinical variables, and an intercept

term. Genetic variables in $\mathbf{X}$ can encode the absence/presence of a genetic mutation as categorical variables or the posterior estimate of allele frequency as a numerical variable. Their magnitude of association with the 3D phenotype can be adjusted for other co-variates, such as age or sex, by including them as additional columns of the matrix $\mathbf{X}$.

- $\boldsymbol{\beta}$ is the $[1, V+1]$ regression coefficient vector to be estimated. $\boldsymbol{\beta}$ can be standardized by normalizing to mean 0 and unit-variance the columns of $\mathbf{X}$ and $\mathbf{Y}$. In this way, $\boldsymbol{\beta}$ will represent the magnitude of variation of $\mathbf{Y}$ in units of its standard deviation when $\mathbf{X}$ is increased by one standard deviation. This enables the comparison of the association effects between different studied variables.

- $\boldsymbol{\epsilon}$ is a $[1, N]$ vector representing the variability of $\mathbf{Y}$ not explained by the linear model, and assumed to be a zero-mean Gaussian process [88].

The same general linear model can be fitted at each surface vertex independently, resulting in a mass univariate regression approach. The regression coefficients β at each vertex and their related p-value thus obtained can be plotted to display, at high resolution on the whole 3D anatomy, the magnitude and spatial distribution of a given association. However, this approach underestimates associations where the signal is more spatially correlated than noise coherence. For this reason non-parametric statistics such as TFCE are valuable to increase the statistical power of the approach.

3.3.2   Threshold-free cluster enhancement on a cardiac atlas

A surface atlas can be parametrized as a 3D mesh composed of non-congruent triangles containing at each vertex p the values of a phenotype variable for the N subjects under study. The value of the TFCE statistic at that vertex p of the atlas under exam for the
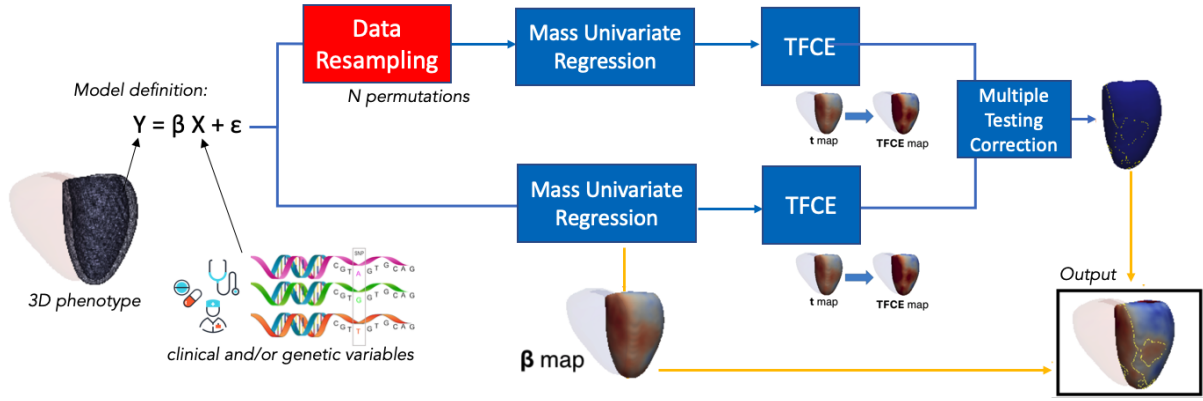
Figure 3.2: Outline of the proposed 3D mass univariate framework. A computational atlas provides 3D vertex-wise phenotypes of geometry and function which association can be tested with clinical and genetic variables via a general linear model. 3D maps representing the degree of association ($\boldsymbol{\beta}$) at each vertex can be derived. Permutation testing coupled with threshold-free cluster enhancement (TFCE) produces vertex-wise p-values weighted to the degree of coherent spatial support. Finally, multiple testing correction is applied to the obtained p-values. Regression coefficients enclosed by significance contours are represented on a model of the left ventricle.

studied model can obtained by computing:

$$TFCE(p) = \int_{h=0}^{h_p} e(h)^E \, h^H \, \delta h \simeq \sum_{h=0}^{h_p} e(h)^E \, h^H \, \Delta h \qquad (3.1)$$

where h is the value of a statistic obtained through mass univariate regression, a t-statistic in scenario under exam, and h_p is the value of the statistic at a vertex p. $e(h)$ is the spatial extent of the biggest cluster with cluster-forming threshold h that contains p, and E and H are two hyperparameters usually set to 0.5 and 2 for empirical and analytical motivations [166]. The integral can be computationally solved by computing its discretized form following the procedure I previously proposed in [252], and which is fully reported in the paper version of this Chapter [260].

Combining TFCE with permutation testing enables the computation of a new set of p-values at each atlas vertex p. The Freedman-Lane procedure [261] is adopted as it provided the best control of statistical power and false positives (type 1 error) in neuroimaging studies [262]. By sampling the data M times, M new models are computed and the

TFCE procedure is computed M times over the obtained statistical maps.

3.3.3   False discovery rate correction

A multiple testing problem arises by testing tens of thousands of statistical hypotheses simultaneously. Control of the family wise error rate at 5% could be derived by extracting the maximum score from each map derived via permutation testing and by using the 95th percentile as a threshold for significance. However, in this context such a correction could be overly conservative as we are rarely interested in the exact number of vertices that reach significance. The main goal is to detect extended areas of coherent signal and therefore we can accept a maximum fixed percentage of false discoveries as provided by false discovery rate (FDR) procedures. In particular, these procedures can be applied to adjust the voxelwise p-values obtained at each vertex by computing the ratio between the number of times in which a TFCE score greater than the measured one has been obtained and the number of permutations N. Also on genetic data, we have found adaptive procedures such as the two-stage Benjamini-Hochberg [171] not suitable for our dataset, since it led to lower p-values and increased areas of significance, as also reported in the neuroimaging literature [263]. For this reason, the original Benjamini-Hochberg (BH) [169] procedure has been employed for this work.

3.3.4   Software

I propose an R package (called *mutools3D*) which enables the reproduction of all the experiments reported in this Chapter. A full package documentation has been also produced to make it easily employable in future research. All the package functions enable their execution in parallel, and all the necessary calculations have been coded using matrix algebra to guarantee fast computations and reduced memory usage in comparison with built-in R functions. Matrices containing the phenotypic data and templates to visualise
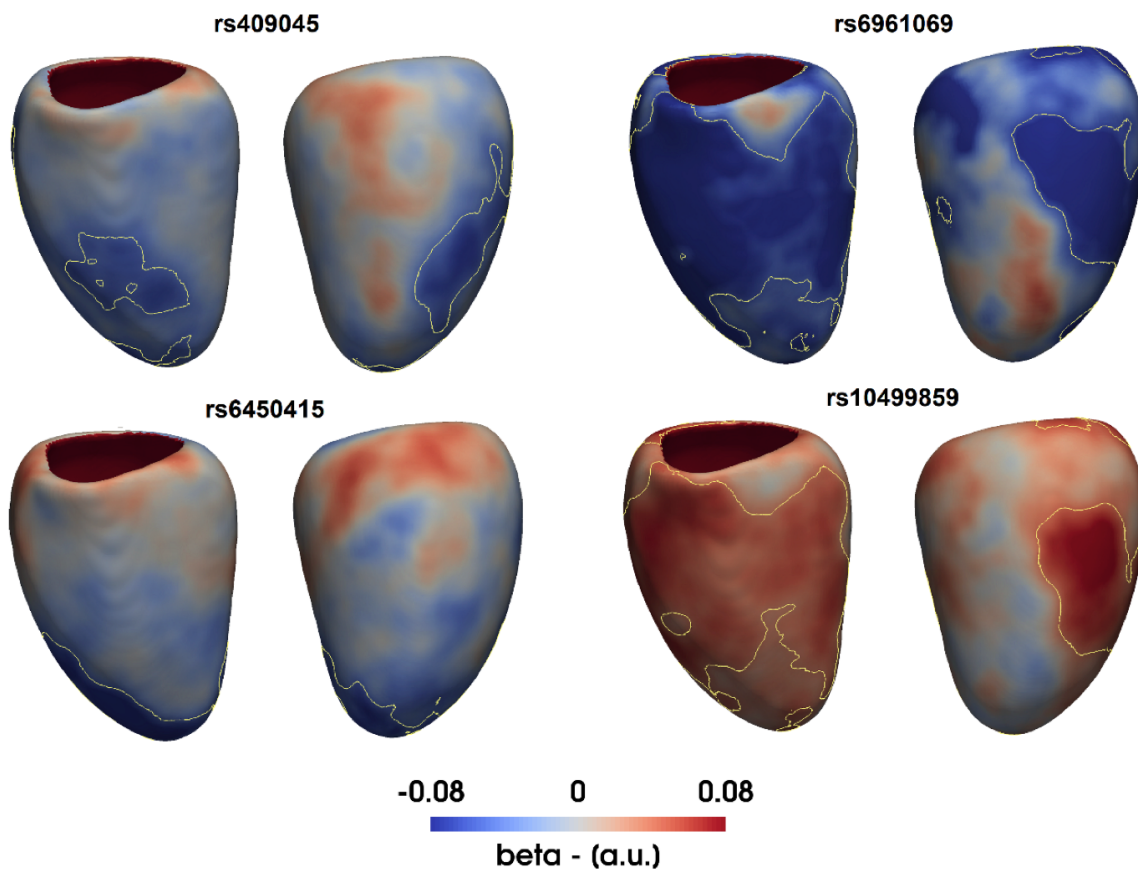
Figure 3.3: Applying three-dimensional analysis to single nucleotide polymorphism (SNP) replication. β coefficients are plotted on the surface of the left ventricle for the effect of 4 distinct SNPs on wall thickness (WT) adjusted for age, gender, body surface area and systolic blood pressure. Yellow contours enclose standardized regression coefficients reaching significance after multiple testing

the 3D models are also available with the R package[2]. Linear regression assumptions must be met in order to obtain reliable inferences [88]. Particularly important in this context are multicollinearity and heteroscedasticity problems which should be checked and solved for each model definition when using the proposed approach. Multicollinearity arises when high correlation between covariates is present, resulting a reduction of statistical power and ambiguous effects [264]. Heteroscedasticity arises when the error ϵ variance σ^2 is identical across observations, giving too much weight to certain subsets of the data when estimating the final βs. For this latter problem, the R package implements mass univariate functions exploiting HC4m heteroscedascity consistent estimators [265].

---

[2]https://github.com/UK-Digital-Heart-Project/mutools3D

3.4   GWAS Replication Study

In this section, the proposed pipeline is employed in a GWAS replication study. In particular, 6 out of 9 exonic SNPs were identified in the UK Digital Heart Project genotypes which have previously shown an association with LV mass in a case-control GWAS, using echocardiography for phenotyping [157]. For each SNP, WT at each vertex in the 3D atlas in 1,124 healthy Caucasian subjects was tested for association with the posterior estimate of the allele frequency by a regression model adjusted for age, gender, body surface area (BSA) and systolic blood pressure (SBP). The posterior estimate of the allele frequency is coded as 1 if the derived (mutant) allele was observed at that site, whereas 0 indicates the ancestral allele was observed. The tested SNPs are rs409045, rs6450415, rs1833534, rs6961069, rs10499859 and rs10483186. Regression diagnosis through Breush-Pagan and White's test showed how the homoscedasticity assumption was violated at a large number of vertices, therefore mass univariate regression was corrected using HC4m heteroscedascity consistent estimators [265]. Regarding the assumption of multicollinearity, the condition number of the model matrix was 2.19 while the variance inflation factor was equal to 1.06, suggesting a very low degree of multicollinearity. All the simulations were executed on a high performance computer (Intel Xeon Quad-Core Processor (30M Cache, 2.40 GHz), 36Gb RAM), using the analysis pipeline and R package proposed in this Chapter (Figure 3.2). A multiple comparisons procedure correcting for the number of vertices and the number of SNPs tested was applied by simultaneously testing in a BH FDR-controlling procedure all the TFCE-derived p-values from all the models as suggested in [266]. The number of permutations was fixed to 10,000 and simulations required less than 3 hours each with the proposed R package. Finally, as a result of a preliminary study we conducted (full details in Section 3.5.3), TFCE parameters E and H were set to 0.5 and 2, as suggested in the original TFCE paper [166], since this choice provides good sensitivity and specificity on a range of synthetic signals.

Four SNPs showed a significant association with WT as reported in Figure 3.3. These are

SNP	β	p-value
rs409045	0.06	0.17
rs6450415	0.01	0.75
rs1833534	-0.05	0.43
rs6961069	-0.01	0.96
rs10499859	0.01	0.84
rs10483186	0.01	0.74

Table 3.2: Regression coefficients and their related p-values of the linear association study between LVM and the posterior estimate of the allele frequency adjusted for age, gender, body surface area (BSA) and systolic blood pressure (SBP) of the presented GWAS replication study.

rs409045 (maximum regression coefficient $\bar{\beta} = -0.1$, percentage of the LV area significant $S = 13\%$), rs6450415 ($\bar{\beta} = -0.11$, $S = 11\%$), rs6961069 ($\bar{\beta} = -0.09$, $S = 44\%$) and rs10499859 ($\bar{\beta} = 0.1$, $S = 41\%$). Table 3.4 reports conventional linear regression analysis using LV mass (LVM) and the same model for all the SNPs. It can be seen that none of the models did not reach statistical significance, even without multiple testing correction.

3.5 Experiments on Synthetic Data

3.5.1 Sensitivity, Specificity and FDR Assessment

In this section, sensitivity, specificity and the rate of false discoveries (FDR) of the proposed pipeline are estimated using synthetic data. A 3D model showing no correlation between WT and the posterior estimate of the allele frequency X_{snp} of an non-associated SNP (rs4288653) adjusted for age, gender, BSA and SBP was used to generate background noise. A synthetic data signal was generated by summing to the WT values of each subject a term I β X_{snp} at each vertex, where I is the signal intensity and β is a map of regression coefficients. Two contrasting β maps (signal A and B) obtained from real clinical data were chosen and are shown in Figure 3.4. Signal A was characterized by non-null β coefficients covering 10% of the total area of the LV and scaled to the (0,1]
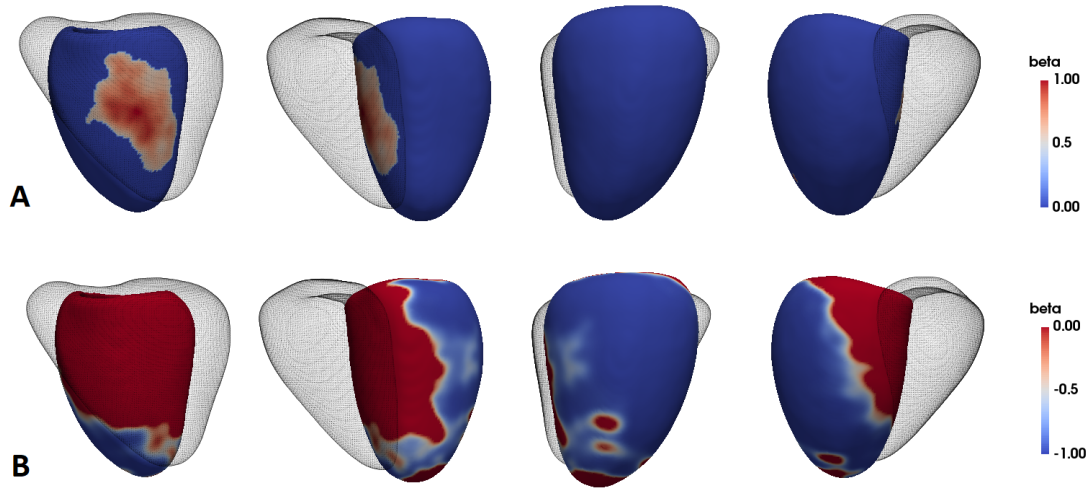
Figure 3.4: The two β maps used to generate the synthetic data for the statistical power calculation experiment - signal A (first row) and B (second row).
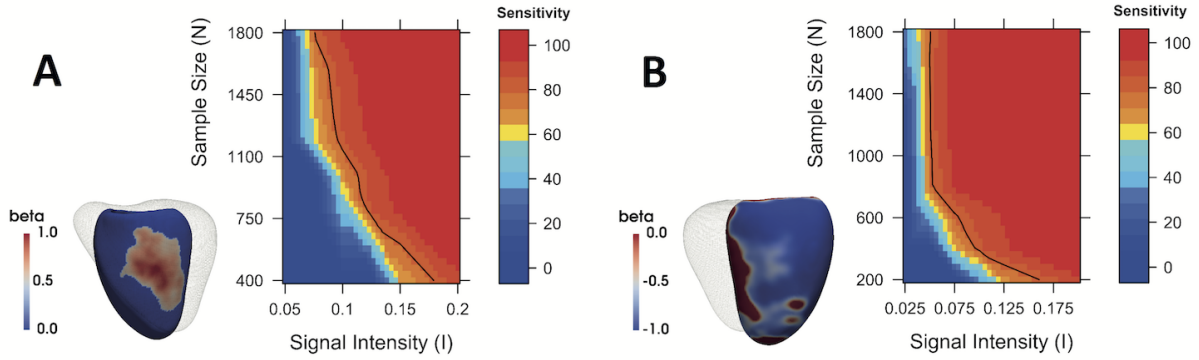
Figure 3.5: Assessment of statistical power using synthetic data. Plots of our framework sensitivity at different sample sizes N and signal intensities I to detect a synthetic signal on A) 10% and B) 60% of the LV surface. A black line on the plots indicates a threshold of 80% sensitivity.

range, while signal B presented non-null regression coefficients scaled to the [-1,0) range in a more extended area covering the 60% of the LV surface. By subsampling the number of subjects N and the signal intensity I, different signals to be detected by the proposed standard mass univariate pipeline were obtained. The number of permutations for each simulation was fixed to 5,000 and results were linearly interpolated and plotted on the contour plots shown in Figure 3.5.

Sensitivity increased at larger sample sizes N and signal intensities I, reaching the greatest values with the most extended signal (signal B) as expected. Given the sample size of our
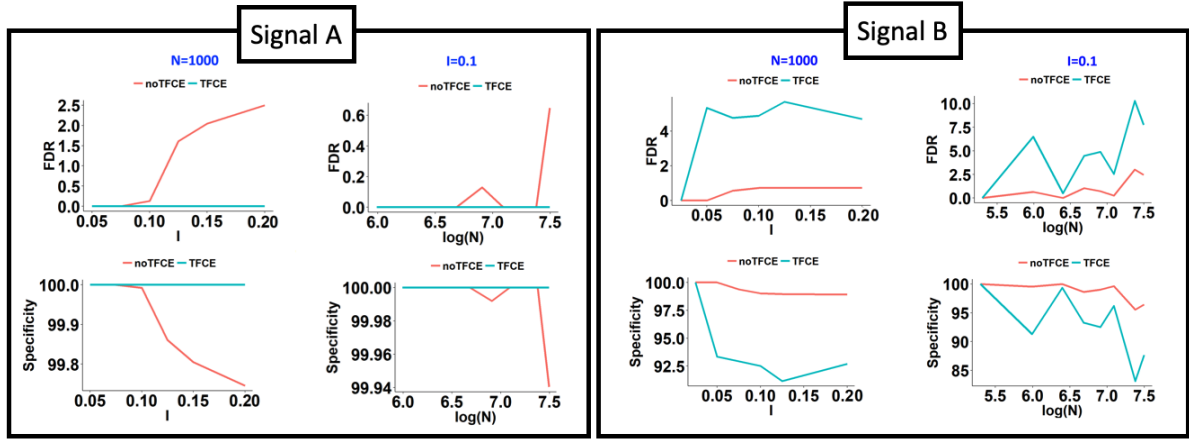
Figure 3.6: Rate of false discoveries and sensitivity of the proposed pipeline with or without TFCE on signal A and B.

GWAS replication study and the intensity of the associations found, these results would assign a sensitivity of 70% for the first two discovered SNPs and more than 90% for the other two. Figure 3.6 reports the model specificity and FDR for different I and N values. It can be noticed how the FDR was 0 for all the results of signal A and below 5% except for few simulations involving signal B and sample sizes greater than 1,600. This effect is due to the large synthetic signal extension, which causes TFCE to extend its support to vertices near the true signal which show the same direction of effect. Hence, this is not considered a major limitation as TFCE will not enhance clusters that originate only from noise.

3.5.2 Importance of TFCE

Figure 3.6 reports the specificity and FDR obtained by our pipeline with and without TFCE on signal A and B at different signal intensities (I) and sample sizes (N). Furthermore, Figure 3.7 reports the two maps obtained for the difference in sensitivity scored by the proposed pipeline using TFCE and without it. From the reported results it can be noticed how TFCE provides a relevant increase of up to 50% in sensitivity, which only comes at the expense of a small decrease in specificity on large extended signals. The increase of sensitivity provided by TFCE was higher for signal B due to its larger spatial
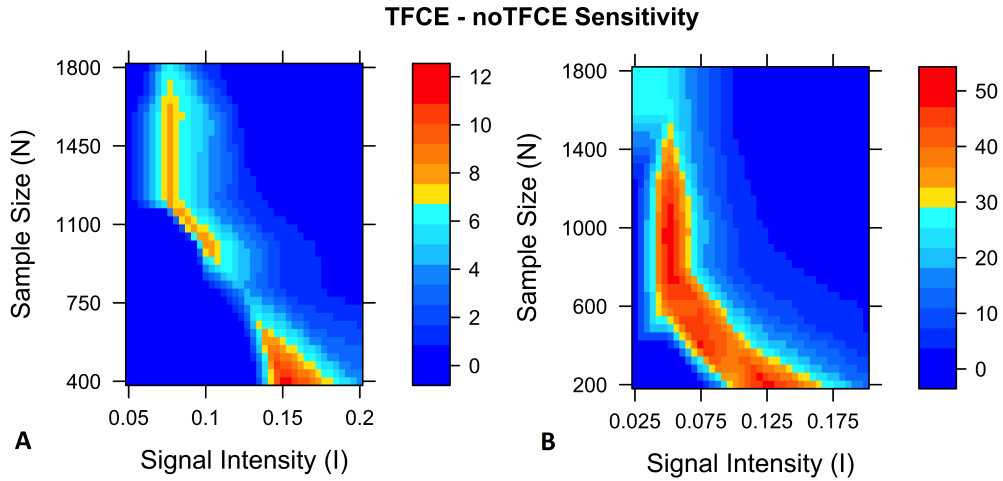
Figure 3.7: 2D maps showing the increase of sensitivity of the proposed pipeline when TFCE is applied on two different synthetic signals (A and B) at different signal intensities I and cohort dimensions (N).
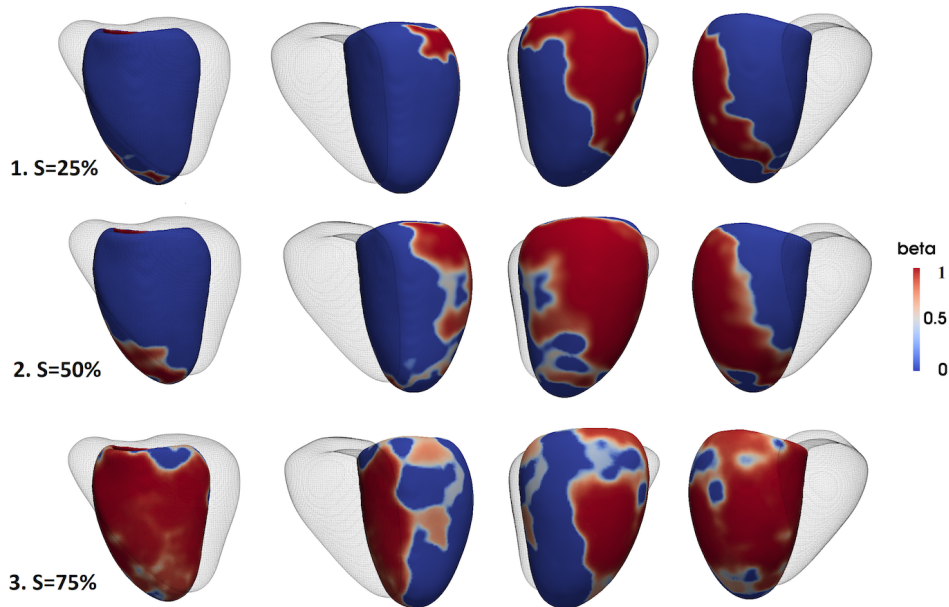
Figure 3.8: The three regression coefficient β maps (signal 1, 2 and 3) used to generate the synthetic signal to be detected the method proposed in this paper with different values of the TFCE parameters E and H.

extension as expected. The difference of sensitivities converged to zero at high I and N as also the sensitivity of the pipeline without TFCE reached 100% sensitivity. Overall, the application of TFCE provides a relevant increase in sensitivity which only comes at the expenses of a little decrease of specificity on largely extended signals.
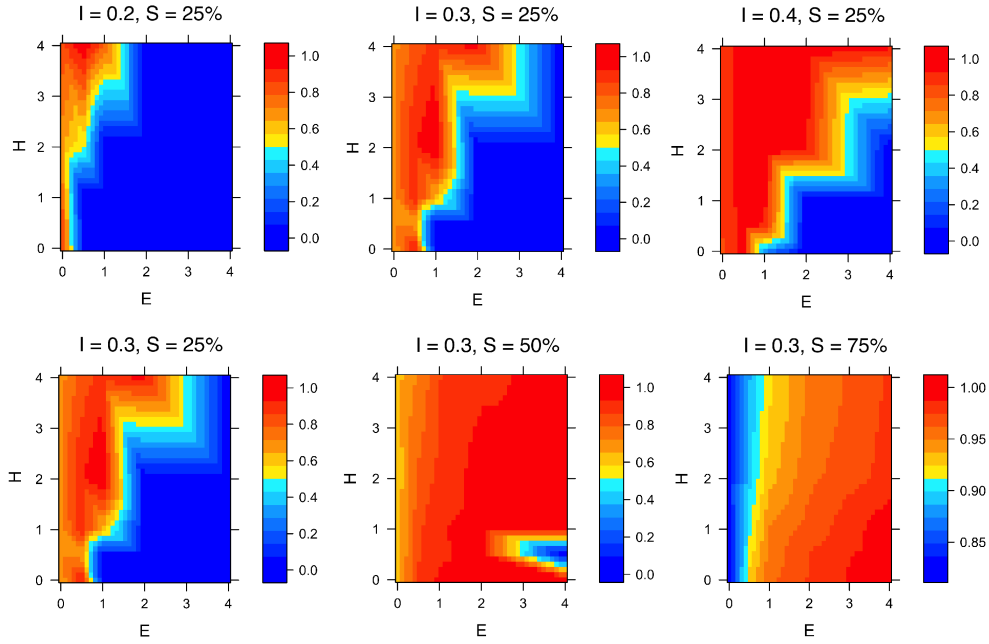
Figure 3.9: Colour plots for the proposed framework sensitivity at different TFCE parameters E and H and for different signal intensities I and signal extension S. In each graph sensitivity values were normalized to the maximum sensitivity detected.

3.5.3 Sensitivity to TFCE hyperparameters

The sensitivity of the proposed pipeline using different values of the TFCE parameters E and H was also assessed using synthetic data. Three new β maps (signal 1, 2 and 3) obtained from real clinical data and characterized by non-null β coefficients scaled to the (0,1] range were employed (Figure 3.8). These covered the 25%, 50% and 75% of total area of the LV respectively and were employed together with three distinct values (0.2, 0.3, 0.4) of signal intensity I. For a given value of the signal intensity I and of the spatial extension S of the synthetic signal, five values of the parameter E and five values of the parameter H were employed by the proposed framework to detect the synthetic signal generated (a total of 25 simulations for each (I, S) couple). The number of subjects N was fixed to 80, the number of permutations for each simulation to 5,000. Sensitivity results were linearly interpolated and normalized for the maximum sensitivity obtained for each (I,S) couple and plotted on the colour plots reported in Figure 3.9.

The first row of Figure 3.9 shows the sensitivity results obtained at a fixed spatial extension

of the generated synthetic signal ($S = 25\%$) and different signal intensities I. It can be noticed how at higher signal intensities the framework sensitivity increases and how on a small extended signal better sensitivity values are obtained when H is higher than E. The second row of Figure 3.9 shows the results obtained at a same signal intensity I when increasing the spatial extension S of the generated signal. In particular, in the bottom left figure it can be noticed how the importance of the signal intensity I is still predominant, while with the increase of the signal extension S the relative importance of the parameter E increases. In all the studied cases, the false discovery rate of the framework was always below the 5% and often equal to 0%, while the sensitivity was always above 99%.

Overall, in this preliminary study the values of $E = 0.5$ and $H = 2$ suggested in the TFCE original paper [166] by theoretical and empirical reasons achieved good sensitivity values. However, the performance of other combinations of E and H such as ($E = 1, H = 3$) show promise.

3.5.4   TFCE vs Standard Cluster-based Thresholding

A comparison between the proposed framework using TFCE or using a standard cluster-extend based thresholding was performed on the same synthetic data used in the previous section. The latter procedure as proposed in [267] has been implemented in the R package developed for this work. This procedure consists of two steps. In a first step, a cluster in the statistical map obtained by mass univariate regression is defined as the group of connected vertices that have a t-statistic value greater than a user-defined threshold h_{thr}. Then, a second threshold h_α is computed via permutation testing as the 95th percentile of the distribution of largest cluster in each permuted map, and used to declare significant the clusters in the original statistical map that are more spatially extended than this threshold h_α. Hence this method depends on the user-defined initial cluster-forming threshold h_{thr}. For this reason, the sensitivity, specificity and FDR of the proposed approach with TFCE parameters $E = 0.5$ and $H = 2$ were compared against the results obtained by the
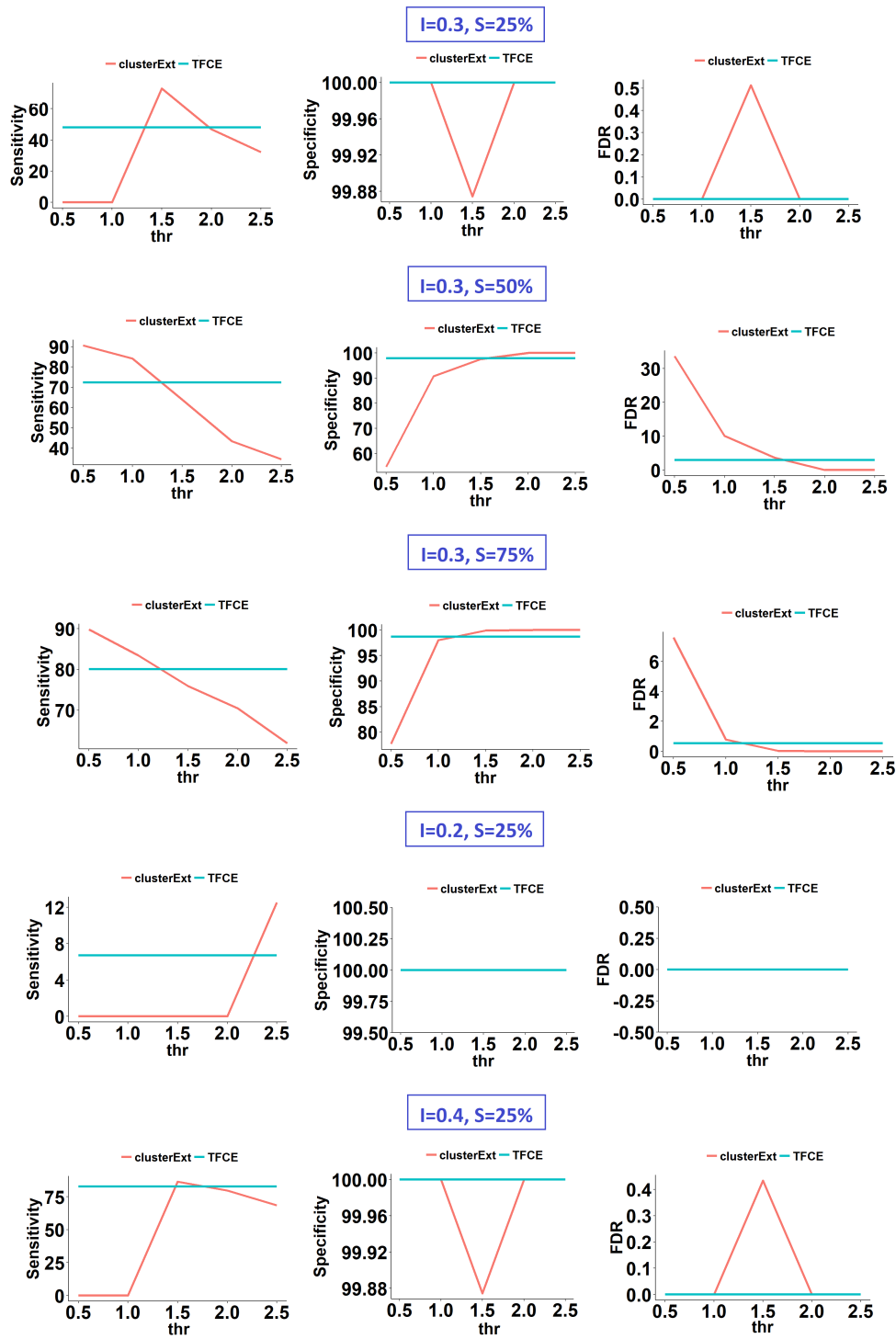
Figure 3.10: Sensitivity, specificity and the false discovery rate (FDR) of the proposed pipeline using either cluster-extent based thresholding or TFCE at different cluster-forming thresholds (thr). At the top of each set of graphs, the intensity I and the spatial extension S of the generated synthetic signal is reported.

same approach using cluster-extent based thresholding with five distinct cluster-forming thresholds h_{thr} $(0.5, 1, 1.5, 2, 2.5)$. The number of subjects N was again fixed to 80, the number of permutations for each simulation to 5,000 and the obtained plots are reported in Figure. 3.10

The sensitivity of the cluster-extent based thresholding method proved to be very dependent on the cluster-forming threshold h_{thr} and its choice had a large impact on the results. Moreover, higher FDR and lower specificity characterised cluster-extent based thresholding results when their sensitivity was comparable or greater than TFCE. These results therefore favour the use of TFCE over cluster-extent based thresholding as also proved in the brain image analysis literature [166].

3.6   Discussion

Despite recent advances in medical imaging and biobank medicine, the genetic and environmental determinants of cardiac physiology and function, especially in the earliest stages of disease, remain poorly characterized [58, 41]. Conventional morphological classification relies on one-dimensional metrics derived by manual image segmentation, largely insensitive to complex phenotypic variations [73, 57]. In contrast, computational cardiac analysis provides precise 3D quantification of shape and motion differences between disease groups and normal subjects [268, 85]. In this chapter, a general linear model framework I previously introduced for standard clinical variables ([252]) was extensively validated on a more challenging imaging-genetics scenario. The approach proved to be capable of providing a powerful approach for modelling the relationship between phenotypic traits and genetic variation using high-fidelity 3D representations of the heart. By translating statistical parametric mapping techniques originally developed for brain mapping to the cardiovascular domain, the proposed methodology exploits spatial dependencies in the data to identify coherent areas of biological effect in the myocardium which would be

otherwise neglected by standard volumetric analysis. Moreover, it also accounts for multiple testing correction at tens of thousands of vertices which is the main drawback of this class of techniques. In particular, the application of TFCE leads to a notable increase in power of the mass univariate approach at the expense of only a slight increase of the false discovery rate in large extended signals.

Genetic association studies using conventional 2D imaging leave much of the moderate heritability of LV mass unexplained [71, 159, 269]. One contribution may be the lack of phenotyping power of conventional imaging metrics, which require manual analysis and are insensitive to regional patterns of hypertrophy [73, 69]. Our simulations on synthetic data show that our approach has the power to detect anatomical regions associated with even small genetic effect sizes. In the reported exemplar application, we replicated the effect of four SNPs discovered in a GWAS for LV mass using a 3D WT phenotype with TFCE applied, while none of the SNPs replicated with conventional LV mass analysis. The genotype-phenotype associations that we report reflect that cardiac geometry is a complex phenotype with a highly polygenic architecture dependent on anatomical patterns of gene expression and spatially-varying adaptations to haemodynamic conditions [270, 271].

One of the main limitations of the presented framework is that high-spatial resolution CMR is not available in all cohorts, although conventional two-dimensional images may be super-resolved to provide similar shape models [215]. A second limitation is that the true association may not be linear in the model parameters and nonlinear models could better fit the data. However, the advantages favouring a general linear model are its simplicity, the ability to easily design and adjust the results for multiple factors and its wide use in biomedical statistics. A third limitation of this work is with regards to the experiments using synthetic data as we only assessed noise in our single centre population and did not generalise this to other cohorts. A general limitation of these approaches is that they do not establish causal relationships, such as the interaction between genetic variants, blood pressure and LV mass, although this may be addressed in future work by

Mendelian randomisation.

In the neuroimaging literature, in the context of brain-wide candidate-SNP analyses, mass univariate approaches are used more extensively than multivariate approaches for a two-fold motivation [88]. Firstly, multivariate methods have been found so far less sensitive to regional association. Secondly, they typically require more observations than the dimension of the response variable, i.e. number of vertices in this case, or the use of dimensionality reduction techniques to summarize the phenotype and which might discard anatomical information of interest. On the other hand, mass univariate approaches do not directly consider the local covariance structure of the data as their multivariate counterparts. However, this is accounted for when Random Field Theory or permutation tests define a threshold for significant activation [272]. For these reasons, the development of the proposed mass univariate approach over multivariate methods has been prioritized.

As the methods are computationally-efficient and require no human input for phenotypic analysis, it is feasible to scale up the pipeline to larger population cohorts such as UK Biobank, which aims to investigate up to 100,000 participants using MR imaging [40]. In the last years, the proposed method and R package already helped to reveal the effect of rare variants on LV geometry in participants without overt cardiomyopathy [273], and to detect specific regional remodeling patterns in aortic stenosis patients [274] and during pulmonary hypertension [109]. All the discovered associations were not detectable by global measures of remodelling, thus showing promise for the discovery of new insight on determinants of healthy and pathological cardiac anatomy in other large-scale imaging studies, such as the above mentioned UK Biobank. In order to characterize genetic susceptibility of cardiac phenotypes in greater depth, vertex-wise genome-wide analyses also represent an interesting direction of future methodological work. In this latter context, multivariate approaches may show promise for modelling high-dimensional imaging and genetic data [150, 275], but specific approaches for the cardiac domain are yet to be proposed. Finally, while we have focused on LV geometry and shape, the same approach

can be applied to time-resolved vertex-wise data to create a functional phenotype for regression modelling.

3.7   Conclusion

In this chapter, a powerful and flexible framework for statistical parametric modelling of 3D cardiac atlases, encoding multiple phenotypic traits, and imaging-genetic data is presented. The approach has been validated the approach on both synthetic and genetic datasets, showing its suitability for detecting genotype-phenotype interactions on LV geometry. More generally, the proposed method can be applied to population-based studies to increase our understanding of the physiological, genetic and environmental effects on cardiac structure and function.

Chapter 4

Learning Interpretable Discriminative Features via Deep Generative Models

This chapter is based on: Biffi, Carlo, *et al.* "Learning interpretable anatomical features through deep generative models: Application to cardiac remodeling." In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 464-471. Springer, Cham, 2018.

4.1 Introduction

Alterations in the geometry and function of the heart (cardiac remodeling) are commonly employed as criteria to diagnose and classify cardiovascular diseases as well as risk-stratify individual patients [34, 42]. For instance, diagnosis of hypertrophic cardiomyopathy (HCM) involves the detection of unexplained left ventricular (LV) hypertrophy patterns, often affecting distinct LV segments and not easily quantifiable [276]. The gold-

standard imaging technique to assess structural shape changes of the heart is magnetic resonance (MR) which enables imaging at high-resolution and in three-dimensions (3D) [17]. Despite the advances in non-invasive imaging, including MR imaging, classification and risk-stratification of cardiac disease patients still rely on the combination of visual interpretation of medical images by a cardiologist, and on the assessment of scalar indexes describing pathological remodeling. However, these indices (LV mass or ejection fraction for instance) have low diagnostic and prognostic value due to their limitations in representing regional or asymmetric effects that occur during pathology [58, 59, 82]. Subjective visual interpretation by a cardiologist often results in considerable observer-dependencies, despite years of training [76].

Due to the recent advancements in machine learning and database technologies, automated data-driven diagnosis approaches able of obtaining and exploiting a deeper quantification of cardiac phenotypes have been increasingly investigated [60, 119, 277]. These approaches aim to fully exploit the diagnostic value of cardiovascular imaging and to provide novel computer-aided diagnosis tools to improve clinician decision-making, promising to improve early detection, risk stratification and therapy of cardiac diseases [76, 17]. Among all the available machine learning methodologies, deep learning approaches seem particularly suited for this task due to their unmatched feature extraction capabilities. However, such models are typically regarded as black-box models due to their difficulties at providing explanations on how and why a certain decision has been made, hence limiting their value in clinical diagnosis [278, 191].

4.1.1 Contributions

In this work, a modification of VAE framework [220], based on 3D convolutional layers, which is employed for classification of cardiac diseases associated with structural remodeling is proposed. The proposed model learns a small set of generative shape features which are concurrently optimized for cardiac shape classification in an end-to-end fash-

ion. This latter property is achieved by employing such features as input of a classification network. As a consequence, the anatomical effect encoded by the learned features can be easily visualised in the original space of the segmentations, making the classification task transparent. Moreover, a novel data-driven method to enable the easy visualisation and quantification of the learned remodeling is proposed. The performance of the proposed approach is evaluated for the classification of healthy volunteers (HVols) and HCM subjects on our own multi-centre cohort (Imperial College dataset) and on the ACDC MICCAI 2017 challenge dataset.

4.1.2 Related Work

Learning-based approaches that can capture complex phenotypic variation and exploit it for data-driven decision making, such as image or shape classification, have been widely studied in the last decade [76, 75, 148]. Of particular interest for this work, recent research has shown the potential of machine learning algorithms in distinguishing benign from pathological hypertrophy from multiple manually-derived cardiac parameters [279, 50] or medical images [280, 281]. However, these methods lack interpretability in feature extraction and decision processes, making it hard to infer which feature best contributes to the outputs and what anatomical information it expresses [278]. In the work of Zheng *et al.* [282], myocardial motion is end-to-end learned to obtain easy-to-interpret features describing cardiac motion. These features are then combined with conventional imaging metrics as input of simple cascaded binary classifiers. The simplicity and the small number of used features, makes the classifier explainable. However, the employed features are not task-specific, as the model proposed in this work aims to obtain, but hand-crafted before classifier training. In the brain imaging domain, Shakeri et al. [229] employed a variational autoencoder (VAE) model based on two fully connected layers to learn a low-dimensional representation of co-registered hippocampal meshes, which is later employed in a multi-layer perceptron (MLP) to classify patients with Alzheimer disease.

In contrast, in this work a deep 3D convolutional neural network architecture is applied directly on segmentation maps to learn a discriminative latent space in an end-to-end fashion. Furthermore, a data-driven method, within the VAE framework, to visualise the anatomical features employed in the classification process is also proposed.

4.2 Material and Methods

4.2.1 Cardiac Datasets

A multi-centre cohort consisting of 686 HCMs patients and 679 healthy volunteers was considered for this work. All subjects underwent cardiac phenotyping at a 1.5-T on Siemens (Erlangen, Germany) or Philips (Best, Netherlands) system using a standard cardiac MR protocol. Cine images were acquired with a balanced steady-state free-precession sequence and included a stack of images in the left ventricular short axis plane (voxel size 2.1x1.3x7 mm^3, repetition time/echo time of 3.2/1.6 ms, and flip angle of 60°). End-diastolic (ED) and end-systolic (ES) phases were segmented using a previously published and extensively validated cardiac multi-atlas segmentation framework [96]. HCM patients were confirmed with reference to established diagnostic criteria [17].

As a first preprocessing step, the quality of the 2D stacks segmentation by a multi-atlas-aided upsampling scheme was improved. For each segmentation, twenty manually-annotated high-resolution atlases at ED and ES were warped to its space using a landmark-based rigid registration. Then a free-form non-rigid registration with a sparse set of control points was applied (nearest-neighbor interpolation) [283] and fused with a majority voting consensus, leading to an upsampled high-resolution segmentation ($2mm$ x $2mm$ x $2mm$). In a second step, all the quality-enhanced segmentations were aligned onto the same reference space at ED by means of landmark-based and subsequent intensity-based rigid registration to remove pose variations. After extracting the LV myocardium label,

	HCM		Hvol	
	Mean	SD	Mean	SD
Age at recruitment / first CMR	54.87	15.97	37.51	12.94
Females (%)	27.1		36.8	
BSA (m^3)	1.91	0.23	1.80	0.19
Left ventricular end-diastolic volume (mm^3)	134.77	36.35	141.77	30.52
Left ventricular end-systolic volume (mm^3)	35.28	17.43	48.53	14.51
Left ventricular ejection fraction (%)	74.45	9.74	66.12	5.04
Left ventricular mass (g)	182.00	64.84	109.84	32.29
Max wall thickness (mm)	18.33	4.85	7.37	3.52

Table 4.1: Table of population characteristics of the Imperial College dataset. Information for 34 HCMs patients were not available. The total number of healthy volunters (HVols) subjects is 451, total number of HCMs is 402.

each segmentation was cropped and padded to $[x = 80, y = 80, z = 80, t = 1]$ dimension using a bounding box centered at the LV's ED myocardium. The latter operation guarantees shapes to maintain their alignment after cropping. Finally, all segmentations underwent manual quality control in order to discard scans with strong inter-slice motion or insufficient LV coverage, resulting in 436 HCM patients and 451 healthy volunteers that were used for the final analysis. This dataset is referred as Imperial College dataset and its population characteristics and standard CMR metrics are reported in Table 4.1). As an additional external testing dataset, ED and ES segmentations from 20 healthy volunteers and 20 HCMs from the ACDC MICCAI'17 challenge training dataset [60] were also used (after undergoing pre-processing using the same high-resolution upsampling pipeline explained above).

4.2.2   Deep Generative Model

A schematic representation of the proposed architecture is shown in Fig. 4.1. At training, the input of the proposed architecture is constituted of a set of N anatomical segmentations $X = \{\boldsymbol{x}_j, \ j = 1, ...N\}$ of a structure of interest from a population S and of their disease class label $Y = \{\boldsymbol{y}_j, \ j = 1, ...N\}$. In this paper, $\boldsymbol{x}_j$ is a pair of 3D LV myocardial segmentations at ED and ES phases and presented as a two-channel input. Our proposed
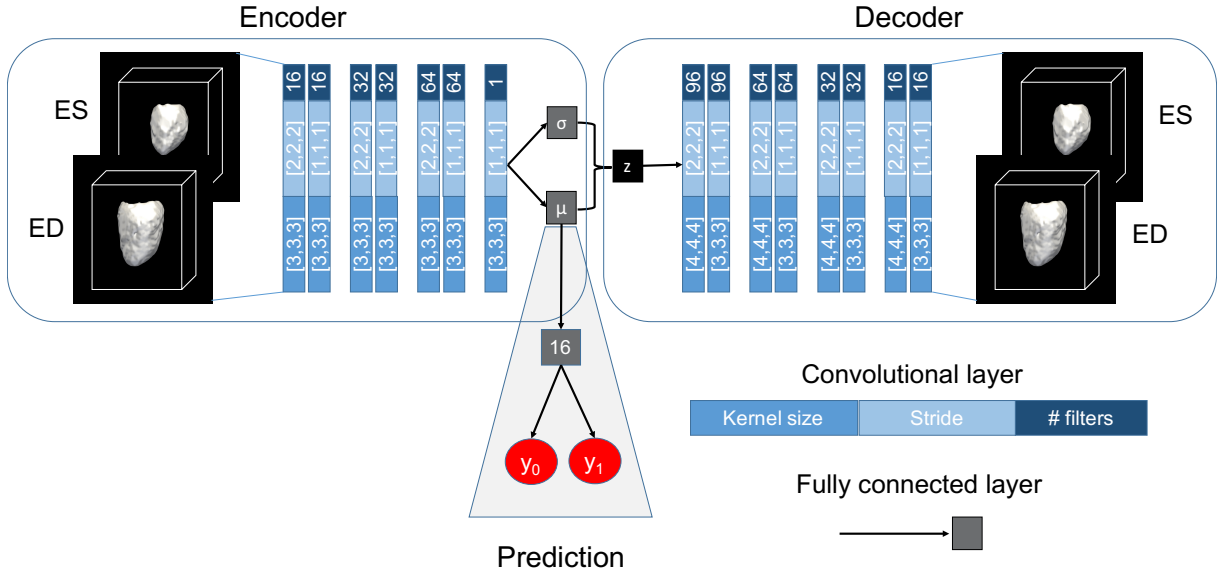
Figure 4.1: Deep generative model architecture. Registered LV segmentations at ED and ES phases are mapped to a low-dimensional latent space. Each latent dimension is forced to be normally distributed with mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$. A decoder network is then used to reconstruct the input segmentation from a low-dimensional vector $\boldsymbol{z}$ sampled from the learned latent distribution (training) or the $\boldsymbol{\mu}$ vector (testing). The $\boldsymbol{\mu}$ latent representation is used as input of a MLP to predict disease status.

architecture consist of a VAE [220], which learns a latent representation $\boldsymbol{z}_j \in \mathcal{R}^p$, where $p \ll d$ and d is the number of pixels/voxels in a segmentation $\boldsymbol{x}_j \in S$, and which is trained concurrently with a shallow multilayer perceptron (MLP) learning $p(\boldsymbol{y}_j|\boldsymbol{z}_j)$. As reported in Section 2.3.4, the VAE learns the distribution $p_\theta(\boldsymbol{x})$ of the population of segmentations $\boldsymbol{x} \in S$ by learning from the training data distribution $p_\theta(\boldsymbol{x}|\boldsymbol{z})$, named decoder network, and $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, encoder network, via the optimization of

$$\mathcal{L}_{\theta,\phi}(\boldsymbol{x}_j) = \underset{q_\phi(\boldsymbol{z}|\boldsymbol{x}_j)}{\mathbb{E}} \left[\log \left[p_\theta(\boldsymbol{x}_j|\boldsymbol{z}) \right] - \mathcal{D}_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}_j)||p_\theta(\boldsymbol{z})) = \mathcal{L}_{rec} + \mathcal{L}_{KL} \quad (4.1)$$

The form of $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2\boldsymbol{I})$ is assumed to be a multivariate Gaussian with a diagonal covariance structure and $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the outputs of the encoder network. Given an input segmentation $\boldsymbol{x}_j$, its disease class label, HVol or HCM in this chapter, is computed by the prediction network using $\boldsymbol{\mu}_j$ as input, i.e. $\boldsymbol{y}_j = p(\boldsymbol{y}_j|\boldsymbol{z}_j = \boldsymbol{\mu}_j)$. With

the addition of a cross-entropy loss term for the training of a prediction network $\mathcal{L}_{MLP}$, the overall loss function becomes $\mathcal{L} = \mathcal{L}_{rec} + \alpha\mathcal{L}_{KL} + \beta\mathcal{L}_{MLP}$. $\mathcal{L}_{rec}$ is a reconstruction loss and it is implemented as a Sorensen-Dice loss between the input segmentations $\mathbf{x_j}$ and their reconstruction $\hat{\mathbf{x}}_{\mathbf{j}}$. $\mathcal{L}_{KL}$ is the Kullback-Leibler divergence loss forcing $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ to be as close as possible to its prior distribution $\mathcal{N}(\mathbf{0}, \mathbf{1})$. The latent space dimension was fixed to $d = 64$. At test time, each input segmentation is reconstructed by passing the predicted $\boldsymbol{\mu}$ to $\mathbf{z}$ (without sampling from the latent space), while the classification is performed as in training time.

4.2.3   Interpreting Learned Features via Navigation in the Latent Space

Our generative model architecture allows for the visualization of the features learned by the network in the original segmentation space. For this purpose, the weights learned by the MLP can be exploited to compute the partial derivative of the disease class label C (y_C) w.r.t. to the latent space representation $\boldsymbol{\mu}_j$ of an input $\mathbf{x_j}$, i.e. $\frac{\partial y_C}{\partial \boldsymbol{\mu}_j}$, by backpropagating the gradient from the class label C to $\boldsymbol{\mu}_j$ using chain-rule. Given a randomly selected healthy shape, the derived gradient can be used to change the latent representation of a subject $\boldsymbol{\mu}$. In particular, using an iterative algorithm, this can be changed along the direction of the latent code variability that maximises the probability of $\boldsymbol{\mu}$ being classified to class C. Starting with the mean latent representation $\boldsymbol{\mu}_0 = \bar{\boldsymbol{\mu}}_j$ of a healthy shape, $\boldsymbol{\mu}_j$ can be iteratively updated at each step t accordingly to Eq. 1:

$$\boldsymbol{\mu}_j^t = \boldsymbol{\mu}_j^{t-1} + \lambda \, \frac{\partial y_C}{\partial \boldsymbol{\mu}_j^{t-1}} \qquad (4.2)$$

Here $\lambda = 0.1$. Finally, each latent representation $\boldsymbol{\mu}_j^t$ at each step t can be decoded back to the segmentation space by passing it to $\mathbf{z_j}$. This enables the visualization of the corresponding reconstructed segmentation $\hat{\mathbf{x}}_{\mathbf{j}}$ on which all the clinical indices currently

employed in the clinic, such as LVM or LVV, can be now also computed.

4.3   Results

The Imperial College Dataset was split into training, evaluation and testing sets consisting of 537 (276 HVols, 261 HCMs), 150 (75 HVols, 75 HCMs) and 200 (100 HVols, 100 HCMs) subjects respectively. The model was developed in Tensorflow, and trained on a Nvidia Tesla K80 GPU using Adam Optimizer, learning rate of 10^{-4} and batch size of 16. After 96k iterations, the total validation loss function stopped improving and the training was stopped. No significant changes in the classification results were found by varying the loss parameters α and β, while α was set to 0.1 as this captured local shape variations without losing the generative model properties. All the 200 subjects in the testing dataset were correctly classified (100% accuracy) by the trained prediction network. The model also correctly classified 36 out of the 40 ACDC MICCAI 2017 segmentations (90% accuracy); of the 4 misclassified cases, 3 did not properly cover the whole LV, which might have been the cause for the error.

By employing the proposed method for latent space navigation, a randomly selected healthy segmentation was deformed from the training set towards the direction that maximizes its probability of being classified as HCM. On the right of Fig. 4.2, the original segmentations of the selected subject at ED and ES phases is reported, their reconstruction from the VAE, and the reconstructed segmentations at four different iterations of the latent space navigation method. On the left of Fig. 4.2, the latent 64-dimensional representation $\boldsymbol{\mu}$ of the training set segmentations together with the latent representations $\boldsymbol{\mu}_t$ obtained at each iteration t were reduced for visualization purposes to a bi-dimensional space using Laplacian Eigenmaps [284]. This technique allows to build a neighborhood graph of the latent representations that can be used to monitor the transformation (light blue points) of the segmentation under study from the HVol cluster to the HCM cluster.
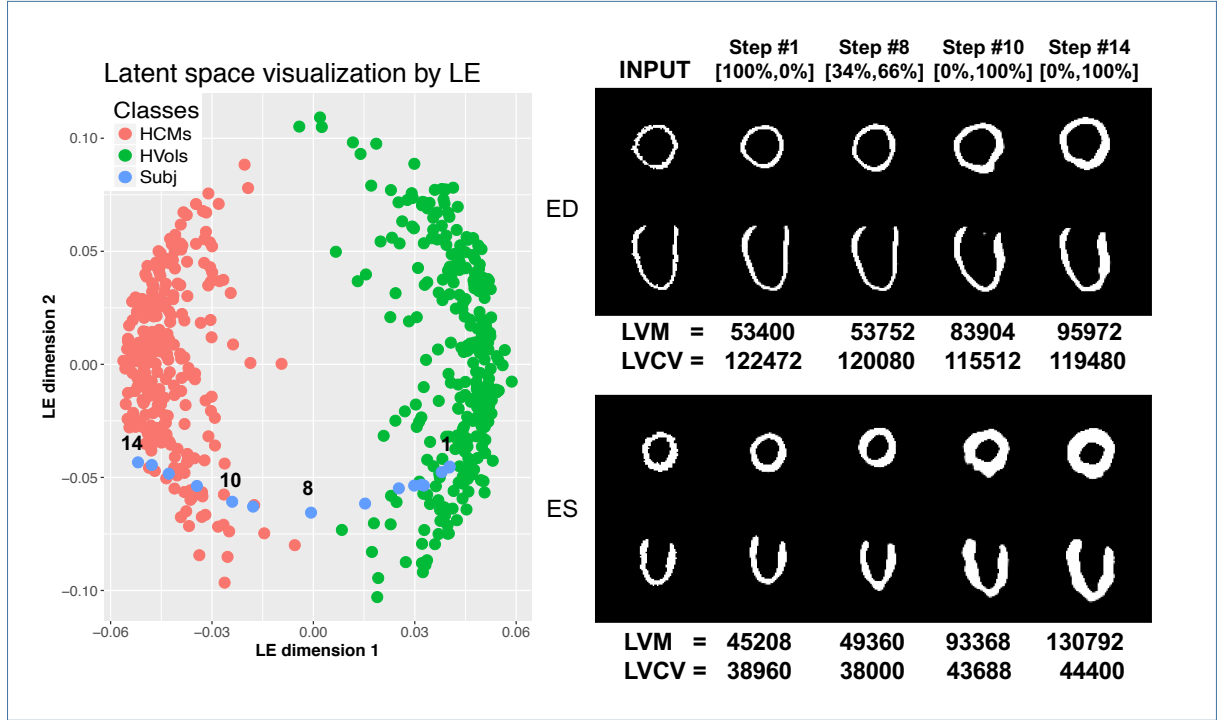
Figure 4.2: On the left, Laplacian Eigenmaps (LE) bi-dimensional representation of the latent $\boldsymbol{\mu}$ of each subject in the training set (red and green dots) and of the $\boldsymbol{\mu}_j^t$ obtained through latent space navigation (light blue dots) for a random healthy shape. This latter is displayed on the right, together with the decoded segmentations corresponding to the sampled $\boldsymbol{\mu}_j^t$ reported on the left at 4 exemplary iterations. The probabilities of class HVols and HCM, and the computed LVM and LVCV are also shown.

At each reported step, LV mass (LVM) from each segmentation was derived by computing the volume of the myocardial wall using the number of voxels labeled as LV myocardium. Moreover, a LV atlas segmentation having also labels for the LV cavity was non-rigidly registered to each segmentation to compute LV cavity volume (LVCV) by summing the number of the voxels labeled as LV blood pool. Finally, for each iteration the probabilities of being an HVol or HCM are also reported as computed by the prediction network. The learned deformations demonstrate a higher LVM and lower LVCV with an asymmetric increase in septal wall thickness in the geometric transition from HVol to HCM - which is the typical pattern of remodeling in this disease [285]. At iteration 8, where the prediction network gives an indeterminate classification probability, LV geometry appears normal at ED but is thickened at ES suggesting that altered contractility may also be a discriminative feature.

4.4 Discussion

In this chapter, a deep generative model for automatic classification of heart conditions associated with cardiac remodeling is presented. The proposed model leverages task-specific latent variables learned directly from 3D segmentations, which are made explainable by design. In particular, the proposed architecture is specifically designed to enable the visualization and quantification of the anatomical effect encoded by the latent variables in the original segmentation space, making the classification decision process interpretable and potentially enabling quantification of disease severity. In this work, a simple but effective method that enables the navigation of the low-dimensional latent manifold learned by the network is also proposed. In the reported exemplar clinical application, the learned features achieved high accuracy in the discrimination of healthy subjects from HCM patients on our unseen testing dataset and on the ACDC MICCAI 17 dataset. Moreover, the reported results show the potential clinical utility of the derived latent representation for tracking and scoring patients against a reference population.

The methodologies presented can be easily extended to other cardiac related clinical tasks by replacing the prediction network with survival, risk score or other clinical models implemented as neural networks. For instance, in later work from our group [286] a supervised denoising autoencoder was used to learn a latent code representation of right ventricular contraction patterns and to concurrently perform survival prediction in pulmonary hypertension patients cohort. Clough *et al.* [287] also extended this approach to use the whole cardiac cycle rather than just the ED and ES frames and employed the concept activation vector method to associate diagnostically meaningful clinical biomarkers to some latent space variables, further increasing the intrepretability of the model. In [288], the authors employ a deep network to estimate conventional cardiac biomarkers used to estimated cardiac function. Then these biomarkers are used as input of a VAE-based model that learns to regress systolic blood pressure (SBP) from the model latent space variables as proposed in this chapter. In this way, by monitoring the changes in the decoded values

of the studied clinical biomarkers when SBP changes, a data-driven description of the relation between cardiac function can SBP is found.

The proposed approach worked successfully on conventional MR acquisitions, showing its potential for using routinely acquired clinical MR imaging for end-to-end automatic diagnosis. Moreover, the by only employing segmentation masks, the proposed model could allow its application to a wider range of CMR images such as multi-site images acquired from different machines and using different imaging protocols. Finally, the external testing dataset was small, future work will focus on evaluating the proposed approach on a bigger unseen dataset from different centres and on various types of cardiomypathies.

4.5   Conclusion

In this work, this chapter demonstrates how a deep generative model can be employed to learn discriminative and generative features for classification of cardiac diseases associated with structural remodeling. Due to the double nature of such features, the anatomical effect they encode can be easily visualised in the original space of the segmentations, making the classification task transparent. The proposed approach is a promising step towards the development of interpretable deep learning classifiers for the medical imaging domain, which may assist clinicians to improve diagnosis and provide new insight into patient stratification. This general approach is not limited to the cardiac domain and can potentially be extended to other image analysis tasks where pathological shape change is prognostically relevant.

Chapter 5

Explainable Anatomical Shape Analysis through Deep Hierarchical Generative Models

This chapter is based on: Biffi, Carlo, *et al.* "Explainable Anatomical Shape Analysis through Deep Hierarchical Generative Models." *IEEE Transactions on Medical Imaging* (2020).

5.1   Introduction

The quantification of anatomical changes and their relationship with disease is a fundamental task in medical image analysis, ultimately leading to new clinical insights and enhanced risk assessment and treatment. Recent improvements in the medical image analysis field have been characterised by an increase of large-scale population-based initiatives [55, 289, 290, 35], such as the UK Biobank study which aims at performing cardiac imaging in up to 100,000 participants [290]. This, together with the development

of automated segmentation pipelines of anatomical structures [291, 292], which recently achieved human-level performance [31, 74], provides a means for unprecedented deep phenotyping of populations. However, alterations in shape and structure of an organ associated with many conditions (pathological remodeling) remain poorly characterised, as typically evaluated with crude global or hand-crafted metrics. Examples include diagnosis of hypertrophic cardiomyopathy (HCM) or Alzheimer's disease (AD), where pathological remodeling is proved to affect distinct local areas of the myocardial or hippocampal shape, but it is often assessed using scalar indices [276, 293, 294]. Therefore, the development of novel data-driven processing tools to enable quantitative assessment of the differences between normal anatomy and pathology has now received significant interest [190, 277, 294].

5.1.1   Related Work

Deep learning methods proved to be powerful features extractors for the classification of clinical conditions from medical images [74, 60]. Despite their tremendous success, a major drawback of deep learning models is their lack of interpretability, which currently hampers their translation to clinical practice. In fact, the physiological reason that drives the classification result is often as important as the classification result itself, making explainability an increasingly requested property for machine learning algorithms [60, 237]. In hippocampus shape analysis, Shakeri *et al.* [229] employed a VAE model to learn a low-dimensional representation of co-registered hippocampus meshes, which was employed in conjuction with a multi-layered perceptron (MLP) to classify healthy subjects from AD patients. The network input consisted of mesh vertices coordinates, and the representation was learned through two fully connected layers. Similarly, in the previous Chapter 4, a modification of the 3D convolutional VAE framework is proposed to learn a low-dimensional latent representation of 3D LV segmentations, which was not only able to encode the 3D segmentations manifold, but also to discriminate different conditions by

performing the classification task in the latent space. In the same work, a latent space navigation method to explore the anatomical variability encoded by the learned latent space was proposed. This consisted in iteratively modifying the latent representation of a segmentation obtained from an healthy subject along the direction that maximized its probability to be classified as pathological. By decoding the different latent representations in the original space of the segmentations, the proposed technique allowed the visualisation of the anatomical changes caused by this transformation. However, the following limitations characterize the work presented in the previous chapter: 1) The learned VAE latent space not only encoded the factors of variation that most discriminate between classes, but also all the other factors of variation that regulate shape appearance. The latent space navigation was thus a necessary step to attempt the offline estimation of the variations linked to the pathological remodeling. This work aims at automatically learning a latent space that encodes only these changes. 2) The previous model required an additional offline dimensionality reduction technique to visualize in two dimensions the clustering obtained in the VAE latent space, which would however not reflect the real distribution of the shapes in the learned latent space. The model proposed in this work aims at directly learning this two-dimensional latent space. 3) The latent space navigation method proposed in the previous model could only obtain subject-specific paths (with no obvious navigation stopping criteria). The model proposed in this work aims at providing a means to extract the more clinically appealing population-based inferences.

In the later work from our group [286], a supervised denoising autoencoder was used to learn a latent code representation of right ventricular contraction patterns and, at the same time, to perform survival prediction. Not being a generative model, the effect of task-specific features learned by the proposed model could not be visualised, making the prediction task not explainable and population based inferences difficult to obtain. In addition, an additional offline dimensionality reduction step was also required to visualise in two-dimensions the distribution of different groups of subjects.

Hierarchical VAEs are a class of generative models that decompose the input data into a hierarchical representation [295], [296]. Although highly flexible, these models have been traditionally difficult to optimise, especially in the training of their higher levels, as often their lowest layer alone can contain enough information to reconstruct the data distribution, and the other levels are ignored. This chapter focuses on the ladder VAE (LVAE) framework [296], which was shown to be capable of learning a deeper and more distributed latent representation by combining the approximate likelihood and the data-driven prior latent distribution at each level of the generative model.

5.1.2   Contributions

I aim to extend the work presented in the previous Chapter 4 with the aim of assisting clinicians in quantifying the morphological changes related to disease and to introduce a deep learning system for explainable anatomical shape analysis. The main contributions of this work can be described as follows:

- It is demonstrated that an interpretable classifier of anatomical shapes can be developed by performing a classification task of interest in the highest level of a LVAE model. In this way, the latent variables of this level automatically encode the most discriminative features for the task under exam, while the other subsequent levels model the remaining factors of anatomical variation in the data.

- It is shown that the LVAE highest latent space can be assumed to be two- or three-dimensional so that the classification space can be directly visualised without further offline dimensionality reduction steps. Furthermore, it is shown how the anatomical variability encoded by this latent space can be visualised in the original space of the segmentations thanks to the generative properties of the model, enabling the visualisation of the anatomical effect of the most discriminative features between different conditions.

- The proposed LVAE-based method achieves high classification accuracy of HCM versus healthy 3D LV segmentations and of AD versus healthy controls 3D hippocampi segmentations. More importantly, it is shown how the proposed model captures and enables the easy visualisation of the most discriminative features between the conditions under exam. Finally, the reported experiments the learned hierarchical representations provide higher reconstruction accuracy compared to single-latent-space VAEs.

- While hierarchical VAEs have been mainly evaluated on benchmark datasets, in this chapter they are successfully applied to two real-world 3D medical imaging datasets. Insights on the model functioning and optimal training are showed, and the implementation of proposed method is made publicly available[1].

5.2 Methods

This section is organised as follows. First, in subsections A and B, the theoretical foundations of the proposed method are summarised. Second, in subsection C the modifications to the original VAE and LVAE frameworks towards explainable shape analysis are reported (graphical models in Fig. 5.1). Then, in subsection D, the datasets used in this work for the classification of healthy subjects versus HCM patients and of healthy controls versus AD patients are described in detail. Finally, in subsection E, a detailed description of the LVAE models used in this work is reported (model summary in Fig. 5.2 for the cardiac application, and in Fig. 5.3 for the brain application).

---

[1]https://github.com/UK-Digital-Heart-Project/lvae_mlp
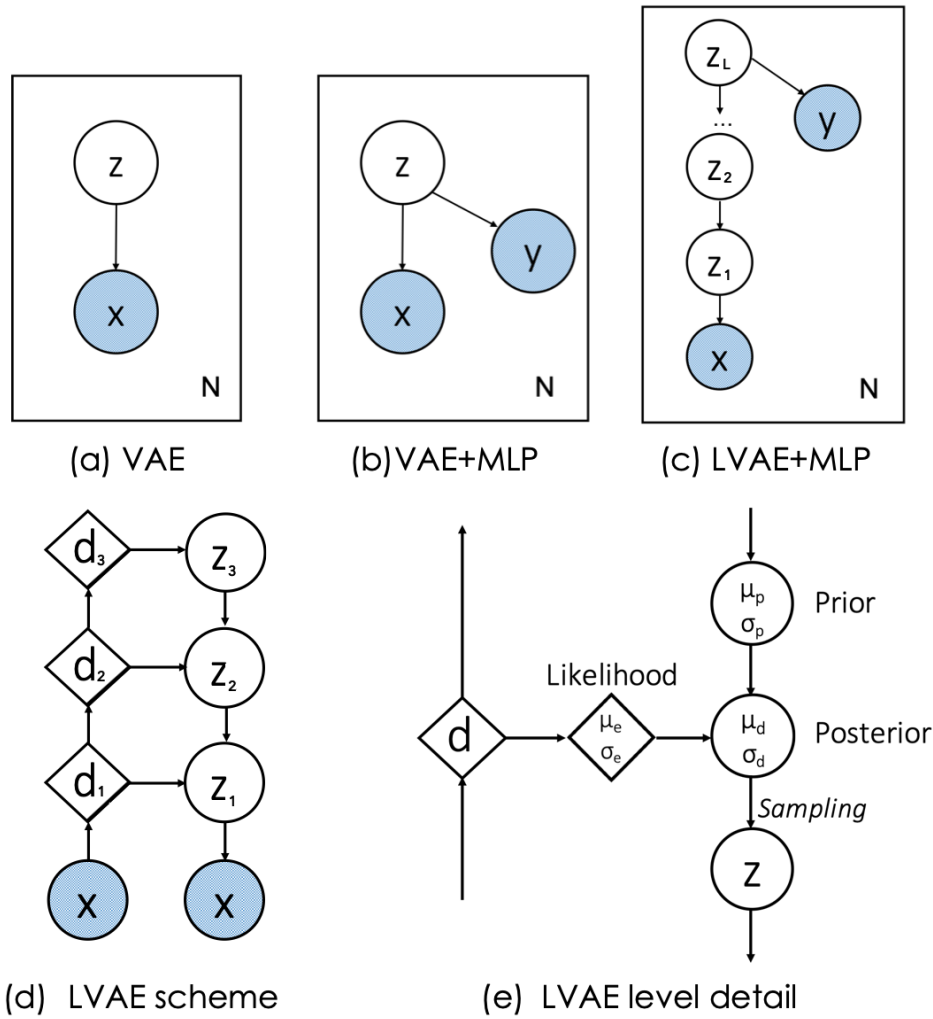
Figure 5.1: Graphical models of a standard VAE (a), the previously proposed method (Chapter 4) (b) and the new LVAE-based approach (c). x represents and anatomical segmentation, y the disease class label and z the latent variables to learn. Schematic representation of a three-level LVAE (d) and of the flow of information (e). Circles represent stochastic variables, diamonds represent deterministic variables. Variables in light blue represent the inputs of the network.

5.2.1   Ladder Variational Autoencoder (LVAE)

Given a training set of N anatomical segmentations $X = \{\boldsymbol{x}_j, \ j = 1, ...N\}$ of a structure of interest from a population S, a VAE [220] aims at learning the distribution $p_\theta(\boldsymbol{x})$ of the population of segmentations $\boldsymbol{x}_j \in S$ under study. As seen in Section 2.3.4, this is achieved by learning from the training data distribution $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ (decoder network) using a model of latent variables $\boldsymbol{z} \in \mathcal{R}^p$, where $p \ll d$ and d is the number of pixels/voxels

in a segmentation $\boldsymbol{x} \in S$, and a variational distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ (encoder network). The VAE graphical model is depicted in Figure 5.1 (a) and it is trained by optimizing:

$$\mathcal{L}_{\theta,\phi}(\boldsymbol{x}_j) = \underset{q_\phi(\boldsymbol{z}|\boldsymbol{x}_j)}{\mathbb{E}} \left[\log \left[p_\theta(\boldsymbol{x}_j|\boldsymbol{z}) \right] - \mathcal{D}_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}_j)||p_\theta(\boldsymbol{z})) \right. \tag{5.1}$$

Conversely, a Ladder VAE (LVAE) [296] is a hierarchical latent variable model that employs a hierarchy of $i = 1, ..., L$ conditional latent variables in the generative model and it is schematised in Fig. 5.1 (d). The total prior distribution $p_\theta(\boldsymbol{z})$ of this model is factorised as:

$$p_\theta(\boldsymbol{z}) = p_\theta(\boldsymbol{z}_L) \prod_{i=1}^{L-1} p_\theta(\boldsymbol{z}_i|\boldsymbol{z}_{i+1}) \tag{5.2}$$

$$p_\theta(\boldsymbol{z}_i|\boldsymbol{z}_{i+1}) = \mathcal{N}(\boldsymbol{z}_i|\boldsymbol{\mu}_{p,i}(\boldsymbol{z}_{i+1}), \boldsymbol{\sigma}_{p,i}^2(\boldsymbol{z}_{i+1})) \quad \forall i < L \tag{5.3}$$

$$p_\theta(\boldsymbol{z}_L) = \mathcal{N}(\boldsymbol{z}_L|\boldsymbol{0}, \boldsymbol{1}) \tag{5.4}$$

where the highest latent space ($i = L$) has a prior distribution $p_\theta(\boldsymbol{z}_L)$ which is typically assumed to be a Gaussian distribution with $\boldsymbol{\mu}_{p,L} = 0$ and $\boldsymbol{\sigma}_{p,L}^2 = 1$ (Eq. 5.4), while the other levels in the hierarchy have their prior values of $\boldsymbol{\mu}_{p,i}$ and $\boldsymbol{\sigma}_{p,i}^2$ that conditionally depend on the upper levels of the ladder (Eq. 5.3).

The LVAE inference model also differs from a standard VAE. In particular, each layer i in the hierarchy of the latent variables is conditioned on the previous stochastic layers and the total inference model $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is specified by the following fully factorised Gaussian distribution:

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}) = q_\phi(\boldsymbol{z}_1|\boldsymbol{x}) \prod_{i=1}^{L-1} q_\phi(\boldsymbol{z}_{i+1}|\boldsymbol{z}_i) \tag{5.5}$$

$$q_\phi(\boldsymbol{z}|\cdot) = \mathcal{N}(\boldsymbol{z}_i|\boldsymbol{\mu}_{d,i}, \boldsymbol{\sigma}_{d,i}^2) \tag{5.6}$$

In contrast with standard hierarchical VAEs [295], where the inference $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and prior distributions $p_\theta(\boldsymbol{z})$ are computed separately with no explicit sharing of information, the LVAE framework introduces a new inference mechanism. As shown in Fig. 5.1 (e), at each level i, an approximate likelihood estimation $\boldsymbol{\mu}_{e,i}$ and $\boldsymbol{\sigma}_{e,i}^2$ of its latent Gaussian distribution parameters is obtained from the encoder branch. This likelihood estimation is combined with the prior estimates $\boldsymbol{\mu}_{p,i}$ and $\boldsymbol{\sigma}_{p,i}^2$ obtained from the generative branch to produce a posterior estimation $\boldsymbol{\mu}_{d,i}$ and $\boldsymbol{\sigma}_{d,i}^2$ of the latent Gaussian distribution at that level i. In particular, this sharing mechanism between the inference (encoder) and generative (decoder) branches is performed at each level $i \neq L$ through a precision-weighted combination of the form:

$$\boldsymbol{\sigma}_{d,i}^2 = \frac{1}{\boldsymbol{\sigma}_{e,i}^{-2} + \boldsymbol{\mu}_{p,i}^{-2}} \qquad \boldsymbol{\mu}_{d,i} = \frac{\boldsymbol{\mu}_{e,i}\boldsymbol{\sigma}_{e,i}^{-2} + \boldsymbol{\mu}_{p,i}\boldsymbol{\sigma}_{p,i}^{-2}}{\boldsymbol{\sigma}_{p,i}^{-2}} \tag{5.7}$$

while $\boldsymbol{\mu}_{d,L} = \boldsymbol{\mu}_{e,L}$ and $\boldsymbol{\sigma}_{d,L}^2 = \boldsymbol{\sigma}_{e,L}^2$. This combination enables to build a data-dependent posterior distribution at each level, $\mathcal{N}(\boldsymbol{\mu}_{d,i}, \boldsymbol{\sigma}_{d,i}^2)$, that is both a function of the values assumed in the higher levels of the generative model and of the inference information derived of the subsequent (lower) levels. The loss function of the LVAE is the same of a VAE (Eq. 5.1) with the only difference that the number of KL divergence terms is equal to the number of levels L in the ladder. These KL divergence terms force the learned prior and posterior distributions at each level to be as close as possible. The sharing of information between the encoder and decoder through Eq. 5.7 promotes the learning of a data-dependent prior distribution better suited for the dataset to be modelled. More-over, this provides a better and more stable training procedure as the inference (encoder) branch iteratively corrects the generative distribution, instead of learning the posterior and prior values separately [296].

The full LVAE generative model has therefore the following formulation:

$$p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} p_\theta(\boldsymbol{x}|\boldsymbol{z}_1)\, p_\theta(\boldsymbol{z}_L) \prod_{i=1}^{L-1} p_\theta(\boldsymbol{z}_i|\boldsymbol{z}_{i+1})\, d\boldsymbol{z} \qquad (5.8)$$

5.2.2   LVAE for Interpretable Shape Analysis

In the previous Chapter (Section 4.2.2), I proposed a modification of the standard VAE to include a classification network $p(\boldsymbol{y}|\boldsymbol{z})$ able to predict the disease class label $\boldsymbol{y}$ associated with a segmentation $\boldsymbol{x}$ by using its latent representation $\boldsymbol{z}$ (the corresponding graphical model is shown in Fig. 5.1 (b)). In this work, I hypothesise that such modification can be extended to the LVAE framework by connecting a MLP $p(\boldsymbol{y}|\boldsymbol{z}_L)$, which classifies the disease status $\boldsymbol{y}$ of an input segmentation $\boldsymbol{x}$ by only using its latent representation on the highest latent space $\boldsymbol{z}_L$ (graphical model in Fig. 5.1 (c)). By end-to-end training the LVAE+MLP architecture, I aim at encoding a very low-dimensional latent space $\boldsymbol{z}_L$ the most discriminative features for the classification task under study, while the other latent spaces will model all the modes of shape variation needed to reconstruct the input segmentations $\boldsymbol{x}$. This approach yields two main advantages over the previous approach:

1. Template shapes for each disease class can be obtained by sampling from the learned distributions in a top-down fashion (starting from the highest level in the hierarchy $p(\boldsymbol{z}_L|\boldsymbol{y})$ and subsequently from every prior $p_\theta(\boldsymbol{z}_i|\boldsymbol{z}_{i+1})$). The posterior $p(\boldsymbol{z}_L|\boldsymbol{y})$ can be estimated by kernel density estimation and, since $\boldsymbol{z}_L$ is typically very low-dimensional, this estimation is straightforward;

2. If the latent space $\boldsymbol{z}_L$ is designed to be 2D or 3D, the distributions $p(\boldsymbol{z}_L|\boldsymbol{y})$ in the classification space can be directly visualised without the need of further offline dimensionality reduction techniques required in previous works [297, 286].

Figure 5.2: Detailed scheme of the LVAE+MLP architecture adopted in this work for the cardiac application. Top: encoder model; Bottom: decoder model. At testing, segmentations class scores y are computed with $z_3 = \mu_{e,3}$. The green arrows indicate the loss function terms used to train the network.

5.2.3   Datasets
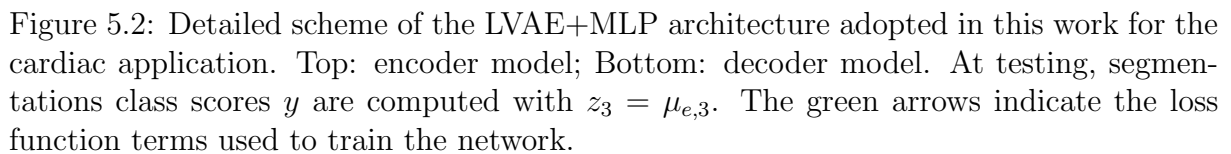
Cardiac Dataset   3D segmentations at ED and ES of the 436 HCM patients and 451 healthy volunteers from the Imperial College Dataset and of 20 healthy volunteers and 20 HCMs from the ACDC MICCAI'17 challenge training dataset [60] used in the previous chapter were also employed in this work. The same pre-processing steps outlined in Section 4.2.1 were adopted, please refer to that section for the full details and population characteristics.

Brain Dataset   A total of 726 3D left and right hippocampus segmentations of healthy controls (HC, $N = 404$, 202 males, median age 74.2 [min=59.8;max=89.6]) and Alzheimer's

disease subjects (AD, $N = 322$, 177 males, median age 75.8 [min=55.1;max=91.4]) from a publicly available repository were also analysed in this chapter. The segmentations were obtained from baseline T1-weighted (T1w) MR brain images from the ADNI-1/-GO/-2 cohorts using a multi-atlas label propagation method with expectation-maximisation based refinement (MALPEM) [292]. Images were automatically segmented individually and no additional pre-processing was performed. All segmentations were rigidly registered to the MNI standard reference space using nearest neighbour interpolation. Shape-based interpolation was applied to upsample each segmentation to 0.75mm x 0.75mm x 0.75mm resolution. Finally, each segmentation was cropped and padded using a bounding box positioned at its centre to obtain 3D segmentations of dimension $[x = 60, y = 60, z = 60, t = 1]$ for both the left and right hippocampus. Moreover, a 3D high-resolution left and right hippocampus volumetric template segmentation was obtained by averaging the upsampled and rigidly registered healthy controls segmentations. By thresholding the template probabilistic segmentation, a template triangular mesh was extracted using marching cubes algorithm and will be used in this work for results visualisation.

5.2.4 Application to Pathological Remodelling - LVAE+MLP model details

A detailed scheme of the three-level ($L = 3$) LVAE+MLP architecture employed in this work for the classification of HCM patients versus healthy subjects is summarised in Fig. 5.2, while the corresponding architecture for the classification of healthy controls versus AD patients is reported in Figure 5.3. For the sake of display clarity the model schemes were split into two rows: the encoder (inference) branch is shown at the top while the decoder (generative) branch is depicted at the bottom, and the two branches are connected by the latent space z_3. In the cardiac application, the input of the encoder branch are the 3D LV segmentations at ED and ES for each subject under study, which are presented as a two-channel input (top-left of Fig. 5.2), while in the brain application,
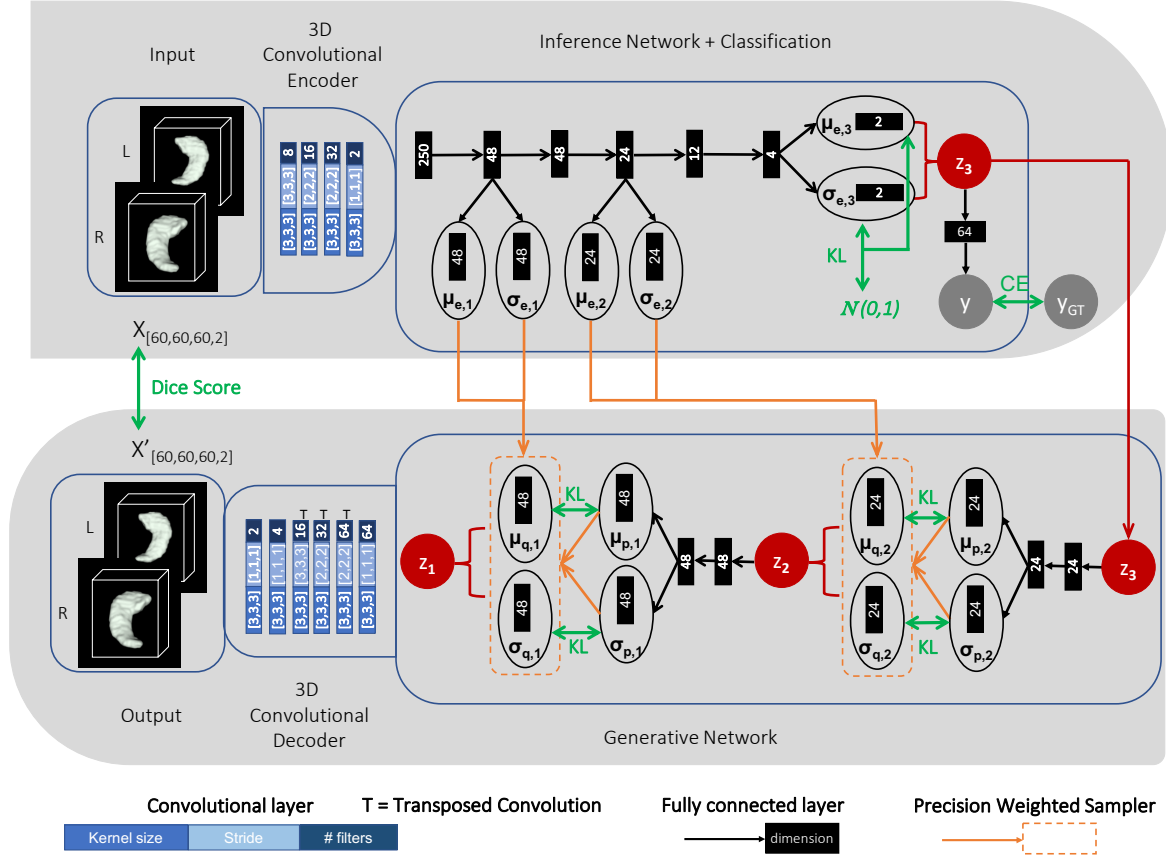
Figure 5.3: Detailed scheme of the LVAE+MLP architecture adopted in this work for the brain application. Top: encoder model; Bottom: decoder model. At testing, segmentations class scores y are computed with $z_3 = \mu_{e,3}$. The green arrows indicate the loss function terms used to train the network.

the input of the encoder branch are the 3D left and right hippocampus segmentations (top-left of Fig. 5.3). In the next paragraph, the flow of data in the network will be explained in detail for the cardiac application, and it is identical for the brain application expect for dimensional details.

A 3D convolutional encoder compresses them into a 250-dimensional embedding through a series of 3D convolutional layers with stride 2. This embedding is used then as input of a deterministic inference network, which computes the likelihood estimates $\boldsymbol{\mu}_{e,i}$ and $\boldsymbol{\sigma}_{e,i}$ for each level i of the hierarchy of latent variables. These estimates are derived by manipulating the input through a series of fully connected layers (black arrows), which are all followed by batch normalisation and *elu* non-linearity with the only exception of the layers computing $\boldsymbol{\mu}_{e,i}$ and $\boldsymbol{\sigma}_{e,i}$. At the highest latent space ($i = 3$ in this case), a shallow

MLP (2 layers) is attached to learn $p(\boldsymbol{y}|\boldsymbol{z}_3)$, i.e. to predict the class (HCM or healthy) label $\boldsymbol{y}$ corresponding to the input segmentation $\boldsymbol{x}$ by just using its latent variable values $\boldsymbol{z}_3$. *ReLu* was used as non-linearity after the first layer. The latent variable values $\boldsymbol{z}_3$ are sampled during training from $\mathcal{N}(\boldsymbol{\mu}_{d,3}, \boldsymbol{\sigma}_{d,3}^2)$ where $\boldsymbol{\mu}_{d,3} = \boldsymbol{\mu}_{e,3}$ and $\boldsymbol{\sigma}_{d,3} = \boldsymbol{\sigma}_{e,3}$ and they are also the starting point of the generative process (bottom-right of Fig. 5.2). At each level i of the generative (decoder) network, the prior distribution terms are computed by modifying the values of the previous latent space $\boldsymbol{z}_{i+1}$ through a fully connected layer followed by batch normalization and *elu* non-linearity and by a second fully connected layer. These prior values are combined with $\boldsymbol{\mu}_{e,i}$ and $\boldsymbol{\sigma}_{e,i}$ through Eq. 5.7 to obtain the posterior estimates $\boldsymbol{\mu}_{d,i}$ and $\boldsymbol{\sigma}_{d,i}$ from which $\boldsymbol{z}_i$ is sampled. Finally, the value of $\boldsymbol{z}_1$ is passed to a 3D convolutional decoder which aims to reconstruct the input segmentations $\boldsymbol{x}$ through a series of upsampling and convolutional layers. After every convolutional and upsampling layer used in the architecture *ReLu* was applied as non-linearity, except at the output of the network where *sigmoid* was applied. All the network weights were randomly initialised from a zero-mean Gaussian distribution ($\boldsymbol{\sigma} = 0.02$). In the brain application, the same architecture was replicated with the only exception of a different layer dimensionality due to a different input dimension.

The training loss function of the LVAE+MLP network is composed of three contributions: 1) two LV segmentation reconstruction accuracy terms (for the ED and ES LV segmentations or for the left and right hippocampus segmentations) and it defined as the overlap (Dice score) between the input segmentation $\boldsymbol{x}$ and its reconstruction $\boldsymbol{x}'$; 2) L KL divergence terms, penalising discrepancies between the learned prior and posterior distributions at each level and 3) a binary classification cross entropy (CE) term for the classification of healthy versus HCM segmentations or healthy controls vs AD patients. All the KL_i divergence terms except the one of the highest level ($i = 3$) were evaluated between the prior distribution $\mathcal{N}(\boldsymbol{\mu}_{p,i}, \boldsymbol{\sigma}_{p,i}^2)$ and their posterior distribution $\mathcal{N}(\boldsymbol{\mu}_{d,i}, \boldsymbol{\sigma}_{d,i}^2)$, while for the highest level the prior distribution was assumed to be a standard Gaussian

$\mathcal{N}(0,1)$. The total loss function is

$$\mathcal{L} = DSC_{ED} + DSC_{ES} + \gamma \left[\sum_{i=0}^{L} \alpha_i \, KL_i + \beta \, CE \right] \tag{5.9}$$

and depends on α_i, which weights the KL terms, on β, which weights the classification loss, and on γ, which is set to increase from 0 to 1 at the beginning of the training. This increase of γ is called deterministic warm-up and it has been commonly found useful in practice to converge to better local minima [296]. The weighting of the KL terms and the use of the Dice Score as a reconstruction metric lead to a different lower bound than standard VAE and LVAE. In the literature, it has been shown that the use of variants of the VAE lower-bound tend to favor better empirical results in various problems [298]. In this work, Dice score was adopted as a reconstruction metric since it was successfully used in the work presented in Chapter 4 and in related work [228] to achieve better reconstruction results on 3D anatomical segmentations.

At testing, a pair of ED and ES LV segmentations (or left and right hippocampus segmentations) are reconstructed by starting from $z_3 = \mu_{d,3}$ and by assigning to z_2 and z_1 the values $\mu_{d,2}$ and $\mu_{d,1}$ computed from $z_3 = \mu_{d,3}$ and $z_2 = \mu_{d,2}$, i.e. no sampling is performed from the posterior distribution at each level. To interpret the anatomical information encoded by the highest latent space, at each level $i \neq 3$, the value of $\mu_{p,i}$ can be assigned to z_i instead of $\mu_{d,i}$ and the segmentations are reconstructed as explained above. In this way, by varying the values of z_3, a set of segmentations can be directly generated for each point in z_3, without using the inference information provided by $\mu_{e,i}$ and $\sigma_{e,i}$. This enables the visualisation of the anatomical information encoded by the highest latent space. Finally, in order to visualise the distribution of a set of segmentations under exam in the highest latent space, the $\mu_{e,3}$ values of each segmentation can be computed through the inference network and directly plotted in a 2D space.

5.3 Cardiac application

5.3.1 Model Training

Healthy and HCM subjects' 3D segmentations at ED and ES from the Imperial College dataset were randomly divided into train, validation and test sets consisting of a total of 537 (276 from healthy volunteers, 261 from HCMs), 150 (75 from healthy volunteers, 75 from HCMs) and 200 (100 from healthy volunteers, 100 from HCMs) segmentations. A 3-level LVAE+MLP model (Fig. 5.2) was adopted since adding more levels in the generative model did neither improve the segmentation reconstruction accuracy, nor the classification accuracy in the clinical application under exam. The model was end-to-end trained on an NVIDIA Tesla K80 GPU using Adam optimiser with learning rate equal to 10^{-4} and batch size of 16. For the first 40k iterations, data augmentation including rotations around the three standard axis with rotation angles randomly extracted from a Gaussian distribution $\mathcal{N}(0, 6°)$ was applied in order to take into account small mis-registrations between the subjects' segmentations. This data augmentation strategy helped the final model to achieve higher reconstruction accuracy (higher Dice score and lower 2D slice-by-slice Hausdorff distance) both at testing and training phases, as it can be seen in Table 5.1.

In the loss function (Eq. 5.9), the KL weights were fixed to $\alpha_1 = 0.02$, $\alpha_2 = 0.001$ and $\alpha_3 = 0.0001$ while γ was set to increase from 0 to 100 by steps of 0.5 every 4k iterations. The relative magnitude and ascending order of the KL weights α_i were chosen as this provided both better classification and reconstruction accuracy in contrast with models having all the weights α_i equal or in descending order (results are reported in Table 5.2). This suggests the higher levels of a LVAE might be more difficult to train, and that a lower KL regularization term helps the generative model training. The model produced similar results when varying these parameters within one order of magnitude, while a further increase in value reduced reconstruction accuracy and a further decrease resulted

Effect of DA and DWU

Augmentation	DSC_{ED}	DSC_{ES}	H_{ED} [mm]	H_{ES} [mm]	ACC [%]
Training					
None	0.75±0.07	0.79±0.05	7.30±1.80	7.08±1.68	51.40%
DA	0.77±0.05	0.80±0.05	6.94±1.62	6.86±1.53	51.40%
DWU	0.82±0.05	0.86±0.04	6.20±1.23	5.93±1.23	99%
DA&DWU	0.85± 0.04	0.88±0.03	5.70±1.12	5.58±1.00	100%
Testing					
None	0.72±0.07	0.76±0.05	8.01±1.99	7.53±1.97	51.40%
DA	0.74±0.06	0.78±0.05	7.62±1.86	7.31±1.82	51.40%
DWU	0.79±0.05	0.83±0.04	6.91±1.79	6.72±1.68	99%
DA&DWU	0.81±0.04	0.85±0.04	6.54±1.62	6.40±1.56	100%

Table 5.1: Dice score (DSC) and average 2D slice-by-slice Hausdorff distance (H) at ED and ES and their standard error for the proposed LVAE+MLP model when Deterministic Warm-Up (DWU) and Data Augmentation (DA) are applied. ACC is the classification accuracy of the different models.

Effect of the KL weights

$[\alpha_1, \alpha_2, \alpha_3]$	DSC_{ED}	DSC_{ES}	H_{ED} [mm]	H_{ES} [mm]	ACC [%]
Training					
$[10^{-4}, 2\,10^{-4}, 10^{-3}]$	0.79± 0.05	0.80± 0.05	6.93±1.62	6.88±1.60	100%
$[10^{-4}, 10^{-4}, 10^{-4}]$	0.80± 0.05	0.83± 0.04	6.50±1.41	6.48± 1.47	100%
$[10^{-3}, 2\,10^{-4}, 10^{-4}]$	0.85± 0.04	0.88± 0.03	5.70±1.12	5.58± 1.00	100%
Testing					
$[10^{-4}, 2\,10^{-4}, 10^{-3}]$	0.75± 0.06	0.78± 0.05	7.64±1.72	7.37±1.68	99%
$[10^{-4}, 10^{-4}, 10^{-4}]$	0.78± 0.05	0.80± 0.04	7.01±1.53	6.94±1.58	100%
$[10^{-3}, 2\,10^{-4}, 10^{-4}]$	0.81±0.04	0.85±0.04	6.54±1.62	6.40±1.56	100%

Table 5.2: Dice score (DSC), average 2D slice-by-slice Hausdorff distance (H) at ED and ES and they standard error of the mean together with classification accuracy (C) for the proposed $LVAE+MLP$ model for different sets of the KL weights α_i in the training loss function. ACC is the classification accuracy of the different models.

in model overfitting.

The classification loss function weight β was instead set to 0.005: experiments showed that a higher β value would have still produced a good model, but at the price of a more unstable training at the early stages. With regards to the number of layers and nodes

adopted in the MLP, adopting a single fully connected layer posed a too strong constraint on the latent space distribution, preventing a successful posterior learning, and using more than two layers caused overfitting. The derived constrain on the latent space distribution is due to the fact that if the MLP is shallow network, it tries to separate healthy and HCM segmentations with a simple, linear separation in the latent space, not being able to differentiate them otherwise with the limited capacity it has. This however makes difficult the matching of the prior and posterior distributions in the latent space. After 220k iterations the training procedure was stopped as the increase of the KL divergence started to interfere with the decrease of the reconstruction and classification losses. In particular, this is due to the fact that in the highest latent space the KL divergence term tries to cluster all the data together, while the classification loss tries to separate the clusters. Hence the relative weight of β and α_3 needs to be tuned in order to obtain a good equilibrium.

Finally, the gradual increase of the classification and KL divergence weights during training through the γ parameter, known as deterministic warm-up [296], proved to be crucial to construct an expressive generative model, as the results reported in Table 5.1 show. As already exposed in previous research [299, 296], the KL regularization can cause many latent variable to become inactive, *i.e.* their posterior matches their prior before their learned a useful representation, and to not model any shape effect. By gradually turning on the Kl weight, this effect can be alleviated.

5.3.2   Classification and Reconstruction Results

All the 200 subjects in Imperial College testing dataset were correctly classified (100% sensitivity and specificity) by the trained prediction network. The same model also correctly classified 36 out of the 40 ACDC MICCAI 2017 segmentations (100% sensitivity and 80% specificity) without the need of any re-training procedure. The four misclassified segmentations correspond to the same ones misclassified by the previous approach
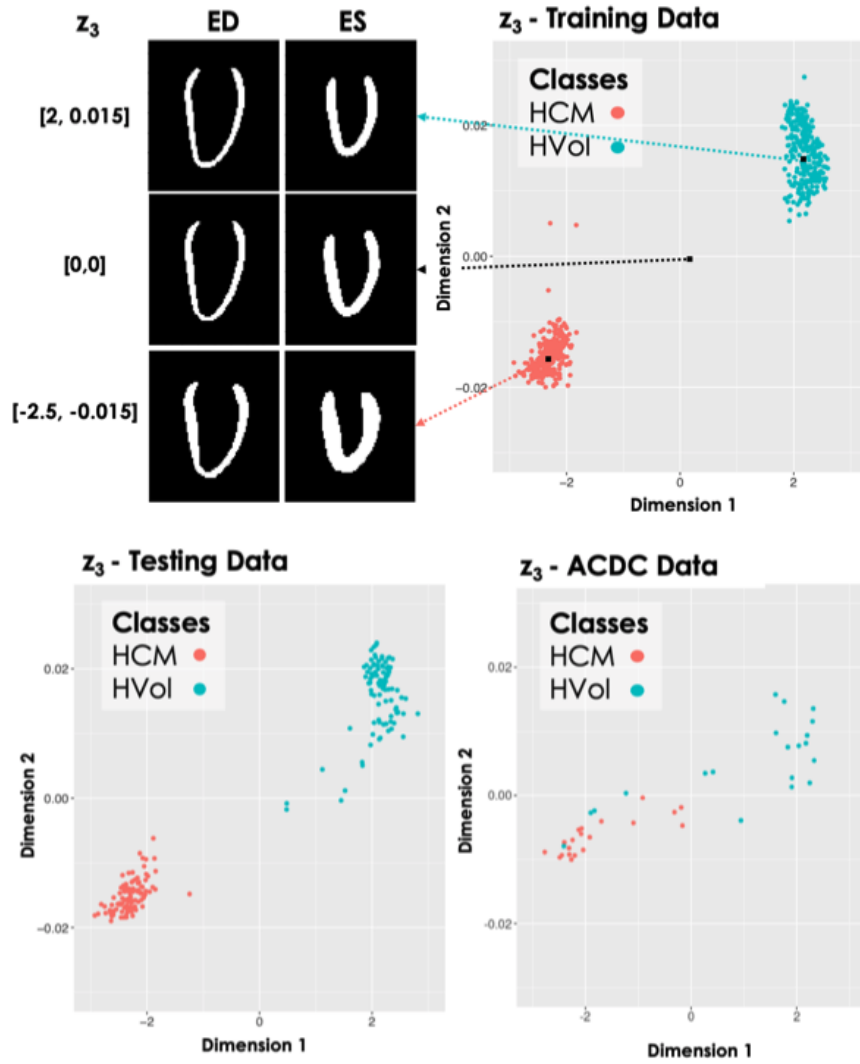
Figure 5.4: Latent space clusters in the highest latent space ($l = 3$) obtained by the proposed LVAE+MLP model on both the in-house training and testing datasets as well as on the ACDC dataset (entirely used as an additional testing dataset). Dimension 1 and 2 represent the two dimensions of $\mu_{e,3}$. On the left, long-axis sections of the reconstructed 3D segmentations at ED and ES obtained by sampling from three points in z_3 are shown.

(Section 4.3), suffering from a lack of coverage of the LV apex which causes them to be classified as pathological. The clustering of the shapes under exam as obtained by the proposed model are shown in Fig. 5.4, where two separated clusters of segmentations have been discovered both on the training and on the testing data. An analogous result was obtained with the VAE-based model (Section 4.3, Figure 4.2): however, the previous version of the model required an additional dimensionality reduction step to visualise in 2D the obtained latent space of segmentations, while in the new proposed framework the
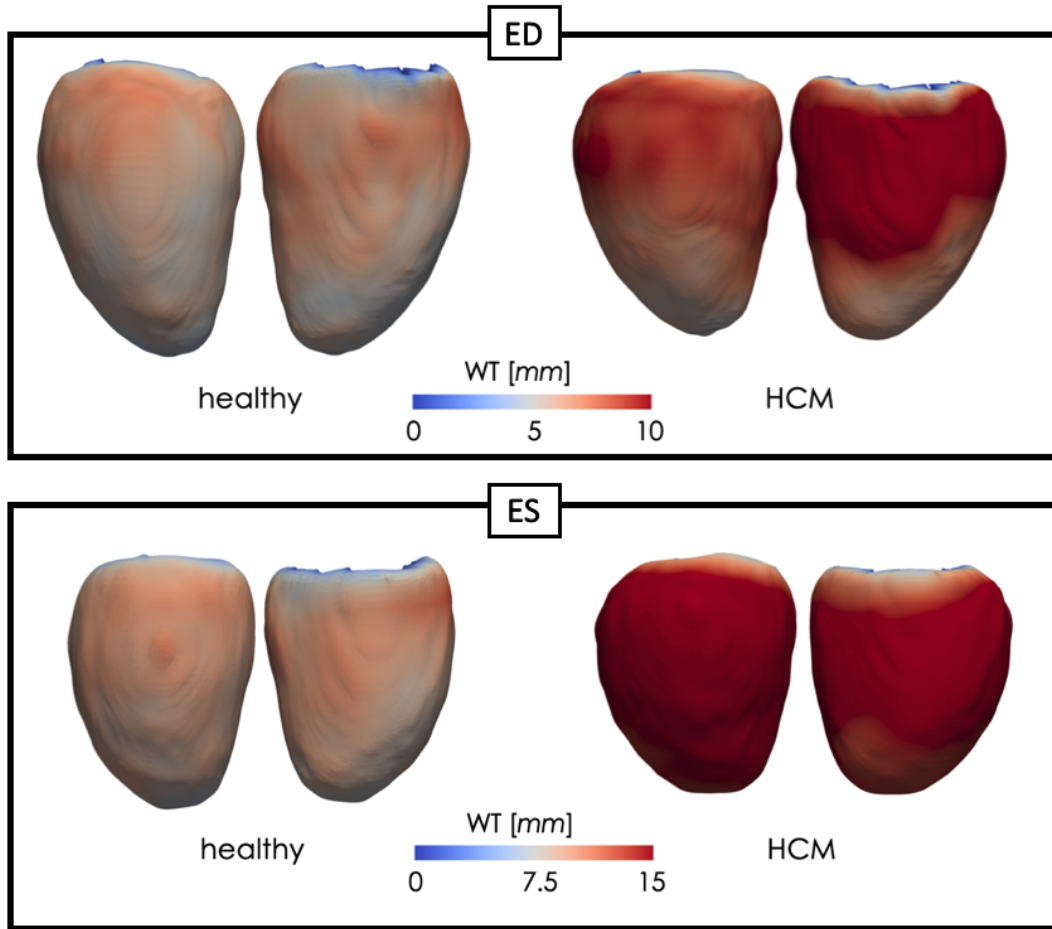
Figure 5.5: Average healthy and HCM shapes at ED and ES sampled from the two clusters in the highest latent space of proposed LVAE+MLP model. The colormap encodes the vertex-wise wall thickness (WT), measured in mm.

VAE+MLP vs LVAE+MLP Reconstruction Accuracy				
Training				
Model	DSC_{ED}	DSC_{ES}	H_{ED} [mm]	H_{ES} [mm]
VAE+MLP	0.81±0.04	0.85±0.04	6.30±1.25	5.96±1.20
LVAE+MLP	0.85±0.04	0.88±0.03	5.70±1.12	5.58±1.00
Testing				
VAE+MLP	0.78±0.04	0.83±0.04	6.98±1.65	6.75±1.61
LVAE+MLP	0.81±0.04	0.85±0.04	6.54±1.62	6.40±1.56

Table 5.3: Cardiac. Dice score (DSC) and average 2D slice-by-slice Hausdorff distance (H) at ED and ES and their standard deviation for the proposed LVAE+MLP model and for the VAE+MLP model proposed in Chapter 4 on training and testing sets.
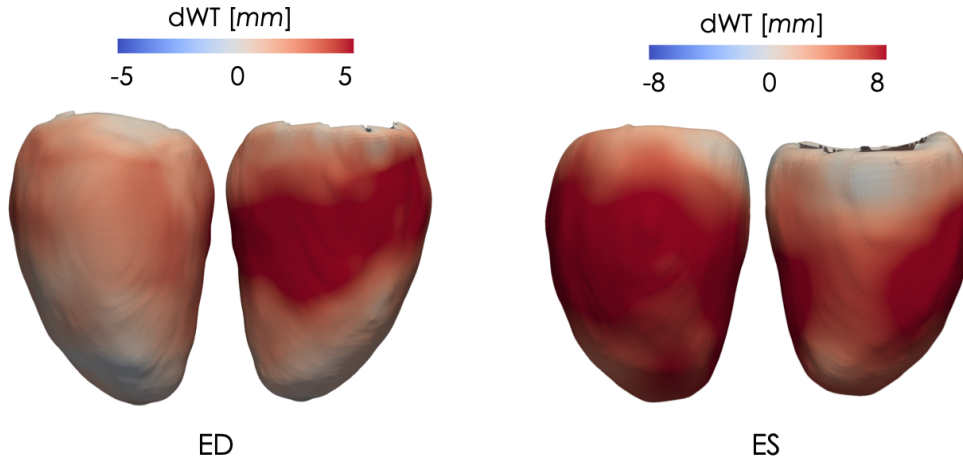
Figure 5.6: Point-wise difference in wall thickness (dWT) at ED and ES between the healthy and the HCM average shapes of Fig. 4. Left - lateral wall; Right - septal wall.

highest latent space is 2D by design.

The new model achieved higher reconstruction accuracy than the previous model, as shown in Table 5.3, suggesting that a better generative model of shapes was learned. In particular, the table shows the reconstruction accuracy in terms of 3D Dice score and average 2D slice-by-slice Hausdorff distance between the 3D original and reconstructed segmentations on the testing and training datasets obtained by the proposed LVAE+MLP model and the previously proposed VAE-based model (VAE+MLP). The VAE+MLP model was constructed with the same 3D convolutional encoder and decoder networks of the LVAE+MLP model and with a single latent space composed of 98 latent variables, which corresponds to the total number of latent variables adopted in the LVAE+MLP model (three levels of 64, 32 and 2 latent variables, respectively). As it can be noticed in the table, the obtained Dice score results at ES are better than at ED for all the models, while the Hausdorff results seem to follow instead an opposite trend. This is probably due to the fact that since the LV is more compact at ES, the Dice score might not be sensitive to small misalignment of the reconstructed shape, which are instead captured by the Hausdorff distance.
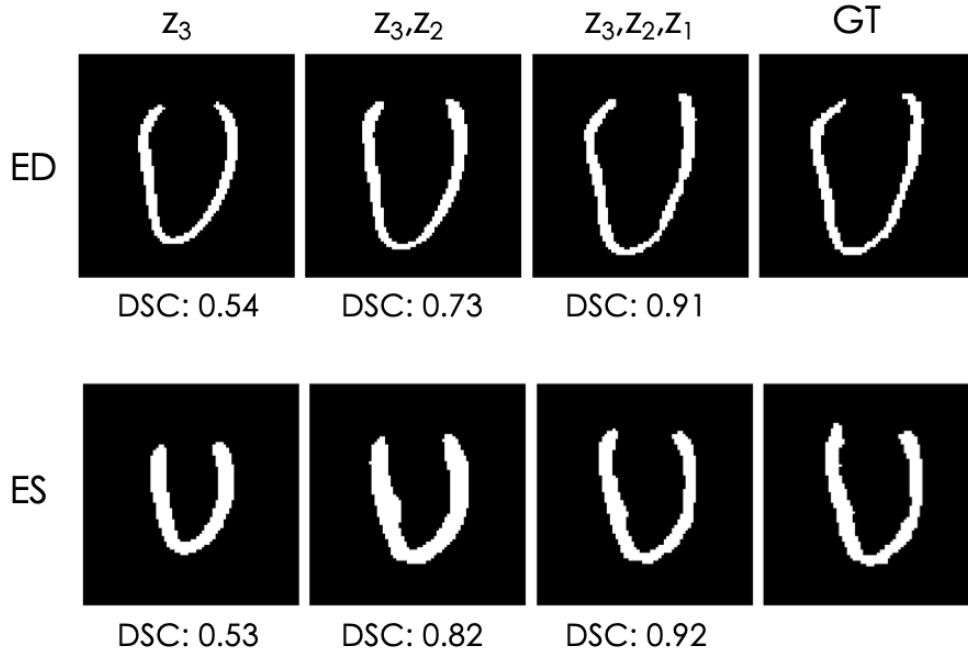
Figure 5.7: Long-axis section of reconstructed LV segmentations at ED and ES by the LVAE+MLP model, using only z_3 information (first column) or also using the posterior information of the other latent spaces (z_2, z_1) . Last column: ground-truth (GT) segmentation. DSC = Dice Score between the segmentation at that column and the GT.

5.3.3 Visualisation of the latent spaces

Due to the properties of the proposed model, the anatomical information encoded by each latent space can be directly visualised, especially the anatomical information embedded in the highest level $(i = 3)$, which encodes the most discriminative features for the classification of healthy and HCMs 3D LV segmentations. For the exemplar application under investigation, little intra-cluster variability between the shapes generated from the latent space z_3 was obtained, while much larger inter-cluster variability between the generated shapes was obtained. This can be seen on the left-side of Fig. 5.4 where the long-axis section of the 3D reconstructed segmentations at ED and ES at three points of the latent space z_3 are reported. In Appendix A a detailed grid visualisation of the shapes encoded by this latent space is also reported.

In Fig. 5.5 the obtained mean average shape for each cluster is shown, represented as a triangular mesh with point-wise wall thickness (WT) values at vertex. This was obtained by
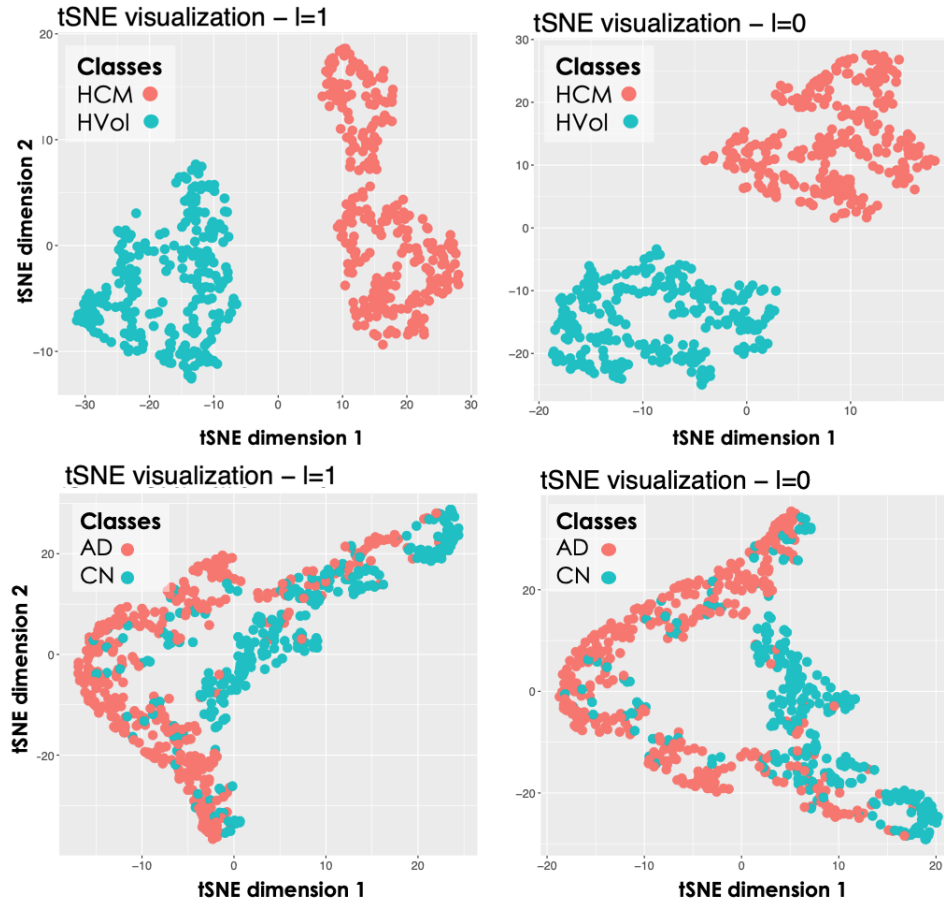
Figure 5.8: tSNE visualisation of the latent spaces z_2 and z_1. Top: cardiac application. Bottom: brain application.

sampling $N = 1000$ segmentations from each cluster in z_3 after estimating its probability density via kernel density estimation. Then, the obtained segmentations for each cluster were averaged to extract the corresponding average segmentation. Finally, a non-rigid transformation between the obtained average segmentation and a 3D high-resolution LV segmentation from the UK Digital Heart project[2] was computed, and the inverse of this transformation was applied to the corresponding 3D high-resolution LV segmentation to warp it to the cluster specific average segmentation. At each of the mesh vertices, WT was then computed as the perpendicular distance between the endocardial and epicardial wall. The results are presented in Fig. 5.5, where it can be noticed that the average HCM shape has higher WT than the corresponding healthy shape and it has a slightly reduced

---

[2]https://digital-heart.org/

VAE+MLP vs LVAE+MLP Reconstruction Accuracy			
	DSC_l	DSC_r	H_r [mm]	H_l [mm]
Training			
VAE+MLP	0.81± 0.05	0.80±0.05	3.35±0.67	3.28±0.69
LVAE+MLP	0.85± 0.04	0.85± 0.03	3.05±0.69	2.96±0.66
Testing			
VAE+MLP	0.79±0.05	0.79±0.05	3.51±0.64	3.49±0.67
LVAE+MLP	0.82±0.03	0.82±0.03	3.31±0.68	3.23±0.65

Table 5.4: Brain. Dice score (DSC) and average 2D slice-by-slice Hausdorff distance (H) for the left (l) and right (r) hippocampus and their standard deviation for the proposed LVAE+MLP model and for the VAE+MLP model proposed in Chapter 4 on training and testing sets.

size. Fig. 5.6 instead reports the point-wise difference in WT between the HCM and the healthy shape, and it can be noticed that the most discriminative anatomical feature to classify an HCM shape consists in an increased WT in the septum, which is in agreement with the clinical literature [285].

Fig. 5.7 shows a long-axis section of the reconstructed segmentations at ED and ES from the LVAE+MLP model when only z_3 posterior information is used (first column) or when also the posterior information in the other levels (z_2, z_1) is exploited: the latent spaces z_2 and z_1 evidently encode different anatomical features that help to refine the structural information provided by z_3. Results for more subjects are reported in Appendix A. Finally, the dimensionality reduction technique tSNE [300] is applied to visualise in two dimensions the distributions of z_1 and z_2 latent spaces. The reported plots show that the latent representations of the two classes of shapes are clustered also at both these levels (plots shown in Fig. 5.8). A possible explanation relies on the fact that the generative process is a conditional: if the data is clustered at the top of the hierarchy, it may be easier for the network to keep the clusters also in the subsequent levels.

5.4   Brain application

5.4.1   Model Training

As an additional benchmark test, the LVAE+MLP model proposed in this work was also tested for the classification of healthy controls (HC) and patients with AD by using only 3D segmentations of the left and right hippocampus. Data was randomly assigned to train, validation and testing sets consisting of a total of 562 (322 HC, 240 AD), 64 (32 HC, 32 AD) and 100 (50 HC, 50 AD) segmentations respectively. A three level LVAE+MLP model was also adopted for this application (Figure 5.3), since adding more levels did not improve classification or reconstruction accuracy. In the model loss function (Eq. 5.9), the KL weights were fixed to the values of $\alpha_1 = 0.03$, $\alpha_2 = 0.003$ and $\alpha_3 = 0.0003$, γ was set to increase from 0 to 100 by steps of 0.5 every 4k iterations and β was instead set to 0.005. The same augmentation strategy and the rationale for the selection of the hyperparameters in the cardiac experiment were adopted. The model training was stopped after 200k iterations.

5.4.2   Classification and reconstruction results

84 out of 100 subjects were correctly classified by the training prediction network (78% sensitivity, 90% specificity). A VAE+MLP model with the same 3D convolutional encoder and decoder networks of the LVAE+MLP model, but with a single latent space of dimension 66 (equal to the total number of latent variables adopted in the LVAE+MLP model) was also trained. This model classified 81 out of 100 subjects correctly (74% sensitivity, 88% specificity) on the same training, testing and validation splits of the previous model. On the same dataset, an accuracy of 78% (75% sensitivity, specificity 80%) for the same classification task was obtained by using left and right hippocampus volume segmentations [292]. Compared to the VAE+MLP model, the LVAE+MLP model achieves
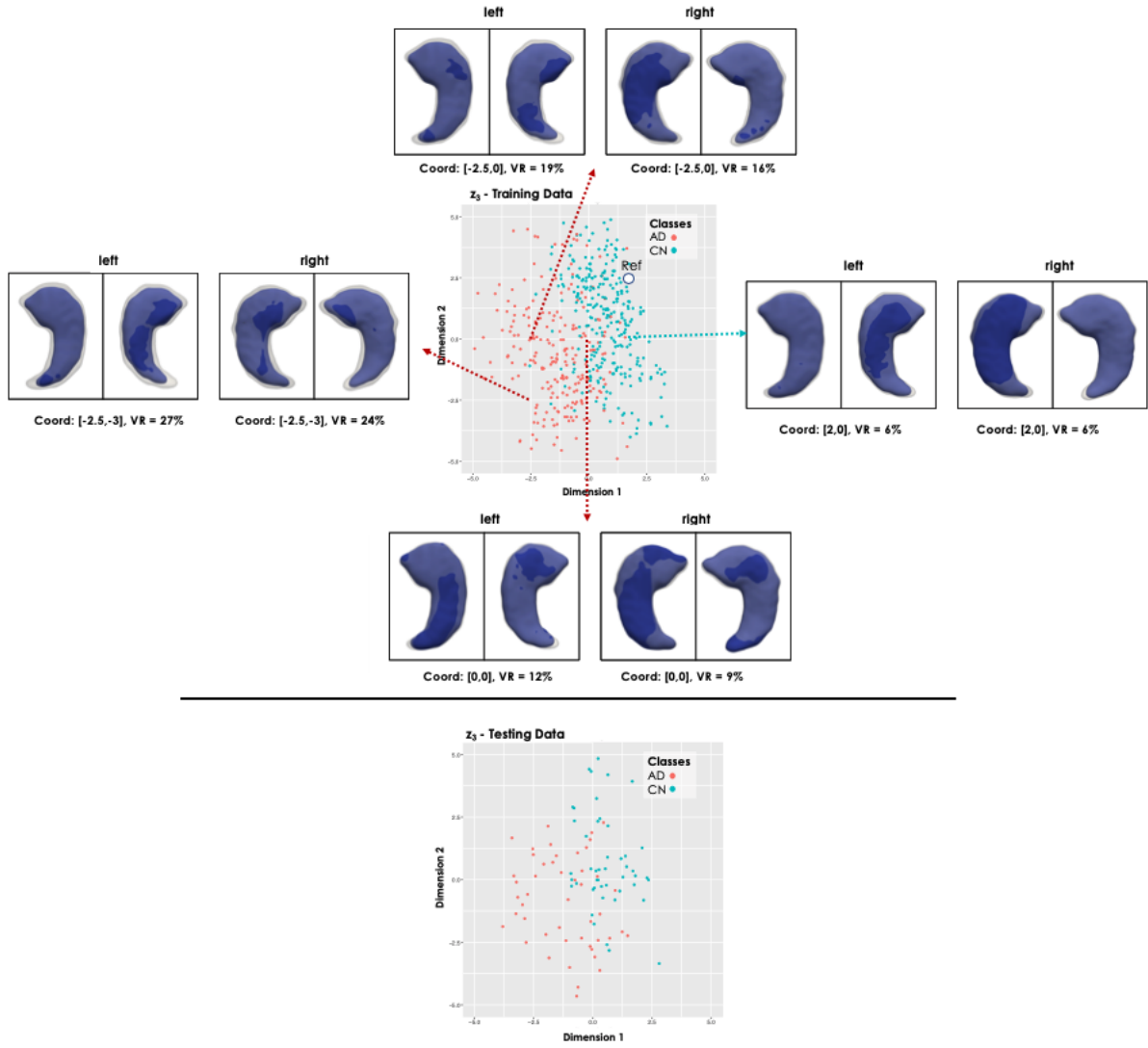
Figure 5.9: Latent space clusters in the highest latent space ($l = 3$) obtained by the proposed LVAE+MLP model on the brain dataset. Left and right hippocampus shapes (in blue) at four points in the latent space have been reconstructed and showed together with a reference shape (in grey and opaque) sampled from the healthy control shapes (Ref, Coord: [2,2]). The first image is a view from the top, second image a view from the bottom.

higher reconstruction accuracy in terms of 3D Dice score and 2D slice-by-slice Hausdorff distance between the original segmentations and the reconstructed ones, these results are reported in Table 5.4.

5.4.3   Visualisation of the latent spaces

Fig. 5.9 shows the distribution of the training and testing 3D hippocampus segmentations in the highest ($i = 3$) latent space for the trained LVAE+MLP model. It can be noticed how the healthy and pathological shapes are not as separated as in the previous application due to the more challenging nature of the new task. However, two clear clusters of healthy and AD shapes can still be identified. Fig. 5.9 also shows the left and right hippocampus segmentations obtained by sampling from four distinct points of this latent space, which are showed together with a reference healthy shape sampled from a point in the healthy cluster (marked as Ref). For each reconstructed segmentation, the rate of hippocampal volume change (VR) with respect to the reference healthy shape was computed ($VR = \left| \frac{V - V_{ref}}{V_{ref}} \right| \times 100$). From the figure, it can be noticed how the AD shapes are characterized by decreased hippocampal volume, reduction that slighty but consistently affects more the left than the right hippocampus, in agreement with the previous findins
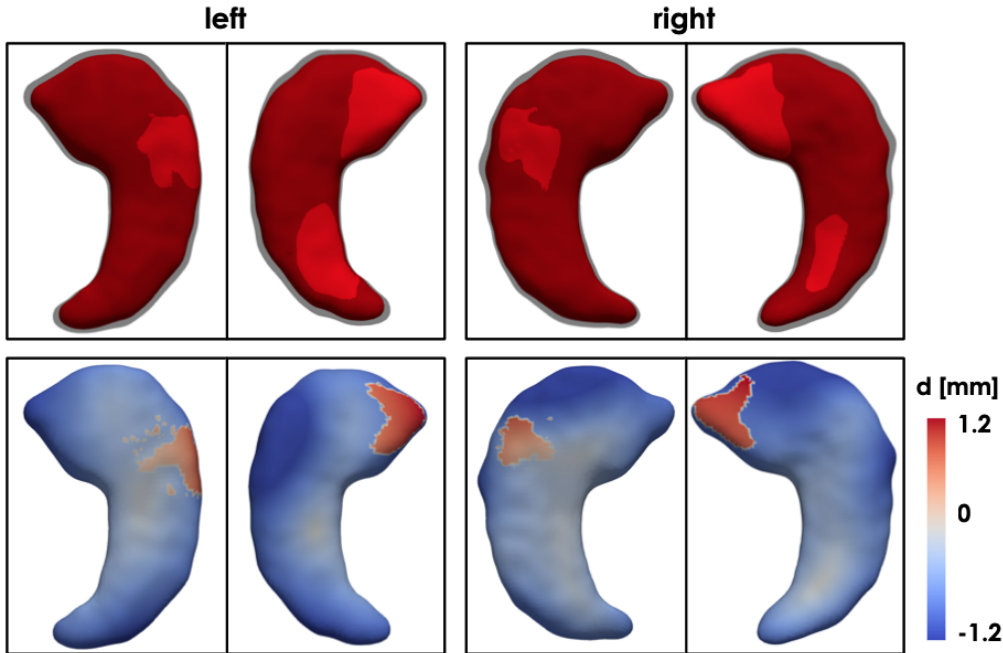
Figure 5.10: First row: Average healthy (in grey and opaque) and AD (in red) left and right hippocampus shapes sampled from the two clusters in the highest latent space of proposed LVAE+MLP model. Second row: vertex-by-vertex L2 distance between the two mean shapes.

on this data [292]. Moreover, a pattern in regional changes in volume can be identified: AD cases closer to the reference healthy shape show atrophy predominantly (if not only) in the tail of the hippocampus, while cases further away from the healthy class and deeper into the AD group show an atrophy pattern more spread throughout the whole hippocampal shape. In Fig. 5.10, the obtained average left and right hippocampus shapes from the healthy and AD distribution represented as triangular meshes are shown. These meshes were obtained by sampling $N = 1000$ segmentations from the healthy and AD distributions in z_3 after estimating their probability density via kernel density estimation. Then, the obtained segmentations for each cluster were averaged to extract the corresponding cluster mean segmentation. Finally, the 3D template hippocampus segmentation was non-ridigly registered to each obtained cluster specific average segmentation and the estimated transformation was applied to the corresponding high-resolution mesh. In the first row of Fig. 5.10, it can be noticed how the reconstructed template AD segmentation (red) which is shown together with the HC segmentation (grey and opaque) is more atrophied and it is characterized by a bending of the head of both the left and right hippocampus. The second row displays the vertex-by-vertex L2 distance between the two mean shapes demonstrating a more pronounced regional atrophy in the hippocampal head consistent with the CA1 and subiculum regional athrophy already reported in the literature [294], [301]. The right hippocampus is characterized by a 13.5% decrease in volume between the healthy shape and the AD shape, while the decrease in volume for the left hippocampus is 14.6%. The volume ratio between the AD right and left hippocampus is 3.6% and 2.5% in the healthy mean shape. Finally, the plots resulting from the application of tSNE dimensionality reduction technique to the z_1 and z_2 training data values are shown at the bottom of Fig. 5.8. Once again, it can be noticed a clear separation between healthy controls and AD patients also in these levels of the ladder.

5.5   Discussion

This work presents a data-driven framework to model a population of 3D anatomical segmentations through a hierarchy of conditional latent variables, encoding at the highest level of the hierarchy the most discriminative features to differentiate distinct clinical conditions. This is achieved by implementing and extending for the first time the LVAE framework to a real-world medical imaging application. In particular, building on the results obtained in the previous chapter, the new proposed model also performs a classification task in a generative latent space, which this time is the highest level of a LVAE model. In this way, this latent space can be forced to encode the most discriminative features for a clinical task under exam, while the other levels encode the other factors of anatomical variation needed to model the manifold of segmentations under analysis. Being a generative model, also this framework provides the advantage of enabling the visualisation and quantification of the remodeling effect encoded by each latent space in the original segmentation space. Hence, the anatomical differences used by the classifier to distinguish different conditions can be easily visualised and quantified by sampling from the posterior distribution of the LVAE highest level computed from a given database of shapes. Moreover, by designing this latent space to be two or three dimensional, no additional offline dimensionality reduction technique is required to visually assess the distribution of these shapes in the latent space. As a consequence, this method not only provides a deep learning classifier that uses a task-specific latent space in the discrimination of different clinical conditions, but more importantly enables the visualisation of the anatomical features encoded by this latent space, making the classification task transparent.

With the aim of assisting the clinicians in quantifying the morphological changes related to disease, the proposed framework was applied for the automatic classification of heart and brain pathologies against healthy controls. In the reported cardiac application, the learned features achieved high accuracy in the discrimination of healthy subjects from

HCM patients on our unseen testing dataset and on a second external testing dataset from the ACDC MICCAI 17 challenge. On the more challenging task of classification of healthy versus AD hippocampi, the model achieved better classification accuracy than using volumetric indices [292] and the previously proposed VAE-based method. Moreover, the visualisation of the features encoded in the highest level of the adopted LVAE+MLP model showed how the proposed model is able to provide the clinicians with a 3D visualisation of the most discriminative anatomical changes for the task under study, making the data-driven assessment of regional and asymmetric remodelling patterns characterizing a given clinical condition possible. On both applications, the reported results show that the proposed LVAE+MLP model allows the construction of a better generative model in comparison to a VAE-based model with a single latent space [297]. To the best of my knowledge, this result confirms for the first time that hierarchical latent spaces provide a more accurate generative model on a real clinical dataset. Moreover, this work also gives insights on the functioning of these models on 3D anatomical segmentations, including how the different levels of latent variables encode different anatomical features, and how to optimally train this class of models for the reconstruction of these 3D anatomical segmentations.

5.5.1 Limitations and Future Work

While this work showed the potential of the proposed method on two common classification tasks, this method is domain-agnostic and could be applied to other classification problems where pathological remodelling is a predictor of a disease class label. However, further work is needed to explore the full potential of this approach, for instance in order to visualize the pathological remodeling of different disease subgroups characterized by different clinical endpoints. Of note, it is expected that on very difficult tasks one or two more dimensions in the highest latent space might be needed, although further increasing the dimensionality will go against the rationale of the proposed approach. In fact, the

proposed model aims at encoding the most discriminative anatomical information for the classification task under exam in the highest latent space, while the other latent spaces are intentionally left to model the remaining factors of variation. Interestingly, Fig. 5.8 shows that the shapes are clustered also in the other latent spaces, probably encoding additional variability of the disease groups not strictly useful for the specific classification task. By specializing the classification task to more categories, it is expected that some information currently encoded in the other latent spaces to be moved and encoded in the highest one. For instance, studying multiple disease subgroups would enable a finer representation of the spectrum of remodeling patterns against which patients can be compared. Presently this was not achievable as the model has been optimized to discriminate only between healthy and diseased subjects, although a step in this direction was taken in Fig. 5.9, showing how different latent space points map to different hippocampal volume measures.

In comparison with the previous (generative) VAE-based model and Bello *et al.* [286] model, the proposed method requires tuning of a few additional hyperparameters, i.e. the number of adopted levels in the ladder and their weights importance in the model loss function. On the other hand, the proposed approach is fully data-driven and it spares the need for further downstream dimensionality reduction and latent space navigation techniques, which would themselves require separate optimization and human intervention, potentially adding bias to the analysis. The proposed method also enables the derivation of population-based inferences (Fig. 5.5, 5.6 and 5.10), which could neither have been obtained from VAE-based model (due to the subject-specific nature of the latent-space navigation), nor from the one of Bello *et al.* (due to the non-generative nature of the model).

Another limitation shared by the current approach is the fact that the input segmentations need to be rigidly registered to train the model. Future work should consider how to extend the proposed method to unregistered shapes, for example with the introduction

of spatial transformer modules inside the architecture [302]. In this work, as the output of the model is binary, Dice score was adopted as reconstruction metric. However, other alternatives exist, for example by modeling the model output with a Bernoulli distribution [303], and they will be investigated in future work. Finally, the prior distribution adopted in the highest latent space is a standard Gaussian distribution $\mathcal{N}(0, 1)$: future work could consider alternative prior distributions which could further favour the clustering of shapes. Even more interestingly, the interpretability and visualisation properties of the proposed method indicate that it could constitute an interesting tool for unsupervised clustering of shapes, for example by learning in the highest level discrete random variables [304] or Gaussian mixtures [305].

5.6 Conclusions

In recent years, the medical image analysis field has witnessed a marked increase both in the construction of large-scale population-based imaging databases and in the development of automated segmentation frameworks. As a consequence, the need for novel approaches to process and extract clinically relevant information from the collected data has greatly increased. In this chapter, a method for data-driven shape analysis which enables the classification of different groups of clinical conditions through a very low-dimensional set of task-specific features is proposed. Moreover, this framework naturally enables the quantification and visualisation of the anatomical effects encoded by these features in the original space of the segmentations, making the classification task transparent. As a consequence, this method could be useful for the study of both normal anatomy and pathology in large-scale studies of volumetric imaging.

Chapter 6

3D High-Resolution Cardiac Segmentation Reconstruction from 2D Views using Conditional Variational Autoencoders

This chapter is based on: Biffi, Carlo, *et al.* 3D high-resolution cardiac segmentation reconstruction from 2D views using conditional variational autoencoders. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 1643-1646. IEEE, 2019.

6.1   Introduction

Accurate assessment of heart morphology is key for the quantitative analysis and diagnosis of cardiac pathology. CMR is the gold-standard technique for reproducible assessment of cardiac morphology and function [306], and multiplanar breath-hold 2D cine sequences

116

constitute standard clinical practice [2]. As outlined in Section 1.2.1, these protocols consist in the acquisition of a stack of breath-hold 2D image sequence in the LV short axis supplemented by long axis images in prescribed planes. Segmentation of the obtained images, more and more often fully-automated, follows image acquisition in order to derive mass and volumetric clinical indices describing the heart clinical status (Section 1.2.2). However, the disadvantages of this acquisition protocol for whole-heart assessment are low through-plane resolution, misalignment between breath-holds and lack of whole-heart coverage, resulting in segmentations not correctly representing the cardiac volume, and causing potential errors in the following analyses and research [56].

High-resolution 3D CMR sequences enable whole-heart structural imaging and, as showed in Chapter 3, proved to be crucial for the construction of integrative statistical models of cardiac anatomy and physiology and promise to further our understanding of disease characterization [96]. However, high-resolution 3D image sequences are time-consuming, expensive to acquire and they often require long breath holds that are not sustainable for patients. The increased availability of annotated large-scale medical imaging databases offers the opportunity to learn in an automated, data-driven way descriptors of anatomical shape variation. Consequently, an end-to-end method to reconstruct a 3D high-resolution segmentation from routinely-acquired 2D cine MR imaging could be highly beneficial - offering high resolution phenotyping robust to artefact in large clinical populations with conventional imaging.

6.1.1 Related Work

In the medical imaging domain, the reconstruction of 3D anatomical structures from a limited number of 2D views has been traditionally studied via the combination of image registration algorithms to align and fuse the acquired 2D views and statistical shape models to infer the final 3D anatomical structure [307, 308, 309]. However, these methods require complex, slow reconstruction procedures and are computationally-intensive.

In recent years, with the advent of learning-based approaches, and in particular of deep learning, a number of alternative strategies have been proposed in the computer vision field. One of the most employed approaches, the TL-embedding network (TL-net), consists of a 3D convolutional autoencoder (AE) which learns a low-dimensional representation of 3D geometries, whereas a second convolutional neural network connected to the AE latent space maps 2D views of the same object to the same low-dimensional representation [310]. In this way, an explicit connection between the 2D views and the 3D shape descriptor is established at training time, and 3D object shapes can be efficiently predicted at inference from their 2D views. The resulting reconstruction process is however deterministic, showing little clues on which shape parts the network considers a plausible reconstruction. Shape generation based on low-dimensional representations of deep generative models has also been largely investigated both using variational inference [225, 202, 185] and adverserial learning [311, 312] for 3D object generation and retrieval from 2D images, and potentially enabling multiple plausible shape predictions from such low-dimensional representations. Cerrolaza *et al.* [228] proposed a convolutional conditional variational autoencoder (CVAE) architecture for the 3D reconstruction of the fetal skull from 2D ultrasound standard planes of the head. Finally, Chapter 4 and 5 showed how the VAE framework can learn a shape segmentation model of LV segmentations and how the learned latent space can be exploited to accurately identify healthy and pathological cases and generate realistic segmentations unseen during training.

6.1.2   Contributions

In this chapter, a CVAE architecture that reconstructs a high-resolution 3D segmentation of the LV myocardium from three segmentations of 2D routinely acquired cardiac views (one short-axis and two long-axis views) is presented. The proposed architecture has the appealing property, unlike deterministic models, of naturally producing confidence maps associated to each reconstruction due to its generative properties.

6.2 Materials and Methods

6.2.1 3D Cardiac Image Acquisition and Segmentation

A dataset of 1,912 healthy volunteers from the UK Digital Heart Project at Imperial College London was used in this preliminary analysis. A high-spatial resolution 3D balanced steady-state free precession cine MR image sequence was acquired for each subject using a 1.5-T Philips Achieva system (Best, the Netherlands) [96]. Left and right ventricles were imaged in their entirety in a single breath-hold (60 sections, repetition time 3.0 ms, echo time 1.5 ms, flip angle 50°, field of view 320 × 320 × 112 mm, matrix 160 × 95, reconstructed voxel size 1.2 × 1.2 × 2 mm, 20 cardiac phases, temporal resolution 100 ms, typical breath-hold 20 s). CMR images were stored in an open-source database (MRIdb, Imperial College London, UK) [253]. For each subject, a 3D high-resolution segmentation of the LV was automatically obtained using a previously reported technique employing a set of manually annotated atlases [96]. In this work, only the end-diastolic (ED) frame was considered.

	Full Cohort ($N = 1,912$)	**Males** ($N = 859$)	**Females** ($N = 1,053$)
Age [*years*]	42.3 ± 13.0 (18-82)	41.0 ± 12.6 (18-75)	41.5 ± 13.4 (20-82)
Ethnicity			
Caucasian	1337 (69.7%)	601 (70.0%)	736 (69.9%)
South Asian	260 (13.6%)	126 (14.6%)	132 (12.5%)
African	195 (10.2%)	73 (8.5%)	122 (11.5%)
Other	120 (6.3%)	58 (6.8%)	62 (5.9%)
BSA [m^2]	1.8 ± 0.2	2.0 ± 0.2	1.7 ± 0.2
SBP [$mmHg$]	118 ± 14	123 ± 13	114 ± 13

Table 6.1: A summary of the 1,912 healthy subjects of the UK Digital Heart Project at Imperial College cohort employed in this project.

6.2.2 Conditional Variational Autoencoder Architecture

The outline of the proposed CVAE architecture is reported in Figure 6.1. The aim of framework is to end-to-end learn a generative model $p(\boldsymbol{x}|\boldsymbol{y})$ of 3D high-resolution LV

segmentations $\boldsymbol{x}$ conditioned on a sparse set of 2D views segmentations $\boldsymbol{y} = \{\boldsymbol{y}_i \mid i = 1, 2, 3\}$, two long axis and one short axis views in the cardiac application examined in this work. The training set of the proposed architecture is composed of a pair (X, Y) of N 3D high-resolution LV segmentations $X = \{\boldsymbol{x}_j, \ j = 1, ...N\}$ and of their corresponding 2D views segmentations $Y = \{\boldsymbol{y}_j, \ j = 1, ...N\}$. For each segmentation $\boldsymbol{x}_j$, a CVAE model learns a generative model via a p-dimensional latent representation $\boldsymbol{z}_j \in \mathcal{R}^p$, where $p \ll d$ and d is the number of pixels/voxels in a segmentation $\boldsymbol{x}_j$, as in a standard VAE. However, the encoder and decoder functions are this time conditioned on the 2D views $\boldsymbol{y}$ values [303], *i.e.* $q_\phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$ and $p_\theta(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{y})$. Therefore, the loss function to be optimized has the form:

$$\mathcal{L}_{\theta,\phi}(\boldsymbol{x}_j|\boldsymbol{y}_j) = \underset{q_\phi(\boldsymbol{z}|\boldsymbol{x}_j,\boldsymbol{y}_j)}{\mathbb{E}} [\log [p_\theta(\boldsymbol{x}_j|\boldsymbol{z}_j,\boldsymbol{y}_j)]] - \mathcal{D}_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}_j,\boldsymbol{y}_j)||p_\theta(\boldsymbol{z}_j)) = \mathcal{L}_{rec} + \mathcal{L}_{KL} \quad (6.1)$$

where $q_\phi(\boldsymbol{z}_j|\boldsymbol{x}_j, \boldsymbol{y}_j))$ is modeled also in this work as a multivariate Gaussian with a diagonal covariance structure, *i.e.* $\mathbf{z} \sim \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2\mathbf{I})$. An important choice in the construction of the encoder and decoder functions, $q_\phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$ and $p_\theta(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{y})$, is how to effectively condition them on the values $\boldsymbol{y}$. As shown in Figure 6.1, in this work a 2D convolutional neural network (CNN) is used to learn a low-dimensional representation $\boldsymbol{c}$ from the views $\boldsymbol{y}$. This latent representation is then concatenated with the output of the 3D CNN encoder, $\hat{\boldsymbol{x}}$, and used as input of a fully connected layer to learn the latent space parameter estimates $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$. Moreover, $\boldsymbol{c}$ is also concatenated with the sampled values of $\boldsymbol{z}$, and used at the input of a fully connected layer to compute $\widetilde{\boldsymbol{x}}$, which is then sent to the 3D CNN decoder to reconstruct the input segmentation $\boldsymbol{x}$. An alternative encoding strategy was proposed in [228], using a separate branch for each conditional input of the model,e.g. a separate branch for each 2D view of the same subject. However, whilst this latter approach proved efficient when the views suffer from large inconsistencies or variability (e.g., free-hand ultrasound scans), the model complexity can be notably reduced by combining the views
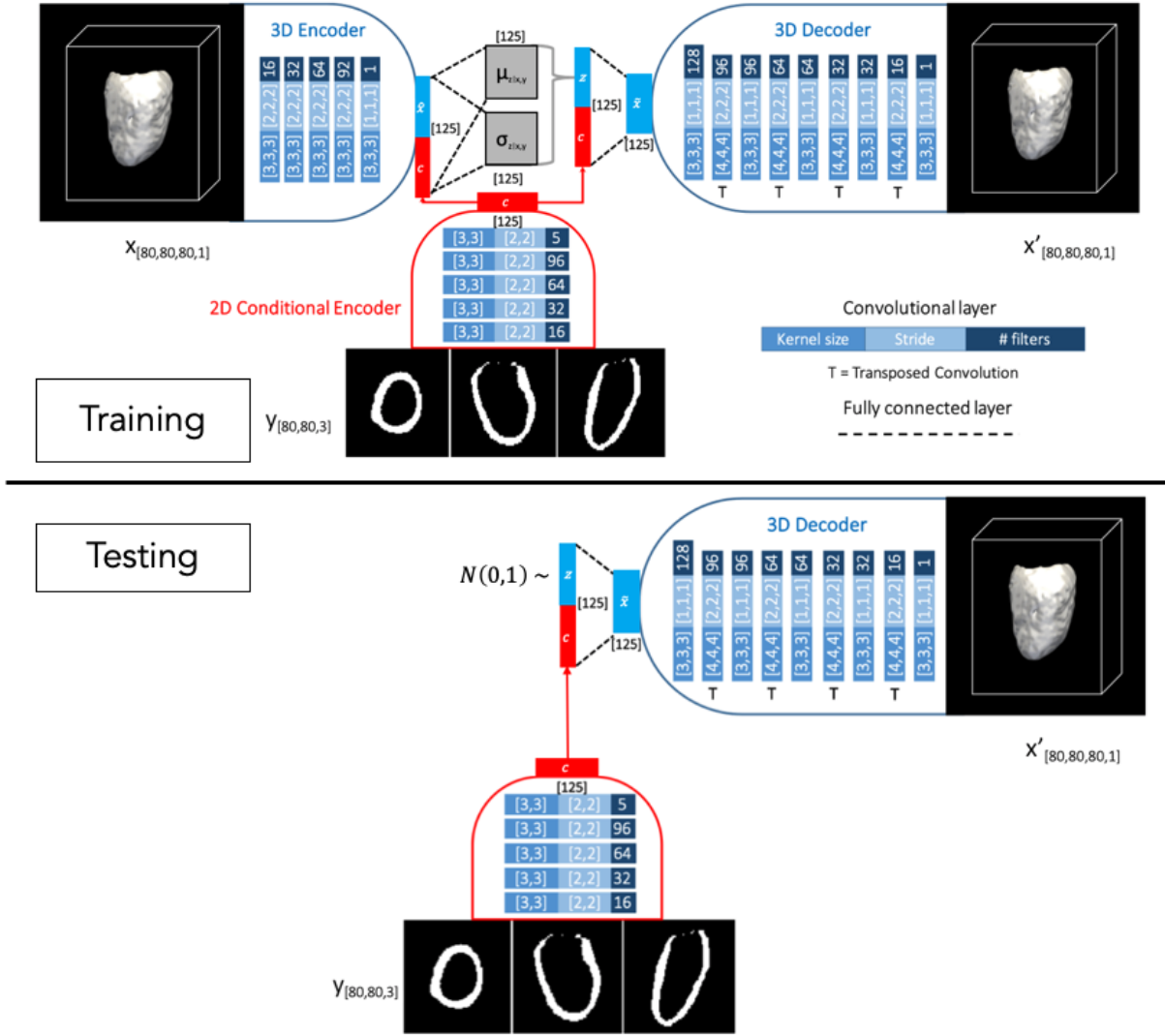
Figure 6.1: The proposed conditional variational autoencoder (CVAE) architecture.

y as a unique three-channel input as these are acquired consistently in clinical routine for this application.

The reconstruction loss L_{rec} in Equation 6.1 is computed as the Dice score (DSC) between $\boldsymbol{x}$ and its reconstruction $\boldsymbol{x}'$, which is the output of the proposed generative model. The regularisation term L_{KL} employs as prior distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{1})$, a p-dimensional normal distribution with zero mean and unit-standard deviation. Building over previous results (Chapter 5), in our implementation the two losses are weighted during training by a parameter α, $\mathcal{L} = \mathcal{L}_{rec} + \alpha \, \mathcal{L}_{KL}$. At testing phase, the 3D encoder branch is disabled.

The reconstruction of a 3D high-resolution segmentation is obtained by setting the latent variables $\boldsymbol{z} = \boldsymbol{0}$ and by concatenating this latent representation with the low-dimensional representation $\boldsymbol{c}$, which is obtained by encoding the 2D views segmentations $\boldsymbol{y}$ via the 2D conditional encoder. The obtained representation is used to computed $\widetilde{\boldsymbol{x}}$ which is then used as input of the 3D CNN decoder.

6.2.3   Experimental Setup and Network Training

In this work, the two long-axis and one short-axis segmentation views acquired in a routine CMR acquisition, $\boldsymbol{y}$, were mimicked via the following steps:

1. All the ground truth 3D high-resolution segmentations $\boldsymbol{y}$ have been rigidly aligned by performing landmark-based and subsequent intensity-based registration (nearest-neighbours interpolation).

2. Only the LV myocardium label was kept and the segmentations were cropped and padded to [x = 80, y = 80, z = 80, t = 1] dimension using a bounding box centered at the centre of mass of the LV myocardium.

3. Finally, three orthogonal views passing through the centre of each segmentation were sampled (an example is shown in Figure 6.1 as input to the 2D conditional encoder).

Thanks to this process, three 2D views segmentations showing the same three LV sections consistently for all subjects were obtained. In the following experiments, the ground truth 3D high-resolution segmentations and their corresponding 2D views were kept all in the same reference space.

The dimension p of the latent space was fixed to 125 as values smaller than 100 provided less accurate reconstruction results, while above 125 no further improvements in

reconstruction accuracy were observed. The dimensionality of the low dimensional representation c was kept equal to the dimensionality of $\mathbf{z}$ to guarantee a balanced contribution to the generative model. Simulations for different values of the parameter α in the loss function were performed: low values of α ($\alpha < 0.5$) provided better reconstruction results on the training data at the expenses of a strong deviation from normality of the latent space distribution (KL term not converging) causing poor reconstruction accuracy at testing on the evaluation set, *i.e.* overfitting. Higher values of α ($\alpha > 2$) penalized the reconstruction term in favour of a strictly normal latent space, hence also providing poorer reconstruction accuracy. In this work, α was set to 1 as this provided good reconstruction accuracy and convergence of the KL term.

Experiments were performed with different numbers of view segmentations $\boldsymbol{y}_i$ as conditions for the proposed model. In particular, referring to the first long-axis view as 1, the second long-axis view as 2 and the short-axis view as 3, the model training was performed using either only one view (which will be indicated as CVAE_1), or a combination or two views (CVAE_12, CVAE_23, CVAE_13), or all the three views (CVAE_123). The feasibility of training a 2D AE on the 3 view segmentations $\boldsymbol{y}$ to then use its encoder as a pre-trained conditional encoder (pCVAE_123) was also studied. Moreover, the reconstruction capability of the proposed architecture was compared with the one of the TL-net [310]. Finally, the reconstruction obtained by a VAE with $\mathbf{z}=0$ (VAE_0) to all our test segmentations was also compared, as this represents the best segmentation that the generative model can reconstruct when no information is provided to it (lower bound for reconstruction accuracy). Results obtained with an autoencoder (AE) are also reported since this model yielded better results than different VAEs with distinct α values as it only optimizes the reconstruction accuracy (upper bound for reconstruction accuracy). All the models share the same 3D encoder and decoder architectures.

The employed dataset was split into training, evaluation and testing sets consisting of 1362, 150 and 400 subjects respectively. Data augmentation included rotation around the

three orthogonal axis with rotation angles randomly extracted from a normal distribution $\mathcal{N}(0, 6°)$ and random closing and opening morphological operations. All the networks were implemented in Tensorflow and training was stopped after 300k iterations, when the total validation loss function had stopped improving (approximately 42 hours per network on an NVIDIA Tesla K80 GPU), using stochastic gradient descent with momentum (Adam optimizer, learning rate $= 10^{-4}$) and batch size of 8.

6.3   Results

6.3.1   Accuracy of 3D Reconstruction

Table 6.2 shows the testing reconstruction accuracy achieved by all the studied architectures in terms of 3D Dice score and 2D slice-by-slice Hausdorff distance between the ground truth 3D high-resolution segmentations and their reconstruction by the network. The reported results indicate that the reconstruction accuracy decreases (lower Dice Score, higher Hausdorff distance) when 2D views segmentations are removed. From the experiments with only two views it can also be inferred how different views have different importance. In particular, the short-axis view seems to have the smallest impact on the reconstruction accuracy. This could be motivated by the fact that the long-axis views contain more information about the regional changes in curvature of the LV, which strongly influences the Dice Score. The results reported in Table 6.2 also show how the proposed architecture significantly outperforms the TL-net by a large amount ($p = 2.2 \cdot 10^{-16}$), and how the pre-training of the 2D CNN encoder network did not help to achieve better results.

LV mass is an important clinical biomarker for diagnosis of many cardiovascular conditions, therefore for each reconstruction its percentage difference in mass with the ground truth was also estimated as $100 \cdot \frac{m(\boldsymbol{x}') - m(\boldsymbol{x})}{m(\boldsymbol{x})}$, where $m()$ sums the number of voxels labelled

Model	DSC	Hausd. [mm]	MassDiff [%]
VAE_0	65.48 ± 0.38	9.32 ± 0.06	35.37 ± 0.70
CVAE_1	78.08 ± 0.33	5.29 ± 0.04	3.94 ± 0.38
CVAE_23	82.90 ± 0.21	4.43 ± 0.04	3.93 ± 0.19
CVAE_12	85.21 ± 0.20	4.46 ± 0.04	3.73 ± 0.19
CVAE_13	83.18 ± 0.18	4.77 ± 0.04	3.69 ± 0.19
CVAE_123	**87.92 ± 0.15**	**3.99 ± 0.03**	**2.70 ± 0.14**
pCVAE_123	87.63 ± 0.16	4.04 ± 0.04	3.05 ± 0.16
TL_net	82.60 ± 0.23	4.66 ± 0.04	3.85 ± 0.19
AE	90.45 ± 0.12	3.46 ± 0.03	1.50 ± 0.10

Table 6.2: Reconstruction metrics together with their standard error of the mean for all the studied models. The Dice score (DSC) and the Hausdorff distance are reported together with the LV mass difference.

as LV myocardium in a segmentation. It can be observed that its trend is in agreement with the other two studied metrics and that the mass difference is systematically overestimated by a small amount, which decreases with the number of views provided. This may be linked to the choice of the reconstruction loss. In fact, we observed that models trained using cross entropy in the loss function yielded a systematic underestimation of the mass, often with reconstructions with missing LV apex, as this loss term tends to favour the background instead of the myocardium, while model trained with Dice score yielded an opposite trend, often overestimating the LV apex.

6.3.2   Visualisation and Uncertainty Estimation

In the first and third rows of Figure 6.2 the reconstructed segmentations obtained with one and three views (in red) are reported and overlaid onto the ground truth segmentation (in black) for one subject of the testing dataset (with Dice score 0.80 and 0.89, respectively). In the second and fourth rows are instead reported the confidence maps obtained for the reconstruction with one and three views - $P(1v)$ and $P(3v)$. These maps have been obtained by sampling N times ($N = 1,000$) $\mathbf{z}$ from $\mathcal{N}(\mathbf{0}, \mathbf{1})$ to reconstruct N segmentations from the same set of views $\mathbf{Y}$. Unlike deterministic architectures (such as the TL-net),
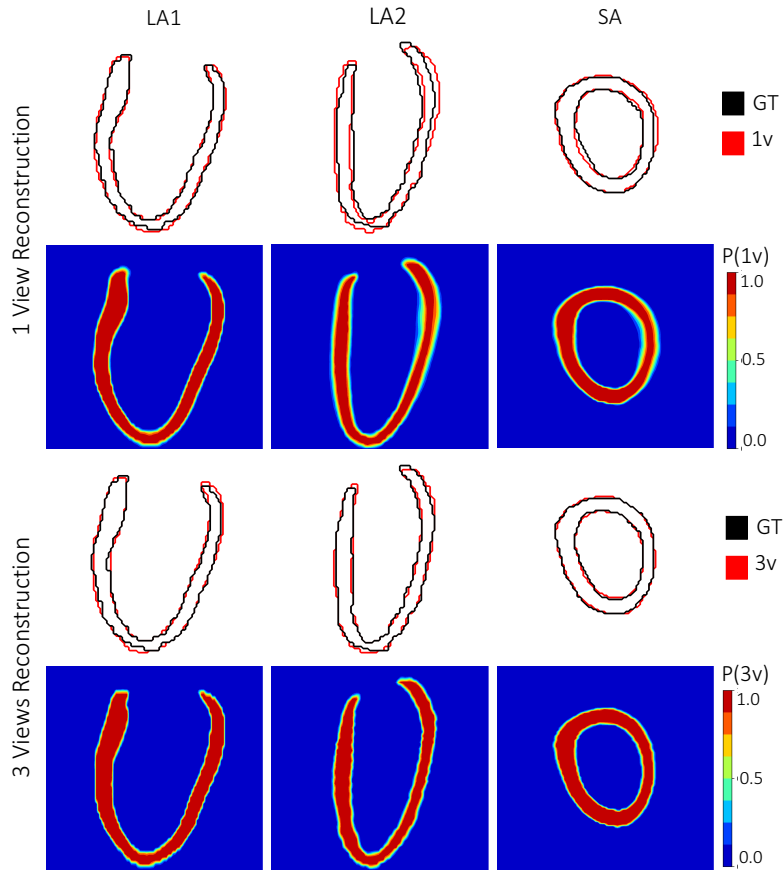
Figure 6.2: First and third rows, reconstructed segmentation obtained with one and three views (in red, 1v and 3v) overlaid onto the ground truth segmentation (in black, GT) for one random subject. Second and fourth rows, confidence maps for the reconstruction with one and three views - $P(1v)$ and $P(3v)$. First and second columns, long-axis views (LA1 and LA2). Third column, short-axis (SA) view.

by averaging these maps the probability of each voxel to be labelled as LV myocardium can be computed, providing to clinicians a richer and more intuitive interpretation of the reconstruction. It can be seen in Figure 6.2 how the confidence map obtained with only 1 view has greater uncertainty than the one obtained with 3 views, which instead shows lower variability. Moreover, the amount of uncertainty in the $P(1v)$ map for the long-axis view 1 is less than for the other two views, as this view was the one provided to the network as condition. Interestingly, in the reconstruction with one view the areas with more uncertainty correspond to the areas where there is less overlap with the ground truth, i.e. the areas where the network is less accurate in predicting the shape.

6.4 Discussion

In this preliminary work, building on the evidence gathered in previous chapters, a simple conditional generative model is introduced to model anatomical 3D high-resolution LV segmentations conditioned on a sparse set of routinely acquired 2D views segmentations. The proposed framework can model shape variations using a low-dimensional latent space and, by conditioning on a conditional vector representing the anatomical variation encoded by the views segmentations, can obtain accurate high-resolution 3D reconstructions by just using such views as input. Furthermore, a simple procedure able to provide reconstruction uncertainty is proposed.

Concurrently with our work, CVAEs have also been used to generate hippocampal shapes conditioned on clinical data [189]. Similarly to this work, shape variation was modeled with a low-dimensional VAE representation, while shapes were parameterised as point clouds in this case. CVAE based methods were also proposed to generate images of the whole fetus from a large number of overlapping image patches [313] and electrical activity maps of the heart conditional to body surface potentials [314], further showing the potential of these models for conditional medical data generation. In future work, this project will focus on benchmarking the proposed model using real standard long-axis views (instead of the simulated ones in this work) with the aim of identifying the smaller sparse set of 2D views necessary to reconstruct with high accuracy the high-resolution segmentations of interest. Invariance towards inter-subject pose variability is a desidered property for the architecture under exam, and currently an important limitation of the presented work, and will be investigated in future work. This can be potentially addresses with a tailored data augmentation strategy, the introduction of spatial transformed modules [302] or by employing other shape parametrisations, such as point clouds [189, 315], which afford for neural network architectures which do not require aligned inputs. Additional direction for future work include reconstructing multiple structures at the same time and extending the proposed framework to pathological datasets, for which acquiring

breath-hold sequences is even more challenging and hence the method can be even more beneficial. Finally, based on very recent advances in shape reconstruction and generative modeling, better generative models architectures, such as the β-VAE [316] or the LVAE [296], should to be investigated in future research together with image-based methods for 3D object reconstruction methods from the computer vision field [225, 226].

6.5   Conclusion

In this chapter, the first deep conditional generative network for the reconstruction of 3D high-resolution LV segmentations from three segmentations of 2D orthogonal views was presented. The reported results showed the potential of this class of models to provide quantitative cardiac models from sparse imaging acquisitions.

Chapter 7

Conclusion and Future Work

In this chapter, a summary of the research aims of this thesis is firstly presented. The achieved methodological contributions are then listed item by item. Finally, the current limitations of the proposed algorithms and interesting possible directions for future research are proposed.

7.1 Summary

The current understanding of the epidemiological risk factors and of the progression of cardiovascular conditions has largely been inferred from population-based studies [37, 39, 38]. The recent improvements in non-invasive imaging modalities and semi- and fully- automated image processing pipelines, together with advances in high-performance computing and big data analytics, have made large-scale population-based imaging studies of the heart feasible [41, 55, 35]. These improvements enable the assessment of heart function and morphology in large cohorts through relatively fast derivation of physiological clinical indices. However, currently adopted indices are largely insensitive to complex modifications in heart morphology, especially at the earliest disease stages [81, 82, 83]. Machine learning approaches showed great promise in addressing big data challenges in

cardiovascular research tasks. This is mainly due to their excellent capability at feature extraction and at learning highly-complex functions modeling a specific task under study [77, 74, 51, 75]. However, machine learning approaches seldom result in image features and biomarkers that can be readily understood by a clinician, and to explain why a specific prediction has been made by a deep model is often difficult. This poses the need for the development of novel methodologies that can successfully employ learning-based strategies or advanced statistical modeling while, at the same time, being explainable [76].

7.2   Achievements

In order to address the aforementioned need for model explainability, all the methodologies proposed in this work afford for the direct visualisation, on a template three-dimensional anatomy, of the learned anatomical remodelling patterns predictive of a disease status or associated with a genetic variation. The main methodological achievements of this PhD project can be summarized as follows:

3D cardiovascular imaging-genetics.   This work demonstrated how regional phenotype-genotype associations can be discovered and visualised via mass univariate analysis of 3D cardiac atlases encoding multiple phenotypic traits from MR imaging (Chapter 3, [260]). An R package providing all the necessary functions to perform this class of studies is also proposed. The R package is validated on synthetic and real genetic data, and the reported results show how the proposed framework enables the detection of small, regional patterns of remodelling with relatively small sample sizes. In particular, in the reported GWAS replication study, the effect of four SNPs previously discovered and never replicated in a GWAS for LV mass was confirmed by the proposed approach, while none of the SNPs was found significant with conventional LV mass analysis.

Learning generative *and* discriminative deep features of cardiac shapes. This thesis demonstrated how automatic classification of anatomical shapes associated with morphological remodelling can be performed from a generative latent space learned by shape auto-encoding. Concurrent classification network training enables the learning of generative and discriminative anatomical shape features. Either by using an iterative latent space navigation (Chapter 4, [297]) procedure or by the sampling from the posterior distribution of the shape classes (Chapter 5, [317]), the results reported in this work show the feasibility of visualising in three-dimension the anatomical effect encoded by learned generative and discriminative shape features. These approaches provide an interpretable classifier of anatomical shapes which achieves high accuracy in the categorisation of healthy and remodelled left ventricles of the heart, and on distinguishing hippocampi from healthy controls and from patients with Alzheimer's disease.

Modeling shape variation in large-scale imaging databases using deep hierarchical generative models. This work introduces a deep learning approach, based on state-of-the-art deep hierarchical generative models, to model a large database of anatomical shapes through a hierarchy of conditional latent variables (Chapter 6, [318]). The proposed approach can learn the most distinctive morphological differences between different clinical populations at the top of the hierarchy, while the subsequent levels in the hierarchy are left to model the other factors of shape variation. Moreover, this latent space can be made two- or three- dimensional, enabling the direct visualisation of the classification space and the estimation of the posterior distribution of the shape classes. Sampling from this posterior distribution affords for the visualisation on a template shape of the most distinctive morphological shape features of the different clinical conditions under exam, hence enabling automated, data-driven population-based inferences.

Learning to generate 3D high-resolution cardiac shape from 2D views. In a preliminary study, this thesis introduces a first, simple deep conditional generative

model architecture able to reconstruct 3D high-resolution LV segmentations from 2D LV segmentations of one short-axis and two long-axis routinely acquired images. The proposed model was evaluated on unseen healthy volunteers, outperforming competing architectures, and could allow for automated reconstruction of three-dimensional high-resolution cardiac phenotypes from routine acquisitions.

7.3   Limitations and Future Work

One of the main focuses of this work is to be able to process a large database of shapes with deep neural networks, aiming at modeling shape variation within a population. A crucial point in shape analysis studies relies in the choice of the shape parametrization to be employed. In the literature, multiple shape representations have been previously explored and concurrently to this work, in the last years, several approaches aiming at processing shape with deep learning methods have been proposed. In this work, a volumetric representation of cardiac shape was adopted, *i.e* a probabilistic distribution of binary variables on a 3D voxel grid. This enabled the full exploitation of 3D convolutional and up-convolutional layers properties, which make deep learning architectures such powerful and efficient feature extractors. However, one important drawback of adopting such volumetric representations is their memory and computational requirements which increase cubically with the input resolution [319]. Shape analysis approaches on point clouds could guarantee a significant reduction of such computational requirements [186, 320]. In the analysis of pathologies of the myocardium, regional changes in volume are of main interest, and point clouds could provide an effective way to study such shape changes, being the classification task primarily interested in regional contraction or expansion of the myocardial wall. Furthermore, deep learning methods on point clouds could not require co-registered input shapes for model training [189, 186, 315], currently another limitation of the proposed approaches. Specific data augmentation strategies in combination with spatial transformer modules [302] to align all the input shapes to a canonical space could

alternatively constitute a viable shortcoming for this limitation.

Geometric deep learning approaches have been introduced to generalize convolution operations to non-Euclidean domains, on which medical shapes naturally live in contrast to 2D/3D images, and to enable learning on graph structured data [321, 195]. An interesting direction for future work consists in encoding at each vertex of a subject mesh multiple 3D structural information together with motion and tissue data. This would enable the construction of a comprehensive representation of the cardiac phenotype as a graph that can be processed with graph CNNs. Many classification and survival prediction tasks are expected to benefit from this richer representation of the cardiac phenotype. The methods based on shape auto-encoding reported in this work could be applied to this new input data by employing graph CNNs in place of regular convolutions. Distance metric learning approaches aiming at separate subjects from different classes and at obtaining population-level saliency maps, which have been recently applied to analyse functional brain networks [322, 323], should also be explored.

In Chapter 5, two common and simple classification tasks were investigated to evaluate the proposed model properties for explainable shape analysis. However, additional work is required to fully explore the full model potential on more complex shape classification tasks. A first and more challenging application could consist in deploying the model to characterize distinct disease subgroups by their clinical endpoints. In this scenario, the database of shapes under exam is expected to be more heterogeneous and the shape changes more challenging to be captured. Secondly, the deployment of the proposed models for end-to-end unsupervised clustering of shapes is another interesting direction of future work. This latter could be achieved by adopting as prior distribution discrete random variables [304] or a Gaussian mixture [305] in the highest latent space. Both the aforementioned directions could provide better patient stratification rules than currently employed global indices, and help the discovery of homogeneous disease subgroups in an automated, data-driven way, while also enabling the visualisation of their morphologi-

cal characteristics. Moreover, it is unrealistic to expect that the proposed models could be employed in the clinical practice without incorporating the information provided by conventional clinical covariates and other non-imaging measurements. Therefore, a relevant direction of investigation consists in being able to include such information, for example by conditioning the latent spaces [324, 325]. This would help the model learn better, prognostically discriminative remodelling patterns, and potentially improve both its sensitivity and specificity.

An end-to-end learned and explainable low dimensional representation of 3D heart shape could more effectively capture the heart phenotype and better inform imaging-genetics studies than simple volumetric measurements. Thus, the application of autoencoded shape representations as novel imaging-derived metrics is another interesting direction for future work, including their employment in multi-trait genome-wide association studies. This class of studies is known for being currently lacking of statistical power, which has not only been attributed to the massive multiple testing correction they require, but also to phenotyping resolution issue [69]. Being able to capture disease-related high-resolution anatomical variation into a low-dimensional representation, and to link it to genetic variables thus constitutes an interesting future application.

Finally, the presented imaging-genetics framework requires high-spatial resolution CMR, which is not available in all populations. Also, the studied models for interpretable shape analysis presented in this work required to process their inputs with up-sampling and motion correction procedures to enhance the quality of segmentations to be studies. Starting from the evidence collected in this work about the feasibility of end-to-end learning of generative shape models, in Chapter 6 a first generative model able to generate such high-resolution phenotypic representation from standard acquisitions was introduced. However, further analysis is required to investigate the feasibility and accuracy of this approach on large-scale imaging cohorts. This includes the training of the proposed approach on real views segmentations and/or images, and testing more sophisticated generative models

and decoder networks [326, 316, 189]. For this latter objective, being able to exploit more computationally efficient shape representations requires additional investigations, and informing the generative process by exploiting non-imaging measurements routinely acquired in the clinical practice should also be considered in this context.

7.4 Conclusion

An important challenge to be addressed to fully exploit the potential of large-scale imaging studies of the heart relies in the development of novel, automated data-driven methods to extract new clinical insights from this unprecedented volume of big heart data. The approaches proposed in this work show that machine learning methodologies can facilitate high-throughput analysis of normal and pathological anatomy of the heart, and of its determinants, without losing clinical interpretability.

List of Publications

Journal Publications

- **C. Biffi**, A. de Marvao, M.I. Attard, T.J.W. Dawes, N. Whiffin, W. Bai, W. Shi, C. Francis, H. Meyer, R. Buchan, S.A. Cook, D. Rueckert, D. P. O'Regan. *Three-dimensional cardiovascular imaging-genetics: A mass univariate framework.* Bioinformatics 34, no. 1 (2018): 97-103.

- **C. Biffi**, J.J. Cerrolaza, G. Tarroni, W. Bai, A. De Marvao, O. Oktay, C. Ledig, L. Le Folgoc, K. Kamnitsas, G. Doumou, J. Duan, S.K. Prasad, S.A. Cook, D.P. O'Regan, D. Rueckert. *Explainable anatomical shape analysis through deep hierarchical generative models.* IEEE Transactions on Medical Imaging (2020). In Print.

- S. Schafer, A. De Marvao, E. Adami, L. R Fiedler, B. Ng, E. Khin, O.J.L. Rackham, S. Van Heesch, C. J. Pua, M. Kui, R. Walsh, U. Tayal, S. K. Prasad, T.J.W. Dawes, N.S.J. Ko, D. Sim, L.L.H. Chan, C. W.L. Chin, F. Mazzarotto, P.J. Barton, F.Kreuchwig, D.P.V. De Kleijn, T. Totman, **C. Biffi**, N. Tee, D. Rueckert, V. Schneider, A. Faber, V. Regitz-Zagrosek, J. G. Seidman, C.E. Seidman, W.A. Linke, J. Kovalik, D.P. O'Regan, J.S. Ware, N. Hubner, S.A. Cook. *Titin-truncating variants affect heart function in disease cohorts and the general population.* Nature Genetics 49, no. 1 (2017): 46.

- J. Duan, G. Bello, J. Schlemper, W. Bai, T.J.W. Dawes, **C. Biffi**, A. de Marvao,

G. Doumoud, D.P. O'Regan, D. Rueckert. *Automatic 3D bi-ventricular segmentation of cardiac images by a shape-refined multi-task deep learning approach.* IEEE Transactions on Medical Imaging 38, no. 9 (2019): 2151-2164.

- G.A. Bello, T.J.W. Dawes, J. Duan, **C. Biffi**, A. De Marvao, L.S.G.E. Howard, J.S.R. Gibbs, M.R. Wilkins, S.A. Cook, D. Rueckert, D.P. O'Regan. *Deep-learning cardiac motion analysis for human survival prediction.* Nature Machine Intelligence 1, no. 2 (2019): 95-104.

- M.I. Attard, T.J.W. Dawes, A. de Marvao, **C. Biffi**, W. Shi, J. Wharton, C.J. Rhodes, P. Ghataorhe, J.S.R. Gibbs, L.S.G.E. Howard, D. Rueckert, M.R. Wilkins, D.P. O'Regan. *Metabolic pathways associated with right ventricular adaptation to pulmonary hypertension: 3D analysis of cardiac magnetic resonance imaging.* European Heart Journal-Cardiovascular Imaging 20, no. 6 (2019): 668-676.

- A.N. Bhuva, T.A. Treibel, A. De Marvao, **C. Biffi**, T.J.W. Dawes, G. Doumou, W. Bai, K. Patel, R. Boubertakh, D. Rueckert, D.P. O'Regan, A.D. Hughes, J.C. Moon, C.H. Manisty. *Sex and regional differences in myocardial plasticity in aortic stenosis are revealed by 3D model machine learning.* European Heart Journal-Cardiovascular Imaging 21, no. 4 (2020): 417-427.

Conference Publications

- **C. Biffi**, O. Oktay, G. Tarroni, W. Bai, A. De Marvao, G. Doumou, M. Rajchl, R. Bedair, S. Prasad, S. Cook, D. O'Regan, D. Rueckert. *Learning interpretable anatomical features through deep generative models: Application to cardiac remodeling.* In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 464-471. Springer, Cham, 2018.

- **C. Biffi**, J.J. Cerrolaza, G. Tarroni, A. de Marvao, S.A. Cook, D.P. O'Regan, D.

Rueckert. *3D high-resolution cardiac segmentation reconstruction from 2D views using conditional variational autoencoders.* In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 1643-1646. IEEE, 2019.

- J.J. Cerrolaza, Y. Li, **C. Biffi**, A. Gomez, J. Matthew, M. Sinclair, C. Gupta, C.L. Knight, D. Rueckert. *Fetal Skull Reconstruction via Deep Convolutional Autoencoders.* In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 887-890. IEEE, 2018.

- J.J. Cerrolaza, Y. Li, **C. Biffi**, A. Gomez, J. Matthew, M. Sinclair, C. Gupta, C.L. Knight, B. Kainz, D. Rueckert. *"3d fetal skull reconstruction from 2dus via deep conditional generative networks.* In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 383-391. Springer, Cham, 2018.

- J. Duan, J. Schlemper, W. Bai, T. JW Dawes, G. Bello, **C. Biffi**, G. Doumou, A. De Marvao, D. P O'Regan, D. Rueckert. *Combining deep learning and shape priors for bi-ventricular segmentation of volumetric cardiac magnetic resonance images.* In International Workshop on Shape in Medical Imaging (pp. 258-267). Springer, Cham.

- C. Chen, **C. Biffi**, G. Tarroni, S. Petersen, W. Bai, D. Rueckert. *Learning shape priors for robust cardiac MR segmentation from multi-view images.* In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 523-531. Springer, Cham, 2019.

- O. Cheng, K. Kamnitsas, **C. Biffi**, J. Duan, and D. Rueckert. *Data efficient unsupervised domain adaptation for cross-modality image segmentation.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 669-677. Springer, Cham, 2019.

- J. Duan, J. Schlemper, C. Qin, C. Ouyang, W. Bai, **C. Biffi**, G. Bello, B. Statton,

D.P. O'Regan, D. Rueckert. *VS-Net: Variable splitting network for accelerated parallel MRI reconstruction.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 713-722. Springer, Cham, 2019.

Appendix A

I report here additional visual results of the method proposed in Chapter 5 for the cardiac application. Figures 7.1 and 7.2 illustrate the anatomical variability encoded by the highest space z_3. Figures 7.3, 7.4, 7.5 and 7.6, 7.7 show additional reconstruction examples of HVol and HCM segmentations.
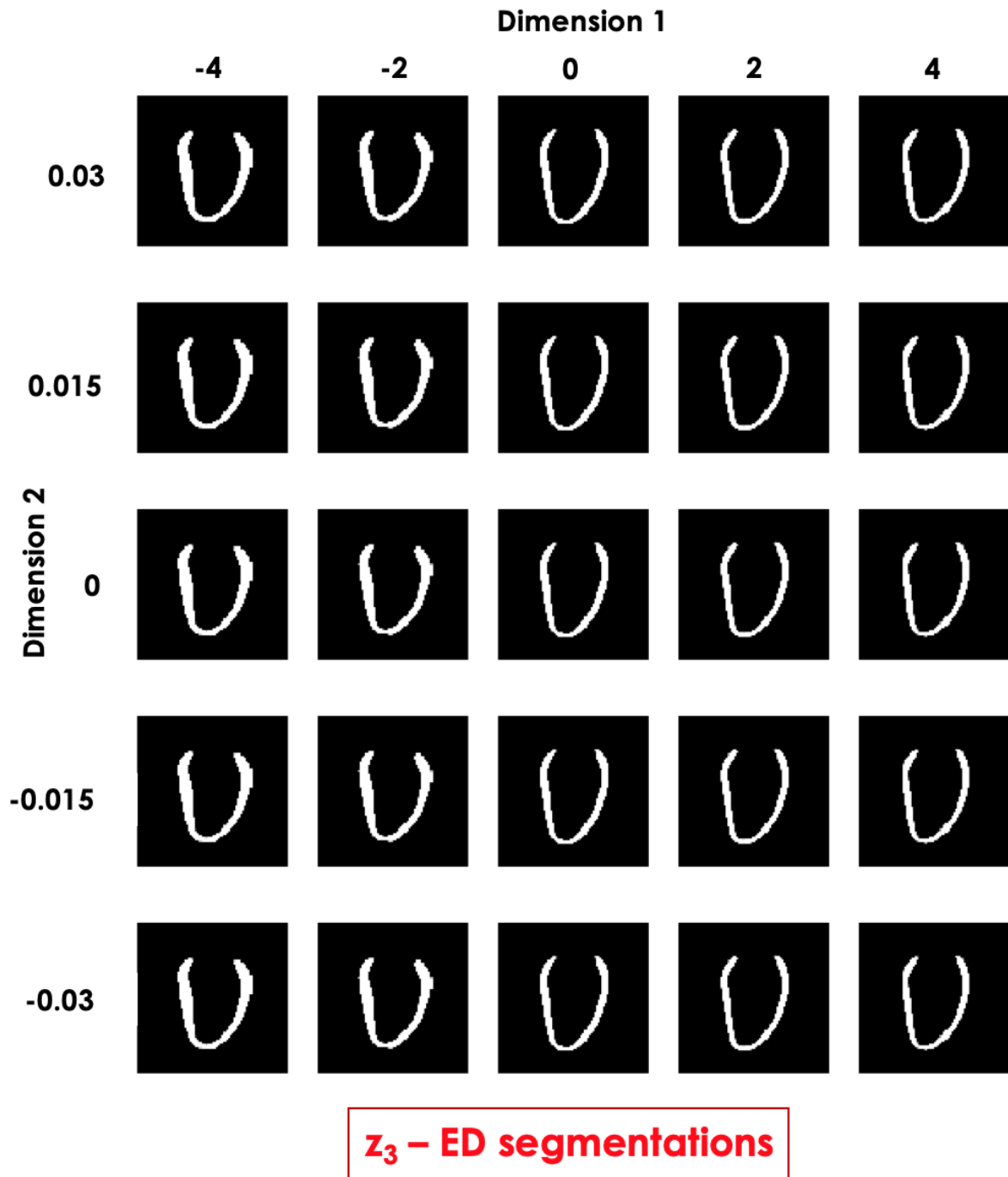
Figure 7.1: Long-axis section of reconstructed segmentations at ED by the LVAE+MLP model by sampling from different points in z_3 and subsequently from the prior distribution of the latent variables z_2 and z_1.
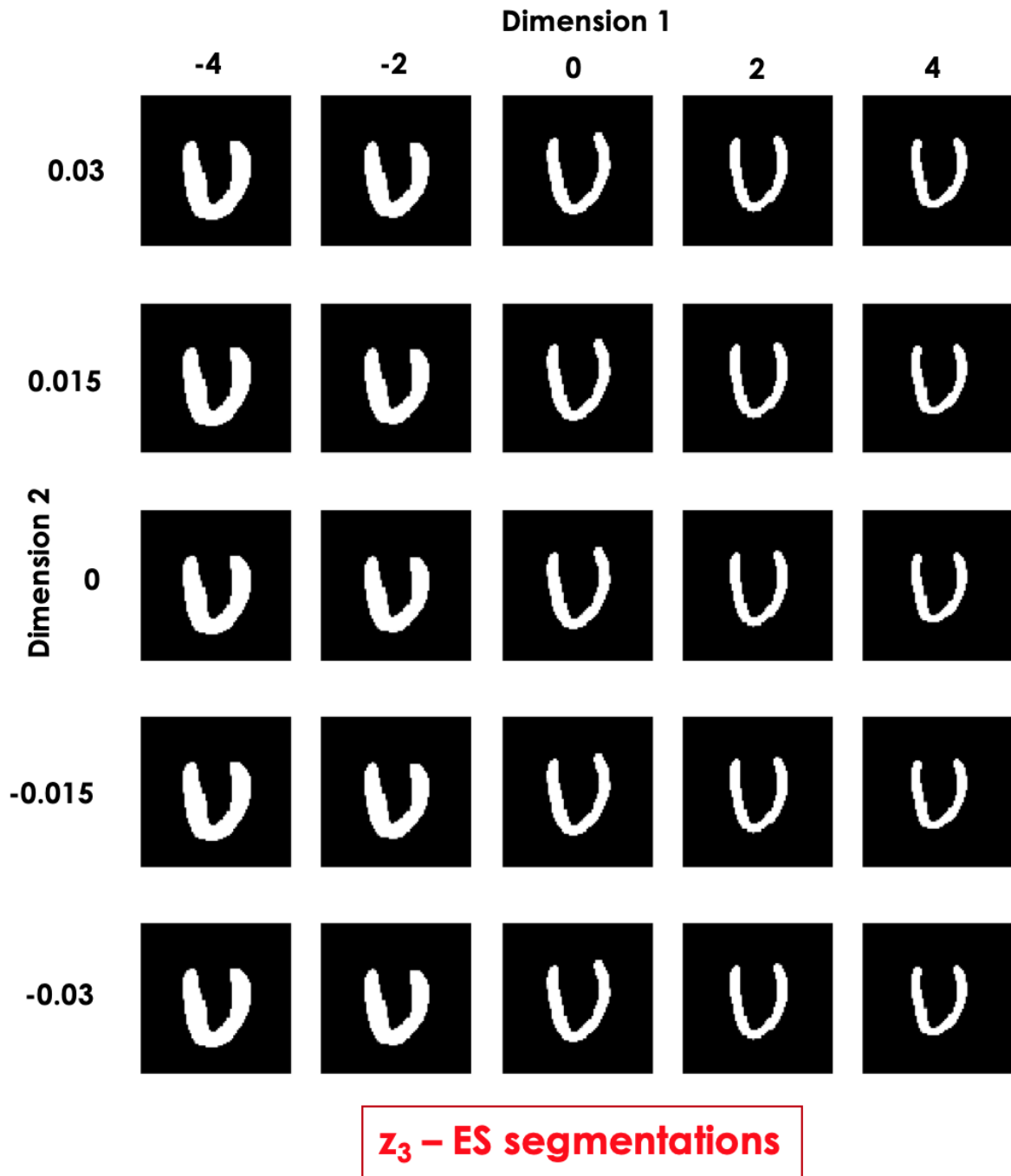
Figure 7.2: Long-axis section of reconstructed segmentations at ES by the LVAE+MLP model by sampling from different points in z_3 and subsequently from the prior distribution of the latent variables z_2 and z_1.

Figure 7.3: Long-axis section of an healthy subject reconstructed segmentations at ED and ES by the LVAE+MLP model using only z_3 information (first column) or also using the posterior information of the other latent spaces (z_2, z_1) . Last column: ground-truth (GT) segmentation. DSC = Dice Score between the segmentation at that column and the GT.
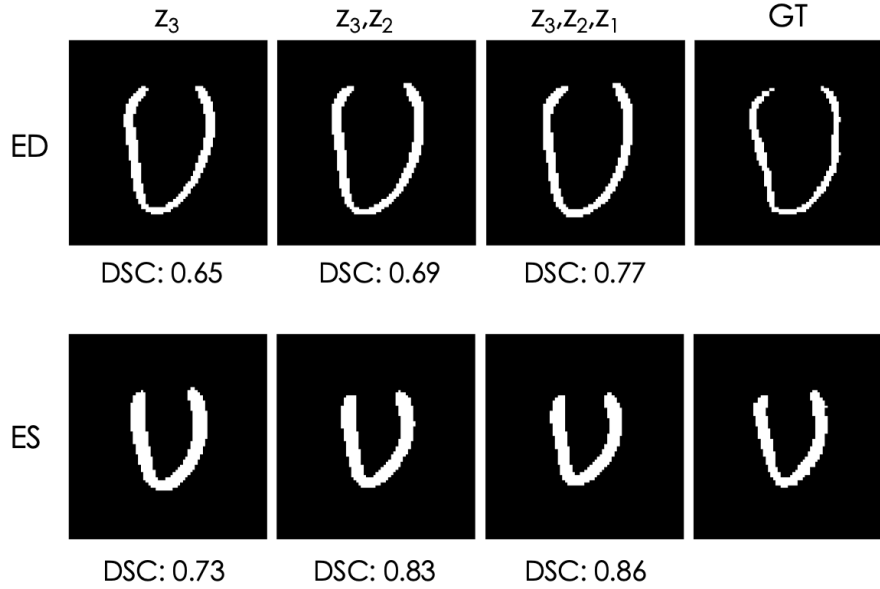
Figure 7.4: Long-axis section of an healthy subject reconstructed segmentations at ED and ES by the LVAE+MLP model using only z_3 information (first column) or also using the posterior information of the other latent spaces (z_2, z_1) . Last column: ground-truth (GT) segmentation. DSC = Dice Score between the segmentation at that column and the GT.

143
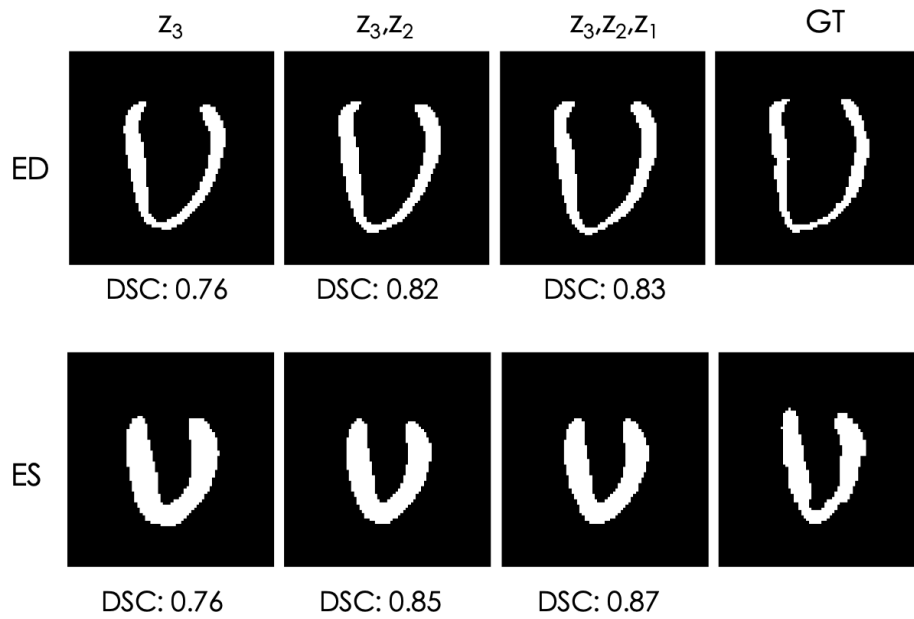
Figure 7.5: Long-axis section of an HCM patient reconstructed segmentations at ED and ES by the LVAE+MLP model using only z_3 information (first column) or also using the posterior information of the other latent spaces (z_2, z_1) . Last column: ground-truth (GT) segmentation. DSC = Dice Score between the segmentation at that column and the GT.
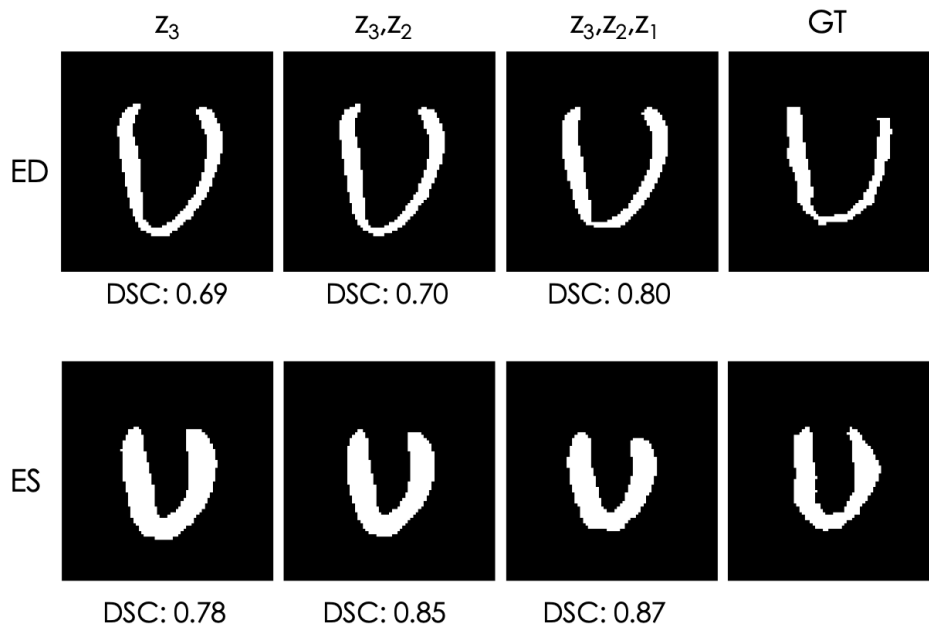
Figure 7.6: Long-axis section of an HCM patient reconstructed segmentations at ED and ES by the LVAE+MLP model using only z_3 information (first column) or also using the posterior information of the other latent spaces (z_2, z_1) . Last column: ground-truth (GT) segmentation. DSC = Dice Score between the segmentation at that column and the GT.
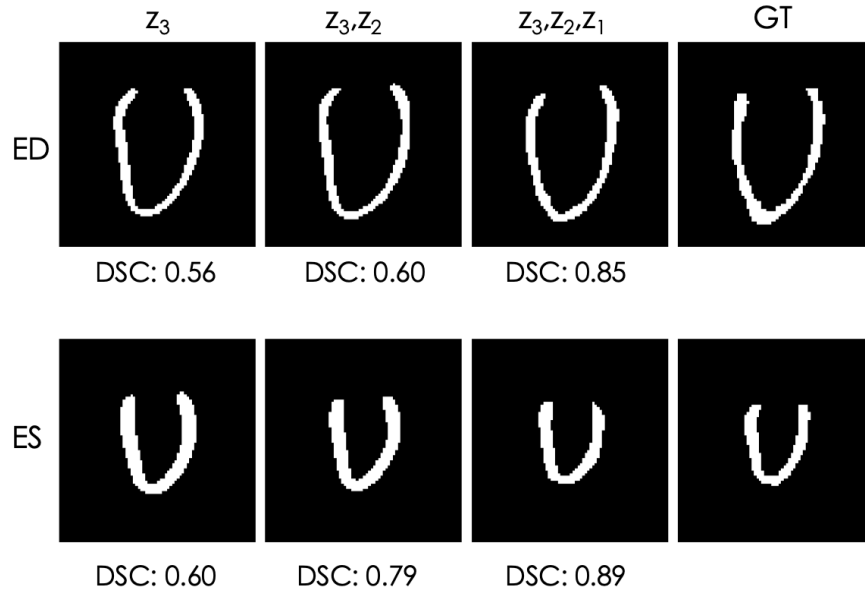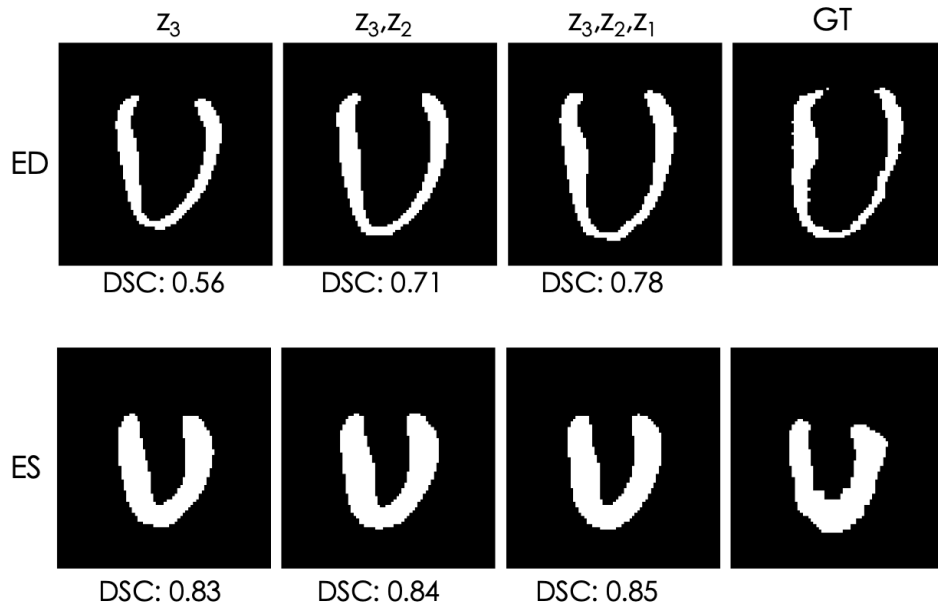
Figure 7.7: Long-axis section of an HCM patient reconstructed segmentations at ED and ES by the LVAE+MLP model using only z_3 information (first column) or also using the posterior information of the other latent spaces (z_2, z_1) . Last column: ground-truth (GT) segmentation. DSC = Dice Score between the segmentation at that column and the GT.

Bibliography

[1] Siew Yen Ho. Anatomy and myoarchitecture of the left ventricular wall in normal and in disease. *European Journal of Echocardiography*, 10(8):iii3–iii7, 2009.

[2] Victor Ferrari. *The EACVI Textbook of Cardiovascular Magnetic Resonance*. Oxford University Press, 2018.

[3] José Luis Zamorano, Jeroen Bax, Juhani Knuuti, Patrizio Lancellotti, and Luigi Badano. *The ESC textbook of cardiovascular imaging*. Oxford University Press, 2015.

[4] American Heart Association Writing Group on Myocardial Segmentation, Registration for Cardiac Imaging:, Manuel D Cerqueira, Neil J Weissman, Vasken Dilsizian, Alice K Jacobs, Sanjiv Kaul, Warren K Laskey, Dudley J Pennell, John A Rumberger, Thomas Ryan, et al. Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart: a statement for healthcare professionals from the Cardiac Imaging Committee of the Council on Clinical Cardiology of the American Heart Association. *Circulation*, 105(4):539–542, 2002.

[5] Eric J Topol and Paul S Teirstein. *Textbook of Interventional Cardiology E-Book*. Elsevier Health Sciences, 2015.

[6] WHO. Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds), 2016. [Online; accessed 19-February-2020].

[7] Gianluigi Savarese and Lars H Lund. Global public health burden of heart failure. *Cardiac failure review*, 3(1):7, 2017.

[8] Barry J Maron, Jeffrey A Towbin, Gaetano Thiene, Charles Antzelevitch, Domenico Corrado, Donna Arnett, Arthur J Moss, Christine E Seidman, and James B Young. Contemporary definitions and classification of the cardiomyopathies: an American Heart Association scientific statement from the council on clinical cardiology, heart failure and transplantation committee; quality of care and outcomes research and functional genomics and translational biology interdisciplinary working groups; and council on epidemiology and prevention. *Circulation*, 113(14):1807–1816, 2006.

[9] Eloisa Arbustini, Navneet Narula, Luigi Tavazzi, Alessandra Serio, Maurizia Grasso, Valentina Favalli, Riccardo Bellazzi, Jamil A Tajik, Robert O Bonow, Valentin Fuster, et al. The MOGE (S) classification of cardiomyopathy for clinicians. *Journal of the American College of Cardiology*, 64(3):304–318, 2014.

[10] Barry J Maron. Hypertrophic cardiomyopathy: a systematic review. *Jama*, 287(10):1308–1320, 2002.

[11] Barry J Maron, William J McKenna, Gordon K Danielson, Lukas J Kappenberger, Horst J Kuhn, Christine E Seidman, Pravin M Shah, William H Spencer, Paolo Spirito, Folkert J Ten Cate, et al. American College of Cardiology/European Society of Cardiology clinical expert consensus document on hypertrophic cardiomyopathy: a report of the American College of Cardiology foundation task force on clinical expert consensus documents and the European Society of Cardiology committee for practice guidelines. *Journal of the American College of Cardiology*, 42(9):1687–1713, 2003.

[12] Arthur AM Wilde and Elijah R Behr. Genetic testing for inherited cardiac disease. *Nature Reviews Cardiology*, 10(10):571, 2013.

[13] Ahmed A Alfares, Melissa A Kelly, Gregory McDermott, Birgit H Funke, Matthew S Lebo, Samantha B Baxter, Jun Shen, Heather M McLaughlin, Eugene H Clark, Larry J Babb, et al. Results of clinical genetic testing of 2,912 probands with hypertrophic cardiomyopathy: expanded panels offer limited additional sensitivity. *Genetics in Medicine*, 17(11):880–888, 2015.

[14] Barry J Maron, Martin S Maron, and Christopher Semsarian. Genetics of hypertrophic cardiomyopathy after 20 years: clinical perspectives. *Journal of the American College of Cardiology*, 60(8):705–715, 2012.

[15] Christine E Seidman and JG Seidman. Identifying sarcomere gene mutations in hypertrophic cardiomyopathy: a personal history. *Circulation research*, 108(6):743–750, 2011.

[16] Bernard J Gersh, Barry J Maron, Robert O Bonow, Joseph A Dearani, Michael A Fifer, Mark S Link, Srihari S Naidu, Rick A Nishimura, Steve R Ommen, Harry Rakowski, et al. 2011 ACCF/AHA guideline for the diagnosis and treatment of hypertrophic cardiomyopathy: a report of the American College of Cardiology Foundation/American Heart Association Task Force on practice guidelines developed in collaboration with the American Association for Thoracic Surgery, American Society of echocardiography, American Society of nuclear Cardiology, Heart Failure Society of America, Heart Rhythm Society, Society for Cardiovascular Angiography and Interventions, and Society of Thoracic Surgeons. *Journal of the American College of Cardiology*, 58(25):e212–e260, 2011.

[17] M Elliott Perry et al. ESC guidelines on diagnosis and management of hypertrophic cardiomyopathy: the Task Force for the Diagnosis and Management of Hypertrophic Cardiomyopathy of the European Society of Cardiology (ESC). *European Heart Journal*, 35(39):2733–79, 2014.

[18] W Gregory Hundley et al. ACCF/ACR/AHA/NASCI/SCMR 2010 expert consensus document on cardiovascular magnetic resonance: a report of the American College of Cardiology Foundation Task Force on Expert Consensus Documents. *J Am Coll Cardiol*, 55(23):2614–2662, 2010.

[19] Martin G St John Sutton and Norman Sharpe. Left ventricular remodeling after myocardial infarction: pathophysiology and therapy. *Circulation*, 101(25):2981–2988, 2000.

[20] Robert C Hendel, Manesh R Patel, Christopher M Kramer, Michael Poon, James C Carr, Nancy A Gerstad, Linda D Gillam, John McB Hodgson, Raymond J Kim, John R Lesser, et al. Accf/acr/scct/scmr/asnc/nasci/scai/sir 2006 appropriateness criteria for cardiac computed tomography and cardiac magnetic resonance imaging: A report of the american college of cardiology foundation quality strategic directions committee appropriateness criteria working group, american college of radiology, society of cardiovascular computed tomography, society for cardiovascular magnetic resonance, american society of nuclear cardiology, north american society for cardiac imaging, society for cardiovascular angiography and interventions, and society of interventional radiology. *Journal of the American College of Cardiology*, 48(7):1475–1497, 2006.

[21] John P Ridgway. Cardiovascular magnetic resonance physics for clinicians: part i. *Journal of cardiovascular magnetic resonance*, 12(1):71, 2010.

[22] Sandra Pujadas, Gautham P Reddy, Oliver Weber, Jennifer J Lee, and Charles B Higgins. Mr imaging assessment of cardiac function. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 19(6):789–799, 2004.

[23] Donald W McRobbie, Elizabeth A Moore, Martin J Graves, and Martin R Prince. *MRI from Picture to Proton*. Cambridge university press, 2017.

[24] Daniel T Ginat, Michael W Fong, David J Tuttle, Susan K Hobbs, and Rajashree C
     Vyas. Cardiac imaging: part 1, mr pulse sequences, imaging planes, and basic
     anatomy. *American Journal of Roentgenology*, 197(4):808–815, 2011.

[25] Klaus Scheffler and Stefan Lehnhardt. Principles and applications of balanced ssfp
     techniques. *European radiology*, 13(11):2409–2418, 2003.

[26] Johan S Van den Brink, Yuji Watanabe, Christiane K Kuhl, Taylor Chung, Raja
     Muthupillai, Marc Van Cauteren, Kei Yamada, Steven Dymarkowski, Jan Bogaert,
     Jeff H Maki, et al. Implications of sense mr in routine clinical practice. *European
     journal of radiology*, 46(1):3–27, 2003.

[27] GS Gulsin, A Singh, and GP McCann. Cardiovascular magnetic resonance in the
     evaluation of heart valve disease. *BMC medical imaging*, 17(1):67, 2017.

[28] PM Pattynama, HJ Lamb, EA Van der Velde, EE Van der Wall, and A De Roos.
     Left ventricular measurements with cine and spin-echo mr imaging: a study of
     reproducibility with variance component analysis. *Radiology*, 187(1):261–268, 1993.

[29] Spencer B Gay, Chris L Sistrom, Chad A Holder, and Paul M Suratt. Breath-holding
     capability of adults. implications for spiral computed tomography, fast-acquisition
     magnetic resonance imaging, and angiography. *Investigative radiology*, 29(9):848–
     851, 1994.

[30] Reza S Razavi, Derek LG Hill, Vivek Muthurangu, Marc E Miquel, Andrew M Tay-
     lor, Sebastian Kozerke, and Edward J Baker. Three-dimensional magnetic resonance
     imaging of congenital cardiac anomalies. *Cardiology in the Young*, 13(5):461–465,
     2003.

[31] Wenjia Bai, Matthew Sinclair, Giacomo Tarroni, Ozan Oktay, Martin Rajchl, Ghis-
     lain Vaillant, Aaron M Lee, Nay Aung, Elena Lukaschuk, Mihir M Sanghvi, et al.
     Automated cardiovascular magnetic resonance image analysis with fully convolu-
     tional networks. *Journal of Cardiovascular Magnetic Resonance*, 20(1):65, 2018.

[32] El-Sayed H Ibrahim. Myocardial tagging by cardiovascular magnetic resonance: evolution of techniques–pulse sequences, analysis algorithms, and applications. *Journal of Cardiovascular Magnetic Resonance*, 13(1):36, 2011.

[33] John J Green, Jeffery S Berger, Christopher M Kramer, and Michael Salerno. Prognostic value of late gadolinium enhancement in clinical outcomes for hypertrophic cardiomyopathy. *JACC: Cardiovascular Imaging*, 5(4):370–377, 2012.

[34] Jay N Cohn, Roberto Ferrari, Norman Sharpe, et al. Cardiac remodeling—concepts and clinical implications: a consensus paper from an international forum on cardiac remodeling. *Journal of the American College of Cardiology*, 35(3):569–582, 2000.

[35] Rahman Attar, Marco Pereañez, Ali Gooya, Xènia Albà, Le Zhang, Milton Hoz de Vila, Aaron M Lee, Nay Aung, Elena Lukaschuk, Mihir M Sanghvi, et al. Quantitative CMR population imaging on 20,000 subjects of the uk biobank imaging study: LV/RV quantification pipeline and its evaluation. *Medical image analysis*, 56:26–42, 2019.

[36] Avan Suinesiaputra, Andrew D McCulloch, Martyn P Nash, Beau Pontre, and Alistair A Young. Cardiac image modelling: breadth and depth in heart disease, 2016.

[37] Thomas R Dawber, Gilcin F Meadors, and Felix E Moore Jr. Epidemiological approaches to heart disease: the framingham study. *American Journal of Public Health and the Nations Health*, 41(3):279–286, 1951.

[38] Annika Rosengren, Steven Hawken, Stephanie Ôunpuu, Karen Sliwa, Mohammad Zubaid, Wael A Almahmeed, Kathleen Ngu Blackett, Chitr Sitthi-amorn, Hiroshi Sato, Salim Yusuf, et al. Association of psychosocial risk factors with risk of acute myocardial infarction in 11 119 cases and 13 648 controls from 52 countries (the interheart study): case-control study. *The Lancet*, 364(9438):953–962, 2004.

[39] Christopher JL Murray, Theo Vos, Rafael Lozano, Mohsen Naghavi, Abraham D Flaxman, Catherine Michaud, Majid Ezzati, Kenji Shibuya, Joshua A Salomon, Safa Abdalla, et al. Disability-adjusted life years (dalys) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010. *The lancet*, 380(9859):2197–2223, 2012.

[40] Steffen E Petersen, Paul M Matthews, Fabian Bamberg, David A Bluemke, Jane M Francis, Matthias G Friedrich, Paul Leeson, Eike Nagel, Sven Plein, Frank E Rademakers, et al. Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of uk biobank-rationale, challenges and approaches. *Journal of Cardiovascular Magnetic Resonance*, 15(1):46, 2013.

[41] Pau Medrano-Gracia, Brett R Cowan, Avan Suinesiaputra, and Alistair A Young. Challenges of cardiac image analysis in large-scale population-based studies. *Current cardiology reports*, 17(3):9, 2015.

[42] Gerard M London, Bruno Pannier, Alain P Guerin, Jacques Blacher, Sylvain J Marchais, Bernadette Darne, Fabien Metivier, Hassan Adda, and Michel E Safar. Alterations of left ventricular hypertrophy in and survival of patients receiving hemodialysis: follow-up of an interventional study. *Journal of the American Society of Nephrology*, 12(12):2759–2767, 2001.

[43] Philipp S Wild, Christoph R Sinning, Alexander Roth, Sandra Wilde, Renate B Schnabel, Edith Lubos, Tanja Zeller, Till Keller, Karl J Lackner, Maria Blettner, et al. Distribution and categorization of left ventricular measurements in the general population: results from the population-based gutenberg heart study. *Circulation: Cardiovascular Imaging*, 3(5):604–613, 2010.

[44] J. N. Bella et al. Heritability of left ventricular dimensions and mass in american indians: the strong heart study. *Journal of hypertension*, 22(2):281–286, 2004.

[45] P. Sharma et al. Heritability of left ventricular mass in a large cohort of twins. *Journal of hypertension*, 24(2):321–324, 2006.

[46] Chad Garner, Edith Lecomte, Sophie Visvikis, Eric Abergel, Mark Lathrop, and Florent Soubrier. Genetic and environmental influences on left ventricular mass: a family study. *Hypertension*, 36(5):740–746, 2000.

[47] Christopher A Miller, Peter Jordan, Alex Borg, Rachel Argyle, David Clark, Keith Pearce, and Matthias Schmitt. Quantification of left ventricular indices from ssfp cine imaging: Impact of real-world variability in analysis methodology and utility of geometric modeling. *Journal of Magnetic Resonance Imaging*, 37(5):1213–1222, 2013.

[48] Qiao Zheng, Hervé Delingette, Nicolas Duchateau, and Nicholas Ayache. 3-d consistent and robust segmentation of cardiac images by deep learning with spatial propagation. *IEEE transactions on medical imaging*, 37(9):2137–2148, 2018.

[49] Jinming Duan, Ghalib Bello, Jo Schlemper, Wenjia Bai, Timothy JW Dawes, Carlo Biffi, Antonio de Marvao, Georgia Doumoud, Declan P O'Regan, and Daniel Rueckert. Automatic 3D bi-ventricular segmentation of cardiac images by a shape-refined multi-task deep learning approach. *IEEE transactions on medical imaging*, 38(9):2151–2164, 2019.

[50] 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. *Medical image analysis*, 51:21–45, 2019.

[51] Piotr J Slomka, Damini Dey, Arkadiusz Sitek, Manish Motwani, Daniel S Berman, and Guido Germano. Cardiac imaging: working towards fully-automated machine analysis & interpretation. *Expert review of medical devices*, 14(3):197–212, 2017.

[52] Diane E Bild, David A Bluemke, Gregory L Burke, Robert Detrano, Ana V Diez Roux, Aaron R Folsom, Philip Greenland, David R JacobsJr, Richard Kronmal, Kiang Liu, et al. Multi-ethnic study of atherosclerosis: objectives and design. *American journal of epidemiology*, 156(9):871–881, 2002.

[53] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3), 2015.

[54] Jane Harman, Evelyn R Walker, Vicki Charbonneau, Ermeg L Akylbekova, Cheryl Nelson, and Sharon B Wyatt. Treatment of hypertension among a frican a mericans: The j ackson heart study. *The Journal of Clinical Hypertension*, 15(6):367–374, 2013.

[55] Carissa G Fonseca et al. The Cardiac Atlas Project: an imaging database for computational modeling and statistical atlases of the heart. *Bioinformatics*, 27(16):2288–2295, 2011.

[56] Antonio de Marvao, Timothy JW Dawes, Wenzhe Shi, Christopher Minas, Niall G Keenan, Tamara Diamond, Giuliana Durighel, Giovanni Montana, Daniel Rueckert, Stuart A Cook, et al. Population-based studies of myocardial hypertrophy: high resolution cardiovascular magnetic resonance atlases improve statistical power. *Journal of cardiovascular magnetic resonance*, 16(1):16, 2014.

[57] Michel G Khouri, Ronald M Peshock, Colby R Ayers, James A de Lemos, and Mark H Drazner. A 4-tiered classification of left ventricular hypertrophy based on left ventricular geometry: the dallas heart study. *Circulation: Cardiovascular Imaging*, 3(2):164–171, 2010.

[58] William H Gaasch and Michael R Zile. Left ventricular structural remodeling in health and disease: with special emphasis on volume, mass, and geometry. *Journal of the American College of Cardiology*, 58(17):1733–1740, 2011.

[59] Alejandro Lopez-Perez, Rafael Sebastian, and Jose M Ferrero. Three-dimensional cardiac computational modelling: methods, features and applications. *Biomedical Engineering Online*, 14(1):35, 2015.

[60] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic MRI cardiac multistructures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

[61] Ralph Knöll, Guido Iaccarino, Guido Tarone, Denise Hilfiker-Kleiner, Johann Bauersachs, Adelino F Leite-Moreira, Peter H Sugden, and Jean-Luc Balligand. Towards a re-definition of 'cardiac hypertrophy'through a rational characterization of left ventricular phenotypes: a position paper of the working group 'myocardial function'of the esc. *European journal of heart failure*, 13(8):811–819, 2011.

[62] Guang Li, Adele Xu, Sopheak Sim, James R Priest, Xueying Tian, Tooba Khan, Thomas Quertermous, Bin Zhou, Philip S Tsao, Stephen R Quake, et al. Transcriptomic profiling maps anatomically patterned subpopulations among single embryonic cardiac cells. *Developmental cell*, 39(4):491–507, 2016.

[63] Antonio de Marvao, Timothy JW Dawes, and Declan P O'Regan. Artificial intelligence for cardiac imaging-genetics research. *Frontiers in Cardiovascular Medicine*, 6:195, 2020.

[64] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell,

et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[65] Caroline S Fox, Jennifer L Hall, Donna K Arnett, Euan A Ashley, Christian Delles, Mary B Engler, Mason W Freeman, Julie A Johnson, David E Lanfear, Stephen B Liggett, et al. Future translational applications from the contemporary genomics era: a scientific statement from the american heart association. *Circulation*, 131(19):1715–1736, 2015.

[66] T. D. Adams et al. Heritability of cardiac size: an echocardiographic and electro-cardiographic study of monozygotic and dizygotic twins. *Circulation*, 71(1):39–44, 1985.

[67] D. Govindaraju et al. Genetics of the framingham heart study population. *Advances in genetics*, 62:33–65, 2008.

[68] David Houle, Diddahally R Govindaraju, and Stig Omholt. Phenomics: the next challenge. *Nature reviews genetics*, 11(12):855–866, 2010.

[69] CA MacRae and RS Vasan. Next-generation genome-wide association studies: time to focus on phenotype? *Circulation. Cardiovascular genetics*, 4(4):334–336, 2011.

[70] Robert M Bilder, FW Sabb, TD Cannon, ED London, JD Jentsch, D Stott Parker, Russell A Poldrack, C Evans, and NB Freimer. Phenomics: the systematic study of phenotypes on a genome-wide scale. *Neuroscience*, 164(1):30–42, 2009.

[71] Ramachandran S Vasan, Nicole L Glazer, Janine F Felix, Wolfgang Lieb, Philipp S Wild, Stephan B Felix, Norbert Watzinger, Martin G Larson, Nicholas L Smith, Abbas Dehghan, et al. Genetic variants associated with cardiac structure and function: a meta-analysis and replication of genome-wide association data. *Jama*, 302(2):168–178, 2009.

[72] Robert Plomin, Claire MA Haworth, and Oliver SP Davis. Common disorders are quantitative traits. *Nature reviews genetics*, 10(12):872–878, 2009.

[73] Evan E Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M Leal, Jason H Moore, and Joseph H Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6):446–450, 2010.

[74] Geert Litjens et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[75] Mir Henglin, Gillian Stein, Pavel V Hushcha, Jasper Snoek, Alexander B Wiltschko, and Susan Cheng. Machine learning approaches in cardiovascular imaging. *Circulation: Cardiovascular Imaging*, 10(10):e005614, 2017.

[76] Carlos Martin-Isla, Victor M Campello, Cristian Izquierdo, Zahra Raisi-Estabragh, Bettina Baeßler, Steffen E Petersen, and Karim Lekadir. Image-based cardiac diagnosis with machine learning: A review. *Frontiers in Cardiovascular Medicine*, 7:1, 2020.

[77] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[78] Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. Deep learning for cardiac image segmentation: A review. *Frontiers in Cardiovascular Medicine*, 7:25, 2020.

[79] Geert Litjens, Francesco Ciompi, Jelmer M Wolterink, Bob D de Vos, Tim Leiner, Jonas Teuwen, and Ivana Išgum. State-of-the-art deep learning in cardiovascular image analysis. *JACC: Cardiovascular Imaging*, 12(8):1549–1565, 2019.

[80] Berkman Sahiner, Aria Pezeshk, Lubomir M Hadjiiski, Xiaosong Wang, Karen Drukker, Kenny H Cha, Ronald M Summers, and Maryellen L Giger. Deep learning in medical imaging and radiation therapy. *Medical physics*, 46(1):e1–e36, 2019.

[81] Kathleen Gilbert, Wenjia Bai, Charlene Mauger, Pau Medrano-Gracia, Avan Suinesiaputra, Aaron M Lee, Mihir M Sanghvi, Nay Aung, Stefan K Piechnik, Stefan Neubauer, et al. Independent left ventricular morphometric atlases show consistent relationships with cardiovascular risk factors: A UK Biobank study. *Scientific Reports*, 9(1):1130, 2019.

[82] Tommaso Mansi, Ingmar Voigt, Benedetta Leonardi, Xavier Pennec, Stanley Durrleman, Maxime Sermesant, Hervé Delingette, Andrew Mayall Taylor, Younes Boudjemline, Giacomo Pongiglione, et al. A statistical model for quantification and prediction of cardiac remodelling: Application to tetralogy of fallot. *IEEE transactions on medical imaging*, 30(9):1605–1616, 2011.

[83] Antonio de Marvao, Timothy JW Dawes, Wenzhe Shi, Giuliana Durighel, Daniel Rueckert, Stuart A Cook, and Declan P O'Regan. Precursors of hypertensive heart phenotype develop in healthy adults: a high-resolution 3D MRI study. *JACC: Cardiovascular Imaging*, 8(11):1260–1269, 2015.

[84] Pau Medrano-Gracia, Brett R Cowan, Bharath Ambale-Venkatesh, David A Bluemke, John Eng, John Paul Finn, Carissa G Fonseca, Joao AC Lima, Avan Suinesiaputra, and Alistair A Young. Left ventricular shape variation in asymptomatic populations: the multi-ethnic study of atherosclerosis. *Journal of Cardiovascular Magnetic Resonance*, 16(1):56, 2014.

[85] Alistair A Young and Alejandro F Frangi. Computational cardiac atlases: from patient to population and back. *Experimental Physiology*, 94(5):578–596, 2009.

[86] M. Varela et al. Novel computational analysis of left atrial anatomy improves prediction of atrial fibrillation recurrence after ablation. *Frontiers in physiology*, 8, 2017.

[87] P. Yu et al. Cortical surface shape analysis based on spherical wavelets. *IEEE transactions on medical imaging*, 26(4):582–597, 2007.

[88] William D Penny, Karl J Friston, John T Ashburner, Stefan J Kiebel, and Thomas E Nichols. *Statistical parametric mapping: the analysis of functional brain images.* Elsevier, 2011.

[89] Mathieu De Craene, Nicolas Duchateau, Catalina Tobon-Gomez, Babak Ghafaryasl, Gemma Piella, Kawal S Rhode, and AF Frange. Spm to the heart: mapping of 4d continuous velocities for motion abnormality quantification. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 454–457. IEEE, 2012.

[90] I. L. Dryden and K. V. Mardia. *Statistical Shape Analysis.* Wiley, Chichester, 1998.

[91] Ulf Grenander and Michael I Miller. Computational anatomy: An emerging discipline. *Quarterly of applied mathematics*, 56(4):617–694, 1998.

[92] Maxime Sermesant, Clément Forest, Xavier Pennec, Hervé Delingette, and Nicholas Ayache. Deformable biomechanical models: Application to 4d cardiac image analysis. *Medical image analysis*, 7(4):475–488, 2003.

[93] Nic Smith, Adelaide de Vecchi, Matthew McCormick, David Nordsletten, Oscar Camara, Alejandro F Frangi, Hervé Delingette, Maxime Sermesant, Jatin Relan, Nicholas Ayache, et al. euheart: personalized and integrated cardiac care using patient-specific cardiovascular modelling. *Interface focus*, 1(3):349–364, 2011.

[94] Daniel Rueckert, Alejandro F Frangi, and Julia A Schnabel. Automatic construction of 3D statistical deformation models using non-rigid registration. In *Interna-*

tional Conference on Medical Image Computing and Computer-Assisted Intervention, pages 77–84. Springer, 2001.

[95] Kanwal K Bhatia, Joseph V Hajnal, Basant K Puri, A David Edwards, and Daniel Rueckert. Consistent groupwise non-rigid registration for atlas construction. In *2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821)*, pages 908–911. IEEE, 2004.

[96] Wenjia Bai, Wenzhe Shi, Antonio de Marvao, Timothy JW Dawes, Declan P O'Regan, Stuart A Cook, and Daniel Rueckert. A bi-ventricular cardiac atlas built from 1000+ high resolution mr images of healthy subjects and an analysis of shape and motion. *Medical image analysis*, 26(1):133–145, 2015.

[97] Alexandre Guimond, Jean Meunier, and Jean-Philippe Thirion. Automatic computation of average brain models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 631–640. Springer, 1998.

[98] Alexandre Guimond, Jean Meunier, and Jean-Philippe Thirion. Average brain models: A convergence study. 1999.

[99] Herve Lombaert, Jean-Marc Peyrat, Pierre Croisille, Stanislas Rapacchi, Laurent Fanton, Farida Cheriet, Patrick Clarysse, Isabelle Magnin, Hervé Delingette, and Nicholas Ayache. Human atlas of the cardiac fiber architecture: study on a healthy population. *IEEE transactions on medical imaging*, 31(7):1436–1447, 2012.

[100] Stephen Marsland, Carole J Twining, and Chris J Taylor. Groupwise non-rigid registration using polyharmonic clamped-plate splines. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 771–779. Springer, 2003.

[101] Sarang Joshi, Brad Davis, Matthieu Jomier, and Guido Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage*, 23:S151–S160, 2004.

[102] Maria Lorenzo-Valdés, Gerardo I Sanchez-Ortiz, Andrew G Elkington, Raad H Mohiaddin, and Daniel Rueckert. Segmentation of 4d cardiac MR images using a probabilistic atlas and the EM algorithm. *Medical Image Analysis*, 8(3):255–265, 2004.

[103] David W Shattuck, Mubeena Mirza, Vitria Adisetiyo, Cornelius Hojatkashani, Georges Salamon, Katherine L Narr, Russell A Poldrack, Robert M Bilder, and Arthur W Toga. Construction of a 3D probabilistic atlas of human cortical structures. *Neuroimage*, 39(3):1064–1080, 2008.

[104] Jörn Diedrichsen, Joshua H Balsters, Jonathan Flavell, Emma Cussans, and Narender Ramnani. A probabilistic MR atlas of the human cerebellum. *Neuroimage*, 46(1):39–46, 2009.

[105] Adam J Lewandowski, Daniel Augustine, Pablo Lamata, Esther F Davis, Merzaka Lazdam, Jane Francis, Kenny McCormick, Andrew R Wilkinson, Atul Singhal, Alan Lucas, et al. Preterm heart in adult life: cardiovascular magnetic resonance reveals distinct differences in left ventricular mass, geometry, and function. *Circulation*, 127(2):197–206, 2013.

[106] Xingyu Zhang, Brett R Cowan, David A Bluemke, J Paul Finn, Carissa G Fonseca, Alan H Kadish, Daniel C Lee, Joao AC Lima, Avan Suinesiaputra, Alistair A Young, et al. Atlas-based quantification of cardiac remodeling due to myocardial infarction. *PloS One*, 9(10):e110243, 2014.

[107] David R Warriner, Tom Jackson, Ernesto Zacur, Eva Sammut, Paul Sheridan, David Rod Hose, Patricia Lawford, Reza Razavi, Steve Alexander Niederer, Christopher Aldo Rinaldi, et al. An asymmetric wall-thickening pattern predicts response to cardiac resynchronization therapy. *JACC: Cardiovascular Imaging*, 11(10):1545–1546, 2018.

[108] Kathleen Gilbert, Nickolas Forsch, Sanjeet Hegde, Charlene Mauger, Jeffrey H
      Omens, James C Perry, Beau Pontré, Avan Suinesiaputra, Alistair A Young, and
      Andrew D McCulloch. Atlas-based computational analysis of heart shape and func-
      tion in congenital heart disease. *Journal of cardiovascular translational research*,
      11(2):123–132, 2018.

[109] Mark I Attard, Timothy JW Dawes, Antonio de Marvao, Carlo Biffi, Wenzhe Shi,
      John Wharton, Christopher J Rhodes, Pavandeep Ghataorhe, J Simon R Gibbs,
      Luke SGE Howard, et al. Metabolic pathways associated with right ventricular
      adaptation to pulmonary hypertension: 3D analysis of cardiac magnetic resonance
      imaging. *European Heart Journal-Cardiovascular Imaging*, 20(6):668–676, 2019.

[110] Guoyan Zheng, Shuo Li, and Gabor Szekely. *Statistical shape and deformation
      analysis: methods, implementation and applications*. Academic Press, 2017.

[111] Alejandro F Frangi, Daniel Rueckert, Julia A Schnabel, and Wiro J Niessen. Au-
      tomatic construction of multiple-object three-dimensional statistical shape mod-
      els: Application to cardiac modeling. *IEEE transactions on medical imaging*,
      21(9):1151–1166, 2002.

[112] Tobias Heimann and Hans-Peter Meinzer. Statistical shape models for 3D medical
      image segmentation: a review. *Medical image analysis*, 13(4):543–563, 2009.

[113] Timothy F Cootes and Christopher J Taylor. Active shape models—'smart snakes'.
      In *BMVC92*, pages 266–275. Springer, 1992.

[114] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Ac-
      tive shape models-their training and application. *Computer vision and image un-
      derstanding*, 61(1):38–59, 1995.

[115] Ian T Jolliffe. Principal components in regression analysis. In *Principal component
      analysis*, pages 129–155. Springer, 1986.

[116] Mikkel B Stegmann and David Delgado Gomez. A brief introduction to statistical shape analysis. *Informatics and mathematical modelling, Technical University of Denmark, DTU*, 15(11), 2002.

[117] Giovanni Biglino, Claudio Capelli, Jan Bruse, Giorgia M Bosi, Andrew M Taylor, and Silvia Schievano. Computational modelling for congenital heart disease: how far are we from clinical translation? *Heart*, 103(2):98–103, 2017.

[118] Pablo Lamata, Ramón Casero, Valentina Carapella, Steve A Niederer, Martin J Bishop, Jürgen E Schneider, Peter Kohl, and Vicente Grau. Images as drivers of progress in cardiac computational modelling. *Progress in biophysics and molecular biology*, 115(2-3):198–212, 2014.

[119] Avan Suinesiaputra, Pierre Ablin, Xenia Alba, Martino Alessandrini, Jack Allen, Wenjia Bai, Serkan Cimen, Peter Claes, Brett R Cowan, Jan D'hooge, et al. Statistical shape modeling of the left ventricle: myocardial infarct classification challenge. *IEEE Journal of Biomedical and Health Informatics*, 22(2):503–515, 2018.

[120] Timothy F Cootes, Andrew Hill, Christopher J Taylor, and Jane Haslam. Use of active shape models for locating structures in medical images. *Image and vision computing*, 12(6):355–365, 1994.

[121] Espen W Remme, Alistair A Young, Kevin F Augenstein, Brett Cowan, and Peter J Hunter. Extraction and quantification of left ventricular deformation modes. *IEEE Transactions on Biomedical Engineering*, 51(11):1923–1931, 2004.

[122] Pablo Lamata, Merzaka Lazdam, Anna Ashcroft, Adam J Lewandowski, Paul Leeson, and Nic Smith. Computational mesh as a descriptor of left ventricular shape for clinical diagnosis. In *Computing in Cardiology 2013*, pages 571–574. IEEE, 2013.

[123] Genevieve Farrar, Avan Suinesiaputra, Kathleen Gilbert, James C Perry, Sanjeet Hegde, Alison Marsden, Alistair A Young, Jeffrey H Omens, and Andrew D Mc-

Culloch. Atlas-based ventricular shape analysis for understanding congenital heart disease. *Progress in pediatric cardiology*, 43:61–69, 2016.

[124] Siamak Ardekani, Robert G Weiss, Albert C Lardo, Richard T George, Joao AC Lima, Katherine C Wu, Michael I Miller, Raimond L Winslow, and Laurent Younes. Computational method for identifying and quantifying shape features of human left ventricular remodeling. *Annals of biomedical engineering*, 37(6):1043–1054, 2009.

[125] Dimitrios Perperidis, Raad Mohiaddin, and Daniel Rueckert. Construction of a 4d statistical atlas of the cardiac anatomy and its use in classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 402–410. Springer, 2005.

[126] Karim Lekadir, Xènia Albà, Marco Pereañez, and Alejandro F Frangi. Statistical shape modeling using partial least squares: application to the assessment of myocardial infarction. In *Statistical Atlases and Computational Models of the Heart*, pages 130–139. Springer, 2015.

[127] Xingyu Zhang, Pau Medrano-Gracia, Bharath Ambale-Venkatesh, David A Bluemke, Brett R Cowan, J Paul Finn, Alan H Kadish, Daniel C Lee, Joao AC Lima, Alistair A Young, et al. Orthogonal decomposition of left ventricular remodeling in myocardial infarction. *GigaScience*, 6(3):gix005, 2017.

[128] T. Dawes et al. Machine learning of three-dimensional right ventricular motion enables outcome prediction in pulmonary hypertension: a cardiac mr imaging study. *Radiology*, 283(2):381–390, 2017.

[129] Stanley Durrleman, Xavier Pennec, Alain Trouvé, and Nicholas Ayache. Statistical models of sets of curves and surfaces based on currents. *Medical image analysis*, 13(5):793–808, 2009.

[130] J. Ashburner et al. Identifying global anatomical differences: deformation-based morphometry. *Human brain mapping*, 6(5-6):348–357, 1998.

[131] J. Ashburner and K. J. Friston. Diffeomorphic registration using geodesic shooting and gauss–newton optimisation. *NeuroImage*, 55(3):954–967, 2011.

[132] M. Vaillant et al. Statistics on diffeomorphisms via tangent space representations. *NeuroImage*, 23:S161–S169, 2004.

[133] X. Pennec et al. Prediction of post-ablation outcome in atrial fibrillation using shape parameterization and partial least squares regression. *Functional Imaging and Modelling of the Heart*, page 311, 2017.

[134] M. Zhang et al. Low-dimensional statistics of anatomical variability via compact representation of image deformations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 166–173. Springer, 2016.

[135] Daniel Rueckert, Alejandro F Frangi, and Julia A Schnabel. Automatic construction of 3-d statistical deformation models of the brain using nonrigid registration. *IEEE transactions on medical imaging*, 22(8):1014–1025, 2003.

[136] Miaomiao Zhang and Polina Golland. Statistical shape analysis: From landmarks to diffeomorphisms, 2016.

[137] J. Modersitzki. Memory efficient lddmm for lung ct. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016*, page 28.

[138] Ndeye Coumba Ndiaye, Mohsen Azimi Nehzad, Said El Shamieh, Maria G Stathopoulou, and Sophie Visvikis-Siest. Cardiovascular diseases and genome-wide association studies. *Clinica Chimica Acta*, 412(19-20):1697–1701, 2011.

[139] Matthew B Lanktree, Reina G Hassell, Piya Lahiry, and Robert A Hegele. Phenomics: expanding the role of clinical evaluation in genomic studies. *Journal of Investigative Medicine*, 58(5):700–706, 2010.

[140] Tian Ge, Gunter Schumann, and Jianfeng Feng. Imaging genetics—towards discovery neuroscience. *Quantitative Biology*, 1(4):227–245, 2013.

[141] Eva Loth, Jean-Baptiste Poline, Benjamin Thyreau, Tianye Jia, Chenyang Tao, Anbarasu Lourdusamy, David Stacey, Anna Cattrell, Sylvane Desrivières, Barbara Ruggeri, et al. Oxytocin receptor genotype modulates ventral striatal activity to social cues and response to stressful life events. *Biological psychiatry*, 76(5):367–376, 2014.

[142] Jason L Stein, Derrek P Hibar, Sarah K Madsen, Mathew Khamis, Katie L McMahon, Greig I de Zubicaray, Narelle K Hansell, Grant W Montgomery, Nicholas G Martin, Margaret J Wright, et al. Discovery and replication of dopamine-related gene effects on caudate volume in young and elderly populations (n= 1198) using genome-wide search. *Molecular psychiatry*, 16(9):927–937, 2011.

[143] Nicola Filippini, Anil Rao, Sally Wetten, Rachel A Gibson, Michael Borrie, Danilo Guzman, Andrew Kertesz, Inge Loy-English, Julie Williams, Thomas Nichols, et al. Anatomically-distinct genetic associations of APOE $\epsilon 4$ allele load with regional cortical atrophy in Alzheimer's disease. *Neuroimage*, 44(3):724–728, 2009.

[144] Jason L Stein, Xue Hua, Suh Lee, April J Ho, Alex D Leow, Arthur W Toga, Andrew J Saykin, Li Shen, Tatiana Foroud, Nathan Pankratz, et al. Voxelwise genome-wide association study (vgwas). *neuroimage*, 53(3):1160–1174, 2010.

[145] Ahmad R Hariri and Daniel R Weinberger. Imaging genomics. *British medical bulletin*, 65(1):259–270, 2003.

[146] Jonathan Flint and Marcus R Munafò. The endophenotype concept in psychiatric genetics. *Psychological medicine*, 37(2):163–180, 2007.

[147] Andreas Meyer-Lindenberg and Daniel R Weinberger. Intermediate phenotypes and genetic mechanisms of psychiatric disorders. *Nature reviews neuroscience*, 7(10):818–827, 2006.

[148] Benjamin SC Wade, Shantanu H Joshi, Boris A Gutman, and Paul M Thompson. Machine learning on high dimensional shape data from subcortical brain surfaces:

A comparison of feature selection and classification methods. *Pattern Recognition*, 63:731–739, 2017.

[149] Hua Wang, Feiping Nie, Heng Huang, Shannon L Risacher, Andrew J Saykin, Li Shen, and Alzheimer's Disease Neuroimaging Initiative. Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics*, 28(12):i127–i136, 2012.

[150] Maria Vounou, Eva Janousova, Robin Wolz, Jason L Stein, Paul M Thompson, Daniel Rueckert, Giovanni Montana, Alzheimer's Disease Neuroimaging Initiative, et al. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. *Neuroimage*, 60(1):700–716, 2012.

[151] Meiyan Huang, Thomas Nichols, Chao Huang, Yang Yu, Zhaohua Lu, Rebecca C Knickmeyer, Qianjin Feng, Hongtu Zhu, Alzheimer's Disease Neuroimaging Initiative, et al. Fvgwas: Fast voxelwise genome wide association analysis of large-scale imaging genetic data. *Neuroimage*, 118:613–627, 2015.

[152] R. Klein et al. Complement factor h polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389, 2005.

[153] A. DeWan et al. Htra1 promoter polymorphism in wet age-related macular degeneration. *Science*, 314(5801):989–992, 2006.

[154] R. Duerr et al. A genome-wide association study identifies il23r as an inflammatory bowel disease gene. *science*, 314(5804):1461–1463, 2006.

[155] A. Herbert et al. A common genetic variant is associated with adult and childhood obesity. *Science*, 312(5771):279–283, 2006.

[156] A. Parsa et al. Hypertrophy-associated polymorphisms ascertained in a founder cohort applied to heart failure risk and mortality. *Clinical and translational science*, 4(1):17–23, 2011.

[157] Donna K Arnett, Na Li, Weihong Tang, Dabeeru C Rao, Richard B Devereux, Steven A Claas, Rachel Kraemer, and Ulrich Broeckel. Genome-wide association study identifies single-nucleotide polymorphism in kcnb1 associated with left ventricular mass in humans: the hypergen study. *BMC medical genetics*, 10(1):43, 2009.

[158] D. Arnett et al. Genetic variation in ncam1 contributes to left ventricular wall thickness in hypertensive families. *Circulation research*, pages CIRCRESAHA–110, 2011.

[159] Ervin R Fox, Solomon K Musani, Maja Barbalic, Honghuang Lin, Bing Yu, Kofo O Ogunyankin, Nicholas L Smith, Abdullah Kutlar, Nicole L Glazer, Wendy S Post, et al. Genome-wide association study of cardiac structure and systolic function in african americans: the candidate gene association resource (care) study. *Circulation: Cardiovascular Genetics*, 6(1):37–46, 2013.

[160] Pankaj Arora and Christopher Newton-Cheh. Blood pressure and human genetic variation in the general population. *Current opinion in cardiology*, 25(3):229, 2010.

[161] Peter A Bandettini. Twenty years of functional MRI: the science and the stories. *Neuroimage*, 62(2):575–588, 2012.

[162] John Ashburner and Karl J Friston. Voxel-based morphometry—the methods. *Neuroimage*, 11(6):805–821, 2000.

[163] IC Wright, PK McGuire, J-B Poline, JM Travere, RM Murray, CD Frith, RSJ Frackowiak, and KJ Friston. A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia. *Neuroimage*, 2(4):244–252, 1995.

[164] Mark W Woolrich, Christian F Beckmann, Thomas E Nichols, and Stephen M Smith. Statistical analysis of fMRI data. In *fMRI techniques and protocols*, pages 179–236. Springer, 2009.

[165] Justin Chumbley, Keith Worsley, Guillaume Flandin, and Karl Friston. Topological fdr for neuroimaging. *Neuroimage*, 49(4):3057–3064, 2010.

[166] Stephen M Smith and Thomas E Nichols. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*, 44(1):83–98, 2009.

[167] Carlo E Bonferroni, C Bonferroni, and CE Bonferroni. Teoria statistica delle classi e calcolo delle probabilita'. 1936.

[168] Christopher R Genovese, Nicole A Lazar, and Thomas Nichols. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15(4):870–878, 2002.

[169] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[170] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[171] Yoav Benjamini, Abba M Krieger, and Daniel Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.

[172] Thomas E Nichols. Multiple testing corrections, nonparametric methods, and random field theory. *Neuroimage*, 62(2):811–815, 2012.

[173] Choong-Wan Woo, Anjali Krishnan, and Tor D Wager. Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *Neuroimage*, 91:412–419, 2014.

[174] Eric Maris and Robert Oostenveld. Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, 164(1):177–190, 2007.

[175] Ben Corden, Antonio de Marvao, Timothy J Dawes, Wenzhe Shi, Daniel Rueckert, Stuart A Cook, and Declan P O'Regan. Relationship between body composition and left ventricular geometry using three dimensional cardiovascular magnetic resonance. *Journal of Cardiovascular Magnetic Resonance*, 18(1):32, 2016.

[176] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[177] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[178] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[179] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[180] Li Deng and Yang Liu. *Deep learning in natural language processing*. Springer, 2018.

[181] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

[182] Frank Rosenblatt. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Lab Inc Buffalo NY, 1961.

[183] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[184] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[185] Charlie Nash and Christopher KI Williams. The shape variational autoencoder: A deep generative model of part-segmented 3D objects. In *Computer Graphics Forum*, volume 36, pages 1–12. Wiley Online Library, 2017.

[186] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[187] Jan L Bruse, Kristin McLeod, Giovanni Biglino, Hopewell N Ntsinjana, Claudio Capelli, Tain-Yen Hsia, Maxime Sermesant, Xavier Pennec, Andrew M Taylor, Silvia Schievano, et al. A statistical shape modelling framework to extract 3D shape biomarkers from medical imaging data: assessing arch morphology of repaired coarctation of the aorta. *BMC medical imaging*, 16(1):40, 2016.

[188] Benjamín Gutiérrez-Becker and Christian Wachinger. Deep multi-structural shape analysis: application to neuroanatomy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 523–531. Springer, 2018.

[189] Benjamín Gutiérrez-Becker and Christian Wachinger. Learning a conditional generative model for anatomical shape analysis. In *International Conference on Information Processing in Medical Imaging*, pages 505–516. Springer, 2019.

[190] Jan L Bruse, Maria A Zuluaga, Abbas Khushnood, Kristin McLeod, Hopewell N
Ntsinjana, Tain-Yen Hsia, Maxime Sermesant, Xavier Pennec, Andrew M Taylor,
and Silvia Schievano. Detecting clinically meaningful shape clusters in medical im-
age data: metrics analysis for hierarchical clustering applied to healthy and patho-
logical aortic arches. *IEEE Transactions on Biomedical Engineering*, 64(10):2373–
2383, 2017.

[191] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller.
Causability and explainability of artificial intelligence in medicine. *Wiley Interdis-
ciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312, 2019.

[192] Sanjeev Muralikrishnan, Vladimir G Kim, Matthew Fisher, and Siddhartha Chaud-
huri. Shape unicode: A unified shape representation. In *Proceedings of the IEEE
Conference on Computer Vision and Pattern Recognition*, pages 3790–3799, 2019.

[193] Eman Ahmed, Alexandre Saint, Abd El Rahman Shabayek, Kseniya Cherenkova,
Rig Das, Gleb Gusev, Djamila Aouada, and Bjorn Ottersten. A survey on
deep learning advances on different 3D data representations. *arXiv preprint
arXiv:1808.01462*, 2018.

[194] Reihaneh Rostami, Fereshteh S Bashiri, Behrouz Rostami, and Zeyun Yu. A survey
on data-driven 3D shape descriptors. In *Computer Graphics Forum*, volume 38,
pages 356–393. Wiley Online Library, 2019.

[195] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and
Philip S Yu. A comprehensive survey on graph neural networks. *arXiv preprint
arXiv:1901.00596*, 2019.

[196] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou
Tang, and Jianxiong Xiao. 3D shapenets: A deep representation for volumetric
shapes. In *Proceedings of the IEEE conference on computer vision and pattern
recognition*, pages 1912–1920, 2015.

[197] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016.

[198] Xiaoguang Han, Zhen Li, Haibin Huang, Evangelos Kalogerakis, and Yizhou Yu. High-resolution shape completion using deep neural networks for global structure and local geometry inference. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 85–93, 2017.

[199] Jun Li, Kai Xu, Siddhartha Chaudhuri, Ersin Yumer, Hao Zhang, and Leonidas Guibas. Grass: Generative recursive autoencoders for shape structures. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017.

[200] Cheng Wang, Ming Cheng, Ferdous Sohel, Mohammed Bennamoun, and Jonathan Li. Normalnet: A voxel-based CNN for 3D object classification and retrieval. *Neurocomputing*, 323:139–147, 2019.

[201] Zhuotun Zhu, Xinggang Wang, Song Bai, Cong Yao, and Xiang Bai. Deep learning representation using autoencoder for 3D shape retrieval. *Neurocomputing*, 204:41–50, 2016.

[202] Qingyang Tan, Lin Gao, Yu-Kun Lai, and Shihong Xia. Variational autoencoders for deforming 3D mesh models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5841–5850, 2018.

[203] Ilke Demir, Camilla Hahn, Kathryn Leonard, Geraldine Morin, Dana Rahbani, Athina Panotopoulou, Amelie Fondevilla, Elena Balashova, Bastien Durix, and Adam Kortylewski. Skelneton 2019: Dataset and challenge on deep learning for geometric shape understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[204] Yi Fang, Jin Xie, Guoxian Dai, Meng Wang, Fan Zhu, Tiantian Xu, and Edward Wong. 3D deep shape descriptor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2319–2328, 2015.

[205] Jin Xie, Guoxian Dai, Fan Zhu, Edward K Wong, and Yi Fang. Deepshape: Deep-learned shape descriptor for 3D shape retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1335–1345, 2016.

[206] Yueqing Wang, Zhige Xie, Kai Xu, Yong Dou, and Yuanwu Lei. An efficient and effective convolutional auto-encoder extreme learning machine network for 3D feature learning. *Neurocomputing*, 174:988–998, 2016.

[207] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3D surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–224, 2018.

[208] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[209] Guoxian Dai, Jin Xie, Fan Zhu, and Yi Fang. Learning a discriminative deformation-invariant 3D shape descriptor via many-to-one encoder. *Pattern Recognition Letters*, 83:330–338, 2016.

[210] Xi Cheng, Li Zhang, and Yefeng Zheng. Deep similarity learning for multimodal medical images. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 6(3):248–252, 2018.

[211] Min Chen, Xiaobo Shi, Yin Zhang, Di Wu, and Mohsen Guizani. Deep features learning for medical image analysis with convolutional autoencoder neural network. *IEEE Transactions on Big Data*, 2017.

[212] Lovedeep Gondara. Medical image denoising using convolutional denoising autoencoders. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 241–246. IEEE, 2016.

[213] Janki Mehta and Angshul Majumdar. Rodeo: robust de-aliasing autoencoder for real-time medical image reconstruction. *Pattern Recognition*, 63:499–510, 2017.

[214] MR Avendi, Arash Kheradvar, and Hamid Jafarkhani. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI. *Medical image analysis*, 30:108–119, 2016.

[215] Ozan Oktay, Enzo Ferrante, Konstantinos Kamnitsas, Mattias Heinrich, Wenjia Bai, Jose Caballero, Stuart A Cook, Antonio De Marvao, Timothy Dawes, Declan P O'Regan, et al. Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. *IEEE transactions on medical imaging*, 37(2):384–395, 2017.

[216] Hariharan Ravishankar, Rahul Venkataramani, Sheshadri Thiruvenkadam, Prasad Sudhakar, and Vivek Vaidya. Learning and incorporating shape models for semantic segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 203–211. Springer, 2017.

[217] Alexandre Bône, Maxime Louis, Olivier Colliot, Stanley Durrleman, Alzheimer's Disease Neuroimaging Initiative, et al. Learning low-dimensional representations of shape data sets with diffeomorphic autoencoders. In *International Conference on Information Processing in Medical Imaging*, pages 195–207. Springer, 2019.

[218] Zehra Camlica, Hamid R Tizhoosh, and Farzad Khalvati. Autoencoding the retrieval relevance of medical images. In *2015 International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 550–555. IEEE, 2015.

[219] Andrew Y Ng and Michael I Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pages 841–848, 2002.

[220] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[221] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *Advances in neural information processing systems*, pages 658–666, 2016.

[222] Martin Simonovsky and Nikos Komodakis. Graphvae: Towards generation of small graphs using variational autoencoders. In *International Conference on Artificial Neural Networks*, pages 412–422. Springer, 2018.

[223] Weidi Xu, Haoze Sun, Chao Deng, and Ying Tan. Variational autoencoder for semi-supervised text classification. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[224] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *European Conference on Computer Vision*, pages 835–851. Springer, 2016.

[225] Shikun Liu, Lee Giles, and Alexander Ororbia. Learning a hierarchical latent-variable model of 3D shapes. In *2018 International Conference on 3D Vision (3DV)*, pages 542–551. IEEE, 2018.

[226] Xianfeng Han, Hamid Laga, and Mohammed Bennamoun. Image-based 3D object reconstruction: State-of-the-art and trends in the deep learning era. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[227] Kerem C Tezcan, Christian F Baumgartner, Roger Luechinger, Klaas P Pruessmann, and Ender Konukoglu. MR image reconstruction using deep density priors. *IEEE Transactions on Medical Imaging*, 2018.

[228] Juan J Cerrolaza, Yuanwei Li, Carlo Biffi, Alberto Gomez, Matthew Sinclair, Jacqueline Matthew, Caronline Knight, Bernhard Kainz, and Daniel Rueckert. 3D fetal skull reconstruction from 2DUS via deep conditional generative networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 383–391. Springer, 2018.

[229] Deep spectral-based shape features for Alzheimer's disease classification, author=Shakeri, Mahsa and Lombaert, Herve and Tripathi, Shashank and Kadoury, Samuel and Alzheimer's Disease Neuroimaging Initiative and others, booktitle=International Workshop on Spectral and Shape Analysis in Medical Imaging, pages=15–24, year=2016, organization=Springer.

[230] Julian Krebs, Hervé E. Delingette, Boris Mailhé, Nicholas Ayache, and Tommaso Mansi. Learning a probabilistic model for diffeomorphic registration. *IEEE Transactions on Medical Imaging*, 2019.

[231] Cheng Ouyang, Konstantinos Kamnitsas, Carlo Biffi, Jinming Duan, and Daniel Rueckert. Data efficient unsupervised domain adaptation for cross-modality image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 669–677. Springer, 2019.

[232] E. Hosseini-Asl et al. Alzheimer's disease diagnostics by adaptation of 3D convolutional network. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 126–130. IEEE, 2016.

[233] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel

Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[234] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015.

[235] Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017.

[236] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015.

[237] Bryce Goodman and Seth Flaxman. Eu regulations on algorithmic decision-making and a "right to explanation". In *ICML workshop on human interpretability in machine learning (WHI 2016), New York, NY. http://arxiv. org/abs/1606.08813 v1*, 2016.

[238] Yarin Gal. Uncertainty in deep learning. *University of Cambridge*, 1:3, 2016.

[239] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017.

[240] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports*, 7(1):1–14, 2017.

[241] Ryutaro Tanno, Daniel Worrall, Enrico Kaden, Aurobrata Ghosh, Francesco Grussu, Alberto Bizzi, Stamatios N Sotiropoulos, Antonio Criminisi, and Daniel C

Alexander. Uncertainty quantification in deep learning for safer neuroimage enhancement. *arXiv preprint arXiv:1907.13418*, 2019.

[242] Shi Hu, Daniel Worrall, Stefan Knegt, Bas Veeling, Henkjan Huisman, and Max Welling. Supervised uncertainty quantification for segmentation with multiple annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 137–145. Springer, 2019.

[243] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In *Advances in Neural Information Processing Systems*, pages 6965–6975, 2018.

[244] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[245] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014.

[246] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):142–158, 2015.

[247] Kristoffer Wickstrøm, Michael Kampffmeyer, and Robert Jenssen. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. *Medical Image Analysis*, 60:101619, 2020.

[248] Moritz Böhle, Fabian Eitel, Martin Weygandt, and Kerstin Ritter. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Frontiers in aging neuroscience*, 11:194, 2019.

[249] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems*, pages 7775–7784, 2018.

[250] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). *arXiv preprint arXiv:1711.11279*, 2017.

[251] Peter N Robinson. Deep phenotyping for precision medicine. *Human mutation*, 33(5):777–780, 2012.

[252] Carlo Biffi. *A composite general linear model approach for population-based association studies using three-dimensional cardiac MRI*. Imperial College London. MRes in Biomedical Research., 2016.

[253] Mark Woodbridge, Gianlorenzo Fagiolo, and Declan P O'Regan. MRIdb: medical image management for biobank research. *Journal of digital imaging*, 26(5):886–890, 2013.

[254] Jeanette Schulz-Menger, David A Bluemke, Jens Bremerich, Scott D Flamm, Mark A Fogel, Matthias G Friedrich, Raymond J Kim, Florian von Knobelsdorff-Brenkenhoff, Christopher M Kramer, Dudley J Pennell, et al. Standardized image interpretation and post processing in cardiovascular magnetic resonance: Society for Cardiovascular Magnetic Resonance (scmr) board of trustees task force on standardized post processing. *Journal of Cardiovascular Magnetic Resonance*, 15(1):35, 2013.

[255] O'Brien et al. Working Group on Blood Pressure Monitoring of the European Society of Hypertension International Protocol for validation of blood pressure measuring devices in adults. *Blood Press Monit*, 7(1):3–17, 2002.

[256] Yik Y Teo, Michael Inouye, Kerrin S Small, Rhian Gwilliam, Panagiotis Deloukas, Dominic P Kwiatkowski, and Taane G Clark. A genotype calling algorithm for the illumina beadarray platform. *Bioinformatics*, 23(20):2741–2746, 2007.

[257] Olivier Delaneau, Jean-Francois Zagury, and Jonathan Marchini. Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods*, 10(1):5, 2013.

[258] B. Howie and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 5(6):e1000529, 2009.

[259] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575, 2007.

[260] Carlo Biffi, Antonio de Marvao, Mark I Attard, Timothy JW Dawes, Nicola Whiffin, Wenjia Bai, Wenzhe Shi, Catherine Francis, Hannah Meyer, Rachel Buchan, et al. Three-dimensional cardiovascular imaging-genetics: a mass univariate framework. *Bioinformatics*, 34(1):97–103, 2018.

[261] David Freedman and David Lane. A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, 1(4):292–298, 1983.

[262] Anderson M Winkler, Gerard R Ridgway, Matthew A Webster, Stephen M Smith, and Thomas E Nichols. Permutation inference for the general linear model. *Neuroimage*, 92:381–397, 2014.

[263] Philip T Reiss, Armin Schwartzman, Feihan Lu, Lei Huang, and Erika Proal. Paradoxical results of adaptive false discovery rate procedures in neuroimaging studies. *NeuroImage*, 63(4):1833–1840, 2012.

[264] Alexandre Andrade, Anne-Lise Paradis, Stéphanie Rouquette, and Jean-Baptiste Poline. Ambiguous results in functional neuroimaging data analysis due to covariate correlation. *Neuroimage*, 10(4):483–486, 1999.

[265] Francisco Cribari-Neto and Wilton Bernardino da Silva. A new heteroskedasticity-consistent covariance matrix estimator for the linear regression model. *AStA Advances in Statistical Analysis*, 95(2):129–146, 2011.

[266] Yoav Benjamini and Daniel Yekutieli. Quantitative trait loci analysis using the false discovery rate. *Genetics*, 171(2):783–790, 2005.

[267] Karl J Friston, Keith J Worsley, Richard SJ Frackowiak, John C Mazziotta, and Alan C Evans. Assessing the significance of focal activations using their spatial extent. *Human brain mapping*, 1(3):210–220, 1994.

[268] Pau Medrano-Gracia et al. Atlas-based analysis of cardiac shape and function: correction of regional shape bias due to imaging protocol for population studies. *J Cardiovasc Magn Reson*, 15(1):80, 2013.

[269] Christopher Semsarian, Jodie Ingles, Martin S Maron, and Barry J Maron. New perspectives on the prevalence of hypertrophic cardiomyopathy. *Journal of the American College of Cardiology*, 65(12):1249–1254, 2015.

[270] Philipp S Wild et al. Large-scale genome-wide analysis identifies genetic variants associated with cardiac structure and function. *J Clin Invest*, 127(5):1798–1812, 2017.

[271] Deepak Srivastava and Eric N Olson. A genetic blueprint for cardiac development. *Nature*, 407(6801):221, 2000.

[272] Joseph D Bronzino and Donald R Peterson. *Biomedical signals, imaging, and informatics*. CRC Press, 2014.

[273] Sebastian Schafer, Antonio De Marvao, Eleonora Adami, Lorna R Fiedler, Benjamin Ng, Ester Khin, Owen JL Rackham, Sebastiaan Van Heesch, Chee J Pua, Miao Kui, et al. Titin-truncating variants affect heart function in disease cohorts and the general population. *Nature genetics*, 49(1):46, 2017.

[274] Anish N Bhuva, Thomas A Treibel, Antonio De Marvao, Carlo Biffi, Timothy JW Dawes, Georgia Doumou, Wenjia Bai, Kush Patel, Redha Boubertakh, Daniel Rueckert, et al. Sex and regional differences in myocardial plasticity in aortic stenosis are revealed by 3D model machine learning. *European Heart Journal-Cardiovascular Imaging*, 21(4):417–427, 2020.

[275] Jingyu Liu and Vince D Calhoun. A review of multivariate analyses in imaging genetics. *Frontiers in neuroinformatics*, 8:29, 2014.

[276] Gabriella Captur, Carolyn Y Ho, Saskia Schlossarek, Janet Kerwin, Mariana Mirabel, Robert Wilson, Stefania Rosmini, Chinwe Obianyo, Patricia Reant, Paul Bassett, et al. The embryological basis of subclinical hypertrophic cardiomyopathy. *Scientific reports*, 6(1):1–11, 2016.

[277] Filippos Triposkiadis, Javed Butler, Francois M Abboud, Paul W Armstrong, Stamatis Adamopoulos, John J Atherton, Johannes Backs, Johann Bauersachs, Daniel Burkhoff, Robert O Bonow, et al. The continuous heart failure spectrum: moving beyond an ejection fraction classification. *European heart journal*, 40(26):2155–2163, 2019.

[278] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.

[279] Sukrit Narula, Khader Shameer, Alaa Mabrouk Salem Omar, Joel T Dudley, and Partho P Sengupta. Machine-learning algorithms to automate morphological and

functional assessments in 2D echocardiography. *Journal of the American College of Cardiology*, 68(21):2287–2295, 2016.

[280] Jelmer M Wolterink, Tim Leiner, Max A Viergever, and Ivana Išgum. Automatic segmentation and disease classification using cardiac cine mr images. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 101–110. Springer, 2017.

[281] Esther Puyol-Antón, Bram Ruijsink, Bernhard Gerber, Mihaela Silvia Amzulescu, Hélène Langet, Mathieu De Craene, Julia A Schnabel, Paolo Piro, and Andrew P King. Regional multi-view learning for cardiac motion analysis: application to identification of dilated cardiomyopathy patients. *IEEE Transactions on Biomedical Engineering*, 66(4):956–966, 2018.

[282] Qiao Zheng, Hervé Delingette, and Nicholas Ayache. Explainable cardiac pathology classification on cine MRI with motion characterization by semi-supervised learning of apparent flow. *Medical image analysis*, 56:80–95, 2019.

[283] Daniel Rueckert, Luke I Sonoda, Carmel Hayes, Derek LG Hill, Martin O Leach, and David J Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE transactions on medical imaging*, 18(8):712–721, 1999.

[284] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[285] Milind Y Desai, Steve R Ommen, William J McKenna, Harry M Lever, and Perry M Elliott. Imaging phenotype versus genotype in hypertrophic cardiomyopathy. *Circulation: Cardiovascular Imaging*, 4(2):156–168, 2011.

[286] Ghalib A Bello, Timothy JW Dawes, Jinming Duan, Carlo Biffi, Antonio De Marvao, Luke SGE Howard, J Simon R Gibbs, Martin R Wilkins, Stuart A Cook, Daniel Rueckert, et al. Deep-learning cardiac motion analysis for human survival prediction. *Nature machine intelligence*, 1(2):95–104, 2019.

[287] James R Clough, Ilkay Oksuz, Esther Puyol-Antón, Bram Ruijsink, Andrew P King, and Julia A Schnabel. Global and local interpretability for cardiac MRI classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 656–664. Springer, 2019.

[288] Esther Puyol-Antón, Bram Ruijsink, James R Clough, Ilkay Oksuz, Daniel Rueckert, Reza Razavi, and Andrew P King. Assessing the impact of blood pressure on cardiac function using interpretable biomarkers and variational autoencoders. *arXiv preprint arXiv:1908.04538*, 2019.

[289] Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics*, 15(4):869–877, 2005.

[290] Steffen E Petersen, Nay Aung, Mihir M Sanghvi, Filip Zemrak, Kenneth Fung, Jose Miguel Paiva, Jane M Francis, Mohammed Y Khanji, Elena Lukaschuk, Aaron M Lee, et al. Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in caucasians from the uk biobank population cohort. *Journal of Cardiovascular Magnetic Resonance*, 19(1):18, 2017.

[291] Nikita Nogovitsyn, Roberto Souza, Meghan Muller, Amelia Srajer, Stefanie Hassel, Stephen R Arnott, Andrew D Davis, Geoffrey B Hall, Jacqueline K Harris, Mojdeh Zamyadi, et al. Testing a deep convolutional neural network for automated hippocampus segmentation in a longitudinal sample of healthy participants. *NeuroImage*, 197:589–597, 2019.

[292] Christian Ledig, Andreas Schuh, Ricardo Guerrero, Rolf A Heckemann, and Daniel Rueckert. Structural brain imaging in alzheimer's disease and mild cognitive impairment: biomarker analysis and shared morphometry database. *Scientific reports*, 8(1):11258, 2018.

[293] Giovanni B Frisoni, Rossana Ganzola, Elisa Canu, Udo Rüb, Francesca B Pizzini, Franco Alessandrini, Giada Zoccatelli, Alberto Beltramello, Carlo Caltagirone, and Paul M Thompson. Mapping local hippocampal changes in Alzheimer's disease and normal ageing with MRI at 3 Tesla. *Brain*, 131(12):3266–3276, 2008.

[294] Kai-kai Shen, Jurgen Fripp, Fabrice Mériaudeau, Gaël Chételat, Olivier Salvado, Pierrick Bourgeat, Alzheimer's Disease Neuroimaging Initiative, et al. Detecting global and local hippocampal shape changes in Alzheimer's disease using statistical shape models. *Neuroimage*, 59(3):2155–2166, 2012.

[295] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic back-propagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

[296] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *Advances in neural information processing systems*, pages 3738–3746, 2016.

[297] Carlo Biffi, Ozan Oktay, Giacomo Tarroni, Wenjia Bai, Antonio De Marvao, Georgia Doumou, Martin Rajchl, Reem Bedair, Sanjay Prasad, Stuart Cook, et al. Learning interpretable anatomical features through deep generative models: Application to cardiac remodeling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 464–471. Springer, 2018.

[298] Tom Rainforth, Adam Kosiorek, Tuan Anh Le, Chris Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better. In *International Conference on Machine Learning*, pages 4277–4285, 2018.

[299] David JC MacKay. Local minima, symmetry-breaking, and model pruning in variational free energy minimization. *Inference Group, Cavendish Laboratory, Cambridge, UK*, 2001.

[300] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[301] Susanne G Mueller, Norbert Schuff, Kristine Yaffe, Catherine Madison, Bruce Miller, and Michael W Weiner. Hippocampal atrophy patterns in mild cognitive impairment and Alzheimer's disease. *Human brain mapping*, 31(9):1339–1347, 2010.

[302] Anil Bas, Patrik Huber, William A. P. Smith, Muhammad Awais, and Josef Kittler. 3D Morphable Models as Spatial Transformer Networks. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

[303] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[304] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[305] Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

[306] Khaled Alfakih, Scott Reid, Tim Jones, and Mohan Sivananthan. Assessment of ventricular function and mass by cardiac magnetic resonance imaging. *European radiology*, 14(10):1813–1822, 2004.

[307] Tristan Whitmarsh, Ludovic Humbert, Mathieu De Craene, Luis M Del Rio Barquero, and Alejandro F Frangi. Reconstructing the 3d shape and bone mineral density distribution of the proximal femur from dual-energy x-ray absorptiometry. *IEEE transactions on medical imaging*, 30(12):2101–2114, 2011.

[308] Sami P Väänänen, Lorenzo Grassi, Gunnar Flivik, Jukka S Jurvelin, and Hanna Isaksson. Generation of 3D shape, density, cortical thickness and finite element

mesh of proximal femur from a dxa image. *Medical image analysis*, 24(1):125–134, 2015.

[309] Nicolas Duchateau, Maxime Sermesant, Hervé Delingette, and Nicholas Ayache. Model-based generation of large databases of cardiac images: synthesis of pathological cine mr sequences from real healthy cases. *IEEE transactions on medical imaging*, 37(3):755–766, 2017.

[310] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *European Conference on Computer Vision*, pages 484–499. Springer, 2016.

[311] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In *Advances in neural information processing systems*, pages 82–90, 2016.

[312] Edward J Smith and David Meger. Improved adversarial systems for 3D object generation and reconstruction. In *Conference on Robot Learning*, pages 87–96, 2017.

[313] Alberto Gomez, Veronika Zimmer, Nicolas Toussaint, Robert Wright, James R Clough, Bishesh Khanal, Milou PM van Poppel, Emily Skelton, Jackie Matthews, and Julia A Schnabel. Image reconstruction in a manifold of image patches: Application to whole-fetus ultrasound imaging. In *International Workshop on Machine Learning for Medical Image Reconstruction*, pages 226–235. Springer, 2019.

[314] Tania Bacoyannis, Julian Krebs, Nicolas Cedilnik, Hubert Cochet, and Maxime Sermesant. Deep Learning Formulation of ECGI for Data-driven Integration of Spatiotemporal Correlations and Imaging Information. In *International Conference on Functional Imaging and Modeling of the Heart*, pages 20–28. Springer, 2019.

[315] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pages 5099–5108, 2017.

[316] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *Iclr*, 2(5):6, 2017.

[317] Carlo Biffi, Juan J Cerrolaza, Giacomo Tarroni, Wenjia Bai, Antonio De Marvao, Ozan Oktay, Christian Ledig, Loic Le Folgoc, Konstantinos Kamnitsas, Georgia Doumou, et al. Explainable anatomical shape analysis through deep hierarchical generative models. *IEEE Transactions on Medical Imaging*, 2020.

[318] Carlo Biffi, Juan J Cerrolaza, Giacomo Tarroni, Antonio de Marvao, Stuart A Cook, Declan P O'Regan, and Daniel Rueckert. 3d high-resolution cardiac segmentation reconstruction from 2d views using conditional variational autoencoders. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1643–1646. IEEE, 2019.

[319] Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Chun-Yu Sun, and Xin Tong. O-cnn: Octree-based convolutional neural networks for 3D shape analysis. *ACM Transactions on Graphics (TOG)*, 36(4):1–11, 2017.

[320] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3D point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9621–9630, 2019.

[321] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[322] Sofia Ira Ktena, Sarah Parisot, Enzo Ferrante, Martin Rajchl, Matthew Lee, Ben Glocker, and Daniel Rueckert. Distance metric learning using graph convolutional networks: Application to functional brain networks. In *International Conference*

on Medical Image Computing and Computer-Assisted Intervention, pages 469–477. Springer, 2017.

[323] Salim Arslan, Sofia Ira Ktena, Ben Glocker, and Daniel Rueckert. Graph saliency maps through spectral convolutional networks: Application to sex classification with brain connectivity. In *Graphs in Biomedical Image Analysis and Integrating Medical Imaging and Non-Imaging Modalities*, pages 3–13. Springer, 2018.

[324] Jack Klys, Jake Snell, and Richard Zemel. Learning latent subspaces in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 6444–6454, 2018.

[325] Hongyoon Choi, Hyejin Kang, Dong Soo Lee, Alzheimer's Disease Neuroimaging Initiative, et al. Predicting aging of brain metabolic topography using variational autoencoder. *Frontiers in aging neuroscience*, 10:212, 2018.

[326] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3577–3586, 2017.