Interpretable machine learning through radiomics and attribute-regularized neural networks for cardiology

Irem Cetin

TESI DOCTORAL UPF / 2022

Directors of the thesis

Prof. Oscar Camara
Prof. Miguel Angel Gonzalez Ballester
Department of Information and Communication Technologies

Directors

Oscar Camara	Full Professor Universitat Pompeu Fabra Barcelona, Spain
Miguel Angel Gonzalez Ballester	ICREA Professor Universitat Pompeu Fabra Barcelona, Spain

Review Committee:

Daniel Rueckert	Full Professor Technical University of Munich Munich, Germany
Gemma Piella	Full Professor Universitat Pompeu Fabra Barcelona, Spain
Alistair Young	Full Professor King's College London London, United Kingdom

This work was carried out in the *Sensing in Physiology and Biomedicine (PhySense)* and *Simulation, Imaging and Modelling for Biomedical Systems (SIMBIOsys)* research groups, at BCN Medtech, Department of Information and Communication Technologies of Universitat Pompeu Fabra, Barcelona, Spain. This thesis was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 825903 (euCanSHare project) and a scholarship of the Department of Information and Communication Technologies at Universitat Pompeu Fabra (DTIC-UPF).

Resumen

El diagnóstico asistido por ordenador de enfermedades cardiovasculares (CVD) con resonancia magnética cardíaca (CMR) es un campo importante de investigación para el fenotipado avanzado de imágenes cardíacas. Las mediciones existentes sobre este tipo de datos, como la fracción de eyección y los volúmenes de las cámaras, son demasiado simples y, a menudo, no son capaces de captar cambios sutiles que afectan a las estructuras cardíacas durante las primeras etapas de la enfermedad. Por su parte, los algoritmos de aprendizaje automático (ML) aplicados a este tipo de datos se basan principalmente en la utilización de índices morfológicos. La radiómica CMR es una técnica emergente para el fenotipado cardíaco más profundo y preciso. Utiliza datos a nivel de píxel para obtener múltiples cuantificadores de forma y textura del tejido. Sin embargo, se enfrenta algunos desafíos como la falta de interpretabilidad y reproducibilidad. Otros tipos de métodos, como los de aprendizaje profundo (DL), revolucionaron las imágenes médicas. A pesar de estos métodos son capaces de aprender representaciones complejas a partir de los datos, una limitación de los mismos es que su naturaleza de "caja negra" provoca falta de explicabilidad e interpretabilidad de los resultados. Recientemente, la inteligencia artificial explicable (XAI) ha mostrado un excelente potencial en esta línea, donde los enfoques de XAI tienen como objetivo explicar cómo se realizan las elecciones de los sistemas de inteligencia artificial (IA). Los modelos basados en representación latente, como los autoencoders variacionales (VAE), tienen el potencial de aliviar las limitaciones anteriormente mencionadas de los modelos basados en DL, ya que pueden codificar atributos ocultos de los datos en un espacio latente de baja dimensión.

Esta tesis presenta enfoques interpretables basados en aprendizaje automático con técnicas de radiómica, aplicadas a imágenes de CMR, y con modelos basados en representación latente regularizada, con el objetivo de identificar cambios en la estructura cardíaca y en la textura del tejido. Usando conjuntos de datos multimodales, esta tesis se esfuerza por generar biomarcadores de imágenes que puedan ser utilizadas para la caracterización de diversas afecciones cardiovasculares. Esta tesis tiene tres contribuciones principales. En primer lugar, se desarrolló una metodología de análisis basada en radiómica sobre imágenes de CMR para cuantificar automáticamente distintos índices estructurales y de función cardiovascular. En segundo lugar, se realizó una de las evaluaciones más grandes y completas del uso de radiómica sobre imágenes de CMR para el fenotipado de las principales enfermedades cardiovasculares, empleando la base de datos de UK Biobank. Por último, se desarrolló una red neuronal regularizada a partir de atributos, con el objetivo de generar explicaciones sobre cardiopatías combinando biomarcadores

clínicos y marcadores radiómicos.

Palabras clave: radiomica, inteligencia artificial explicable, enfermedades cardiovasculares, interpretación, resonancia magnética cardiaca.

Translated from english by Guillermo Jiménez Pérez.

Abstract

Computer-aided diagnosis of cardiovascular diseases (CVD) with cardiovascular magnetic resonance (CMR) is an important research topic for advanced cardiac image phenotyping. Existing quantifiers, such as ejection fraction and chamber volumes, are overly simplistic and often do not capture subtle and complex changes that affect the heart structures at early disease stages. Machine learning (ML) approaches have primarily been concerned with shape indices. CMR radiomics is an emerging technique for deeper and more accurate cardiac phenotyping. It uses pixel-level data to derive multiple quantifiers of tissue shape and texture. Yet, it faces some challenges, including the lack of interpretability and reproducibility. Deep learning (DL) methods, on the other hand, revolutionized medical imaging. While these methods are capable of learning complex representations from data, a limitation of many of these models is that their black-box nature suffers from lack of explainability and interpretability. Recently, explainable artificial intelligence (XAI) has shown excellent potential in this line, where XAI approaches aim at explaining how the choices of artificial intelligent (AI) systems are made. Latent representation based models, such as Variational Autoencoders (VAEs) have the potential to alleviate the limitation of DL-based models as they are able to encode hidden attributes of the data in a low-dimensional latent space.

This thesis presents interpretable machine learning-based approaches through CMR radiomics and regularized latent-representation based models for identifying changes in cardiac structure and tissue texture due to various cardiovascular conditions from multi-modal datasets and endeavors to generate explanations from different imaging biomarkers. The contributions of this thesis are three-fold. Firstly, a CMR radiomics-based pipeline was developed to quantify cardiovascular conditions automatically. Secondly, one of the largest and most comprehensive assessments of CMR radiomics for image phenotyping of important cardiovascular diseases was carried out employing the UK Biobank dataset. Thirdly, a DL-based attribute-regularized network is proposed to generate explanations from cardiovascular pathological cases combining clinical biomarkers and radiomics signatures.

Keywords: radiomics, explainable artificial intelligence, cardiovascular diseases, interpretation, CMR.

This thesis would not have been possible without the inspiration and support of many wonderful people. First and foremost, I am incredibly grateful to my supervisors, Oscar and Miguel, for their invaluable advice, continuous support, and patience during my Ph.D. Thanks for offering advice and encouragement with a perfect blend of insight and humor.

I want to thank all the members of PhySense and SIMBIOsys. Thank you all for a cherished time spent together in the office and social settings and for providing tremendous support during my time in the lab.

I would like to acknowledge all my co-authors for their contribution to this work.

My gratitude extends to the great team at UPF; researchers, developers, and of course, administrative staff who have been incredibly friendly, helpful, and fun. Thank you for all the beautiful moments shared.

I would like to thank my friends and my Schlatzi, who supported me in this adventure; no words can express my gratitude - big hugs for you, for all the joyful moments.

Special thanks to all the people who, along the way, believed in me.

And finally, to my mom, to whom I dedicate this milestone. You deserve endless gratitude. This journey would not have been possible without you. Thank you for your love, support, and unwavering faith in me. - Bu tezi kendisine adadığım anneme, bu başarı sensiz mümkün olmazdı. Bu yolda hiçbir fedakârlıktan kaçınmayıp bana her anlamda destek olduğun ve bana olan inancını bir an bile kaybetmediğin için sana çok teşekkür ederim.

"Let the future tell the truth, and evaluate each one according to his work and accomplishments. The present is theirs; the future, for which I have really worked is mine"

Nikola Tesla

Contents

List of Figures

List of Tables

List of Acronyms

ACDC	Automated Cardiac Diagnosis Challenge
AF	atrial fibrillation
AI	artificial intelligence
ARV	abnormal right ventricle
AUC	area under the curve
BCE	binary cross-entropy
BMI	body mass index
bSSFP	balanced steady-state free precession
CAD	coronary artery disease
CHD	congenital heart disease
CMR	cardiovascular magnetic resonance imaging
CNN	convolutional neural network
CO	cardiac output
CRT	cardiac resynchronization therapy
CT	computed tomography
CVD	cardiovascular diseases
DCM	dilated cardiomyopathy

DL	deep learning
ECG	electrocardiography
ED	end-diastole
EDV	end-diastole volume
EF	ejection fraction
ES	end-systole
ESV	end-systole volume
GAN	generative adversarial network
GLCM	gray-level co-occurence matrix
GLDM	gray-level dependence matrix
GLRLM	gray-level run-length matrix
GLSZM	gray-level size-zone matrix
HCM	hypertrophic cardiomyopathy
HHD	hypertensive heart disease
IBSI	image biomarker standardization initiative
IHD	ischemic heart disease
KL	Kullback-Leibler
LA	left atrium
LASSO	least absolute shrinkage and selection operator
LIME	local interpretable model-agnostic explanation
LR	logistic regression
LV	left ventricle
LVCI	left ventricle cardiac index
LVCO	left ventricle cardiac output
LVEDV	left ventricle end-diastolic volume
LVEF	left ventricle ejection fraction
LVESV	left ventricle end-systolic volume
LVM	left ventricle mass
LVMi	indexed left ventricle mass
LVSV	left ventricle stroke volume
LVSVi	indexed left ventricle stroke volume

MAE	mean absolute error
MI	mutual information
MIG	mutual information gap
MINF	myocardial infarction
ML	machine learning
MLP	multilayer perceptron
MMD	maximum mean discrepancy
MR	magnetic resonance
MRI	magnetic resonance imaging
MYO	myocardium
NGTDM	neighbouring gray tone difference matrix
PET	positron emission tomography
RA	right atrium
RF	random forest
RFE	recursive feature elimination
RNN	recurrent neural network
ROC	receiver operating curve
ROI	region of interest
RV	right ventricle
RVEDV	right ventricle end-diastolic volume
RVEF	right ventricle ejection fraction
RVSV	right ventricle stroke volume
RVSVi	indexed right ventricle stroke volume
SAP	separated attribute predictability
SCC	spearman correlation coefficient
SFFS	sequential forward feature selection
SHAP	SHapley Additive exPlanations
SPECT	single-photon emission computed tomography
SSFP	steady-state free precession
SVM	support vector machines
US	ultrasound
VAE	variational autoencoder

VM ventricular mass

WT wall thickness

XAI explainable artificial intelligence

CHAPTER

1 Introduction

Machine learning (ML) is having a transformational effect in many sectors, including healthcare. Recent advances in artificial intelligence methodologies, such as deep learning (DL), are expected to revolutionize the way health monitoring and care is approached. However, such data-driven methods often lack interpretability and explainability, which is crucial for health applications. In this thesis, we focus on the development of novel interpretable machine learning methods, based on radiomics and DL, and their application for cardiovascular disease diagnosis and stratification.

This chapter introduces the motivation for the research undertaken in this thesis and aims at providing the reader with all the necessary background including, clinical and technical context. This chapter firstly starts introducing cardiovascular diseases (CVD) in Section 1.1.1. In Section 1.1.2 the role of cardiac imaging with a special focus on cardiovascular magnetic resonance imaging (CMR) is explained. This is followed by imaging-derived indices of cardiovascular function in Section 1.1.4. Section 1.2.1 introduces ML-based cardiovascular analysis and its challenges. After that radiomics based analysis is described, explaining its use both in cardiology and its limitations in Section 1.2.2. DL-based cardiovascular analysis is introduced in Section 1.2.3. explainable artificial intelligence (XAI) is introduced in Section 1.2.4. Finally, this chapter is concluded with a summary of the objectives and contributions of this research work and with an overview of thesis content in Section 1.3.

1.1. Clinical context

1.1.1. Cardiovascular diseases

Cardiovascular diseases are the leading cause of morbidity and mortality worldwide, representing 32% of all deaths globally (17.9 million), being 85% of these deaths due to heart attack and stroke, according to the World Health Organization (WHO) [1]. The decline in age-standardized mortality rates and in the incidence of coronary artery disease (CAD), also known as ischemic heart disease (IHD), in many countries illustrates the potential for prevention of premature deaths and for prolonging life expectancy.

The cardiovascular system consists of the heart, which is an anatomical pump with four chambers and an equal number of valves (see Figure 1.1). By contracting and relaxing in turns, it transports blood to different parts of the body through the vessels [2]. Heart muscle (myocardium) cells need oxygen to function properly. The oxygen is provided by the blood that comes from the coronary arteries. The reduction of blood supply to coronary arteries is known as ischemia. An ischemic event is generally caused by atherosclerosis in the coronary arteries, which occurs when the arteries become clogged, restricting blood flow to the myocardium. Without adequate blood flow from the coronary arteries, the heart cannot get enough oxygen and vital nutrients to work properly. Myocardial infarction (MINF), which is an irreversible damage to the myocardial tissue, occurs when there is a complete blockage of the arteries [3]. MINF can cause abnormal loading conditions in the myocardium. In time, these abnormalities result in shape alterations of the heart such as localized thinning of the wall and also, in some extreme cases might result ventricle aneurysm [4].

Heart failure being the final common stage of several CVD such as MINF, atrial fibrillation (AF), and valvular heart diseases, indicates a dysfunction of the heart's pumping ability [6]. Congenital heart disease (CHD), on the other hand, is an abnormality of the structure of the heart and usually occurs at birth. A variety of conditions can cause these abnormalities during embryonic development. These conditions can be treated via surgical intervention. Some examples of congenital heart disease are: cardiac shunts, valve abnormalities, aortic coarctation, and transposition of the great vessels [7].

Cardiomyopathy is a group of diseases that affects the heart muscle. The classification of cardiomyopathies can be done by separating them into genetic, mixed, and acquired [8]. The most common genetic cardiomyopathy is hypertrophic cardiomyopathy (HCM) which is a condition in which the ventricles and the septum (the wall separating the left and the right side of the heart) as seen in Figure 1.1,

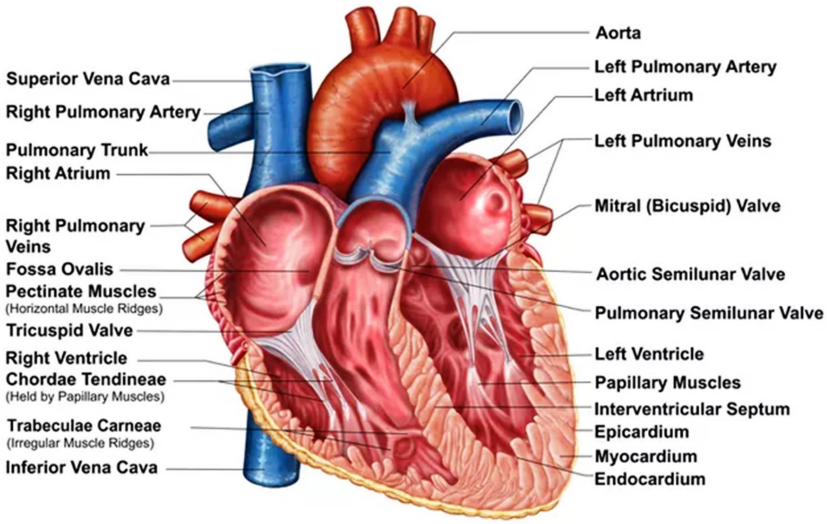

Figure 1.1: Internal view of the heart. Source: [5]

becomes thicker without an obvious reason. The thickened areas cause narrowing or blockage in the ventricles and make it harder for the heart to function properly [9]. HCM is the most common cause of sudden cardiac death in young people, especially in athletes. The most common mixed (genetic and non-genetic) cardiomyopathy is dilated cardiomyopathy (DCM) which is characterized by ventricular chamber enlargement (dilation) and systolic dysfunction with normal left ventricle wall thickness. The causes of DCM include infectious agents, particularly viruses, often producing myocarditis. DCM is also caused by a number of mutations in other genes encoding cytoskeletal/sarcolemma, nuclear envelope, sarcomere, and transcriptional coactivator proteins [8].

New therapeutic options for prevention and treatment of CAD have resulted in an increasing number of patients who survive a cardiovascular event. In developed countries the burden has shifted from the middle-aged to the elderly, and the prevalence of CAD increases exponentially with aging [10]. Therefore, early diagnosis of CVD plays an important role to reduce mortality and improve the quality of life for patients. In this regard, medical imaging is a major asset due to its unique capability for capturing in vivo the most subtle structural and functional changes in diseased organs.

1.1.2. Cardiovascular imaging

Non-invasive cardiac imaging techniques allow to diagnose and monitor the structure and function of the heart in a safe manner. These techniques aim to avoid

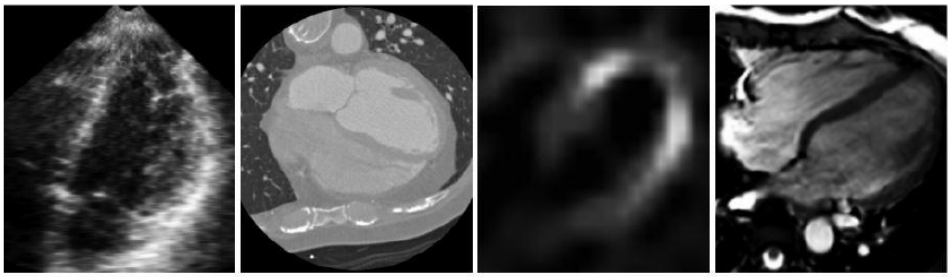

Figure 1.2: Images from different modalities: 3DUS= Three-dimensional ultrasound, CT= Computed Tomography, SPECT= Single Photon Emission Computed Tomography, MRI = Magnetic Resonance Imaging. Figures are taken from wikipedia.org

unnecessary invasive procedures, which require catheters to be inserted in the heart. As opposed to invasive techniques, non-invasive techniques are safe and can be used to diagnose a wide range of CVD [11]. Common modalities for cardiac imaging include computed tomography (CT), positron emission tomography (PET), single-photon emission computed tomography (SPECT), ultrasound (US) and magnetic resonance imaging (MRI) [12]. Images from different cardiac imaging modalities can be seen in Figure 1.2.

CT is based on tomographic reconstruction methods, which generate a high-resolution 3D volume from sets of X-ray images. Anatomical information of the heart and coronary arteries is clearly visible in CT. It is relatively inexpensive and provides fast image acquisitions. However, the disadvantage of this imaging modality is that it is based on X-rays, so patients are exposed to radiation [13].

SPECT and PET are both functional imaging techniques that measure radioactive tracers that are injected intravenously to the subject under examination. In cardiology, SPECT is mostly used to study myocardial perfusion, which investigates the function of the heart muscle. This technique is useful to quantify CAD and it is usually performed in conjunction with CT or MRI. The strength of SPECT is that it directly assesses tissue viability. However, the weakness of this modality is the low spatial and temporal resolution of the generated images [14].

Cardiac US, also known as echocardiography is the fastest, least expensive and least invasive modality to visualize structure of the heart. It uses high frequency sound waves (ultrasound) that can provide a real-time visualization of heart chambers. The low costs and absence of radiation associated with US makes it the most common technique to analyze the heart status for the first assessment. Its limitations include low signal-to-noise ratio and poor contrast between different tissues.

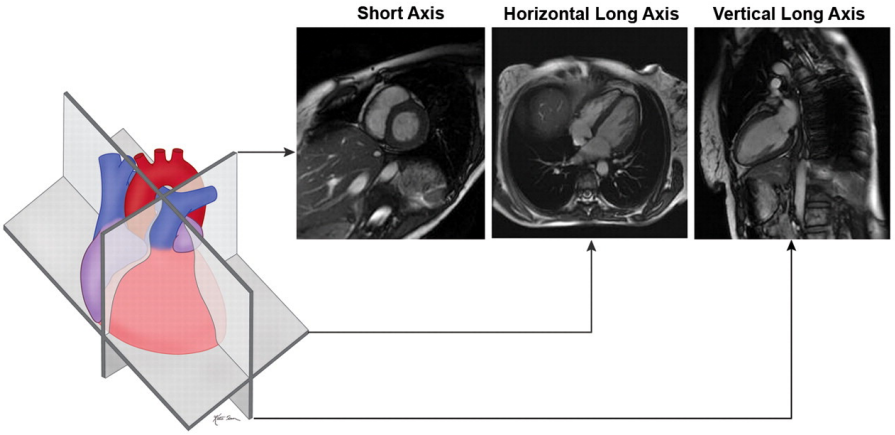

Figure 1.3: Orientation of major cardiac planes with respect to the heart. Image is taken from [21].

MRI is a technique that uses strong magnetic fields and radio-frequency waves to produce detailed images of the organs and tissues in the body. As US, MRI does not use ionising radiation and in contrast to US, MRI provides images with higher soft tissue contrast. These aspects make of MRI an accurate, reproducible modality and convenient for population studies, despite its high cost [15, 16].

Among the many different modalities of cardiac imaging, this thesis deals mostly with data from MRI, specifically cardiovascular magnetic resonance imaging (CMR), which represents the current gold standard for the assessment of cardiac structure and function.

1.1.3. Cardiovascular magnetic resonance imaging (CMR)

CMR is the reference imaging modality for assessment of cardiac structure and function; accordingly, its use in clinical practice is increasingly widespread [17]. It allows easy discrimination of soft tissues and blood pool without using any contrast agent. The main advantages of using CMR are image quality, noninvasiveness, accuracy and no exposure to radiation. CMR imaging is able to provide new insights to understand the pathophysiologic processes of underlying cardiac diseases, such as tissue damage from heart attack, reduced blood flow in myocardium, heart valve disorders, and abnormal right and left ventricular function [18–20].

CMR uses MRI pulse sequences techniques within a single study, leading to a comprehensive assessment of the cardiovascular system [23]. Most CMR images are acquired in breath hold and synchronized to the electrocardiography (ECG)

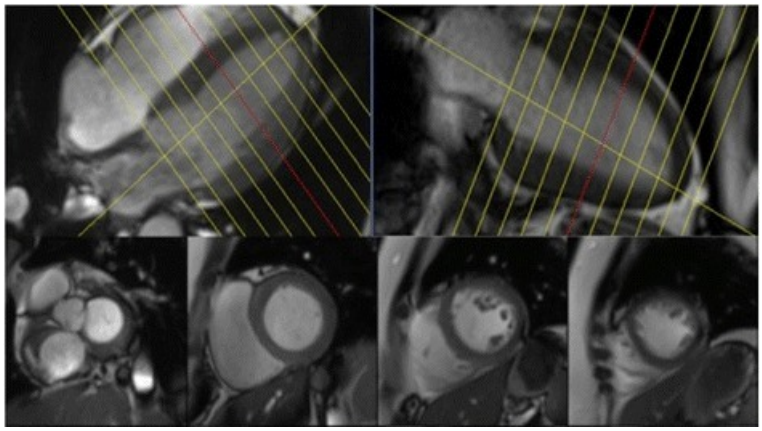

Figure 1.4: Planning used for the short-axis SSFP cine stack shown on 4- chamber and 2-chamber slices (top panel), with examples of some short-axis slices (bottom panel). Figure is taken from [22].

using cardiac gating [24]. Currently, the preferred MRI sequence for myocardial assessment is cine balanced steady-state free precession (bSSFP) [25].

CMR acquisition is acquired using stack of 2D slices from different imaging planes in different cardiac time frames (e.g. end-diastole (ED) and end-systole (ES) frames) to include the whole cardiac cycle. The standard cardiac planes include short axis, horizontal long axis (four-chamber view), and vertical long axis (two-chamber view) [22] as it be seen in Figure 1.3. Example 2D cine BSSFP CMR acquisitions for two- and four-chamber views can be found in Figure 1.4. The main cardiac structures that can be identified with CMR include right atrium (RA), left atrium (LA), right ventricle (RV), left ventricle (LV) and coronary arteries. For quantification, LV epicardial and endocardial contours are usually drawn at ED and ES frames (see Figure 1.5) which provides a way to calculate left and right ventricular volumes, mass and function [26].

1.1.4. Clinical indices of cardiac function

Heart performance, i.e., cardiac function is usually divided; systolic and diastolic. Systolic function refers to contraction of the ventricles in order to push blood into the arteries. Diastolic function refers to the relaxation of the ventricles to receive blood from the atria (filling). Different quantitative indices have been suggested to evaluate cardiac function at global and regional levels for both ventricles and are based on ventricular volumes at end-diastole (EDV) and end-systole (ESV).

For the clinical assessment, global indices include:

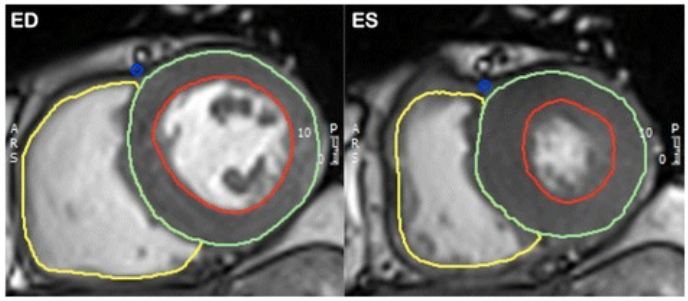

Figure 1.5: Epicardial (green), endocardial (red) and right ventricular (yellow) contours at end-diastole (ED) and end-systole (ES) [22].

- Ventricular mass (VM) is a global indicator of cardiac function and is calculated based on the volume contained within epicardial borders minus the chamber volume, multiplied by the density of the muscle tissue.

- Ejection fraction (EF) quantifies the quantity of the blood pumped out of the heart in each beat as a percentage. Usually to diagnose many CVD (e.g. cardiomyopathy, remodeling after MINF), left ventricle ejection fraction (LVEF) is used [2]. For example, high LVEF can often be seen in LV hypertrophy (e.g. hypertrophic cardiomyopathy (HCM)) [27]. In addition, right ventricle ejection fraction (RVEF) may also be decreased after myocardial infarction including parts of the RV.

- Cardiac output (CO) refers to the amount of systemic flow per minute. When the CO is normalized by body mass index (BMI), it is referred as left ventricle cardiac index (LVCI). Left ventricle cardiac output (LVCO) and LVCI are decreased in the case of congestive heart failure [28].

- Wall thickness (WT) is the thickness of the myocardium (MYO), and it is used to assess the systolic performance. It may be increased in conditions with increased afterload, such as hypertension. Some conditions also affect WT as a regional increase (with or without increased left ventricle mass (LVM)) typically referred to as asymmetric hypertrophy, such as in HCM [2]. In contrast, some cardiovascular diseases cause regional changes in wall thickness, for example MINF leads to a regional thinning as a consequence of cardiac remodeling [29]. LV wall thickening indicates the change of myocardial WT during systole.

1.2. Methodological context

1.2.1. Machine learning in medical image analysis

Machine learning (ML) methods in medical imaging show promise to automatize many medical image processing tasks. ML is a set of techniques that enable the extraction of meaningful patterns from examples [30]. Specifically, it can learn rules and identify patterns progressively from larger datasets without being explicitly programmed or given any prior assumptions [31]. ML techniques can perform either classification where discrete labels are determined, or regression, where continuous variables are estimated.

ML is now being applied to many areas of medical imaging, such as segmentation of anatomical structures [32], computer-aided diagnosis [33], medical image reconstruction [34] and clinical decision support [35]. ML methods can be devised for the segmentation and analysis of CMR. For example, ML methods can be used to determine LVEF from a CMR study [36].

In a ML model, important characteristics or *features* for a certain task are extracted from the images, by training on an example dataset (*training phase*). In *testing phase* the trained model is used to make a prediction on the data not seen previously in the training (called *testing data*). To evaluate performance of the model, it is of paramount importance to keep training data, which is used during model development, separated from the testing data which is used to evaluate the performance. However, another dataset (called *validation* dataset) is also used during the training phase to help determine the optimal design of the ML model. The validation dataset is used to optimize the parameters of the model and to ensure that the model does not overfit. Overfitting is a phenomenon that is seen when a trained model performs extremely well on training data but shows poor performance on unseen data (testing data) [36, 37].

The availability of reference labels in training data determines the type of ML as supervised or unsupervised [31, 38]. In supervised ML methods, the ground truth labels are provided in training data, e.g., cases with clinical cardiovascular indices (EF, ventricular volume, or mass), cases with imaging features (intensity, edges, or shape), cases with pathological status and images with ground truth segmentations. Supervised ML techniques are the most commonly used approaches, as learning from expert annotated data is the most intuitive way to mimic human performance in comparison to unsupervised ML methods where training data are given without ground truth labels [36]. With unsupervised learning, the data are processed with the aim of separating or clustering the training data into groups based on a measure of similarity or distance [30, 31].

The ability of ML techniques to analyze high dimensional data has facilitated the emergence of a novel field called *radiomics*. As this thesis explored the use of radiomics analysis in cardiac imaging, the next section will focus on radiomics.

1.2.2. Radiomics analysis

Radiomics is a novel image analysis technique, whereby digital medical images are converted to minable high dimensional data extracting hand crafted imaging features of shape and tissue character (referred to as *radiomics features*) [39]. These features may be used as predictor variables in ML models for diagnosis or clinical outcome prediction. Radiomics features are extracted from regions of interest (ROIs) and they can be roughly divided into shape-based, signal intensity-based (also called histogram-based), texture-based and transform-based features [40].

Shape-based radiomics features quantify the morphological characteristics and are extracted directly from the segmented region. Most of the shape-based features are conceptually simple, such as axes, minimum or maximum 2D and 3D diameters. Yet, some surface- and volume-based features use meshes and are more complex. For example, features include sphericity and compactness, that defines how the shape differs from a sphere [41].

Signal intensity-based radiomics features define the global gray-level histogram-based properties such as minimum, maximum, mean, median, percentiles and variance of gray-level values within the image region defined by the ROI [42]. More sophisticated features that quantify the shape of the intensity distribution of data including, skewness which defines asymmetry of the distribution of intensity values about mean value (negative skew: below the mean and positive skew: above the mean), and kurtosis which measures the tailedness of intensity distribution relative to a Gaussian distribution due to outliers (higher kurtosis: the mass of the distribution is concentrated towards the tail rather than towards the mean and lower kurtosis implies the reverse) [43]. Other features include histogram uniformity, standard deviation, mean absolute deviation and root mean squared.

Texture-based radiomics features are derived from different types of gray-level intensity matrices. These features are obtained using second- or higher-order statistics, and they can be categorized into several subgroups depending on the type of gray-level intensity matrix they are derived from (see Figure 1.6). Gray-level co-occurence matrix (GLCM) is a second-order histogram matrix that defines the spatial neighborhood relationship of pairs of pixels or voxels, with intensity values, in defined directions (horizontal, vertical, or diagonal for a 2D analysis or 13 directions for a 3D analysis), with a predefined distance between the pixels

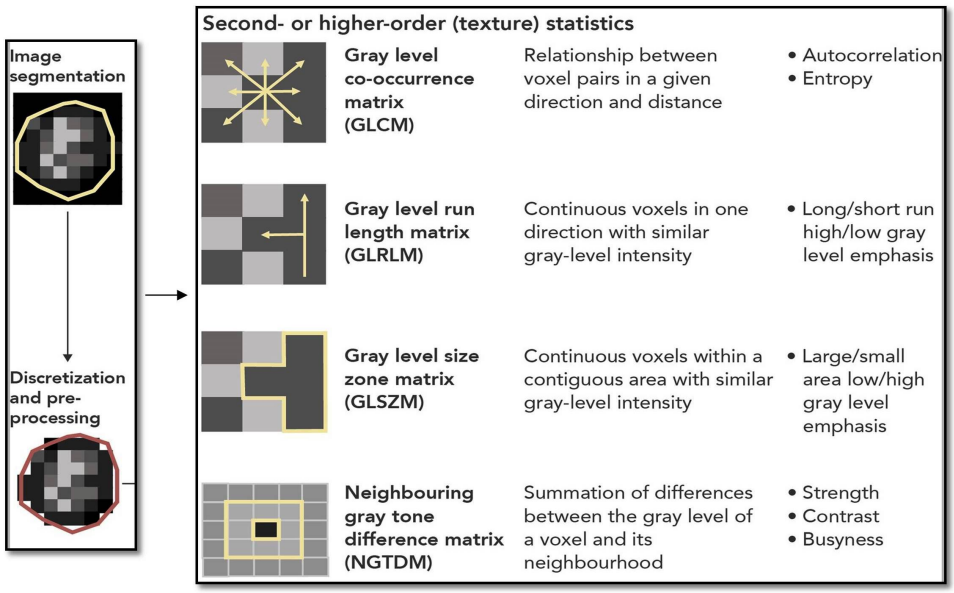

Figure 1.6: Characterization of textural features. For a given ROI, differences in the underlying histological structure will result in different texture patterns that can be described using higher-order features that reflect the unique spatial arrangement of voxels and their attenuation on computed tomography. Histogram-based first-order features only reflect the voxel attenuation distribution. Different texture patterns (same number of voxels with similar attenuation values but different location) may still have identical histogram and therefore similar first-order statistics.Figure is depicted from [44].

or voxels [41] as it can be seen in Figure 1.6. GLCM features include joint entropy, which is a measure of randomness or variability in neighborhood intensity values, maximum correlation coefficient assessing the complexity of the texture, inverse difference moment (also called homogeneity), which reflects gray-level homogeneity or order and contrast emphasizes gray-level variability between pixels or voxels belonging to a predefined pixel or voxel pair [45]. Gray-level size-zone matrix (GLSZM) quantifies the number of connected voxels that share the same gray-level intensity (so-called gray-level zone) [46]. GLSZM is not computed for different directions as in GLCM, but rather calculated for different pixel or voxel distances that define the neighborhood in 2 dimensions (8 neighboring pixels) or in 3 dimensions (26 neighboring voxels). GLSZM features include fractions, large and small area emphasis, that assess the percentage of pixels or voxels that are part of gray-level zones, defining the type of texture (fine or coarse). Other GLSZM features include gray-level non-uniformity, which measures homogeneity of the image, and size-zone non-uniformity, which measures the variability of size-zone volumes in the image. Gray-level run-length matrix (GLRLM) as

described in [47], quantifies gray-level runs, which are defined as the length in number of pixels, of consecutive pixels that share the same gray-level value, in one or more directions and in 2 or 3 dimensions. Following GLSZM definitions, GLRLM features also include fraction (the percentage of pixels or voxels within the ROI that are part of gray-level runs and thus reflects graininess [41]); short-run (greater value indicates shorter run length and more fine texture) and long-run emphasis (greater value implies longer run length and more coarse texture). Other features include, for example, gray-level and run-length non-uniformity, that assess the similarity of run lengths over different gray-level values and run lengths, respectively. Gray-level dependence matrix (GLDM) defines gray-level relationship between a central pixel or a voxel and its neighborhood within a predefined distance [48]. In this definition, a neighboring pixel or voxel is considered as being connected to the central pixel or voxel only if it meets certain dependence criterion in terms of a defined range of gray-level differences [41, 48]. GLDM features comprise of dependence non-uniformity, dependence variance, dependence entropy, and as in GLSZM and in GLRLM, features that represent the fraction, such that small dependence emphasis measures the distribution of small dependencies (greater value is an indication of less homogeneous texture) and large dependence emphasis which is a measure of the large dependences in the image (greater value implies more homogeneous texture). Finally, Neighbouring gray tone difference matrix (NGTDM) quantifies the difference between a gray-level value of a pixel or voxel and the average gray-level value of its neighbors within a predefined distance [49]. Key features include coarseness which quantifies the gray-level difference between the central pixel or voxel and its neighbors, and thus it is an indication of spatial rate of change in gray-level intensities, with a higher value indicating a lower spatial change rate and a locally more uniform texture. Contrast and busyness are both measure of gray-level intensity change where contrast assesses global change which is dependent on overall gray-level dynamic range and busyness quantifies rapid gray-level changes between the central pixel or voxel and its neighbors, so that an image with many small areas of different gray-level values will results in a greater busyness.

Transform-based features analyze gray-level patterns from the images that are transformed using different techniques, including Fourier, Gabor, and wavelet transforms [50]. Fourier transform [51] analyzes the spatial frequency information of the image, ignoring temporal and spatial representation, by converting the image in the spatial domain into a set of sine and cosine components. Gabor transform [52], on the other hand, describes textural patterns by sinusoidal functions and allows the spatial, temporal, and frequency representation of the signal [53]. Wavelet transforms allow a more detailed analysis of the image components, quan-

tify the frequency content of an image at different scales by decomposing it into high frequency (heterogeneity) and low frequency (homogeneity) regions [54, 55]. Wavelet decomposition of an image is done with a pair of so-called quadrature mirror filters, a high-pass and a low-pass filter [56].

Radiomics workflow

Standard radiomics workflow as illustrated in Figure 1.7 consists of several steps. Radiomics can be applied to any standard medical imaging modality without a defined requirement for dedicated acquisitions or imaging protocols.

Once the images to be analyzed are determined, for any radiomics framework, delineation of the ROI is a crucial step, as the ROI defines the region in which radiomics features are calculated. There are several approaches to determine the segmentations for the images; it can be done manually, semi-automatically (it can be done using segmentation algorithms such as thresholding and then applying manual corrections) and fully automatically (with ML or nowadays DL techniques). Manual and semi-automatic segmentations are the most used methods especially in radiomics analysis, however they are time-consuming and observer bias. DL-based image segmentation is rapidly emerging, and many different algorithms have already been proposed to segment various structures in medical images. Although DL-based automated image segmentation is the best option as it avoids intra- and inter-observer variability of radiomics features, the generalizability of trained algorithms is currently the biggest drawback and applying those algorithms on a different dataset often results in complete failure [57]. For this reason, further research is needed to develop robust and generalizable automated image segmentation.

Image preprocessing is the step just before radiomics feature extraction which aims to homogenize images from which radiomics features are extracted. This can be done with respect to pixel spacing, gray-level intensities, and bins of the gray-level histogram. Preliminary results have shown that the robustness of the extracted radiomics features depends mostly on the segmentations and image processing techniques [59–61]. For this reason, the selection of these techniques is crucially important for reproducible research [57]. After image preprocessing, radiomics features are extracted within ROI. Extraction of features can be performed using dedicated open-source software packages such as pyRadiomics [62]. As there are many different ways to calculate those features, use of the Image Biomarker Standardization Initiative (IBSI) guideline is recommended [42]. This guideline offers a consensus for standardized feature calculations from all radiomic feature matrices.

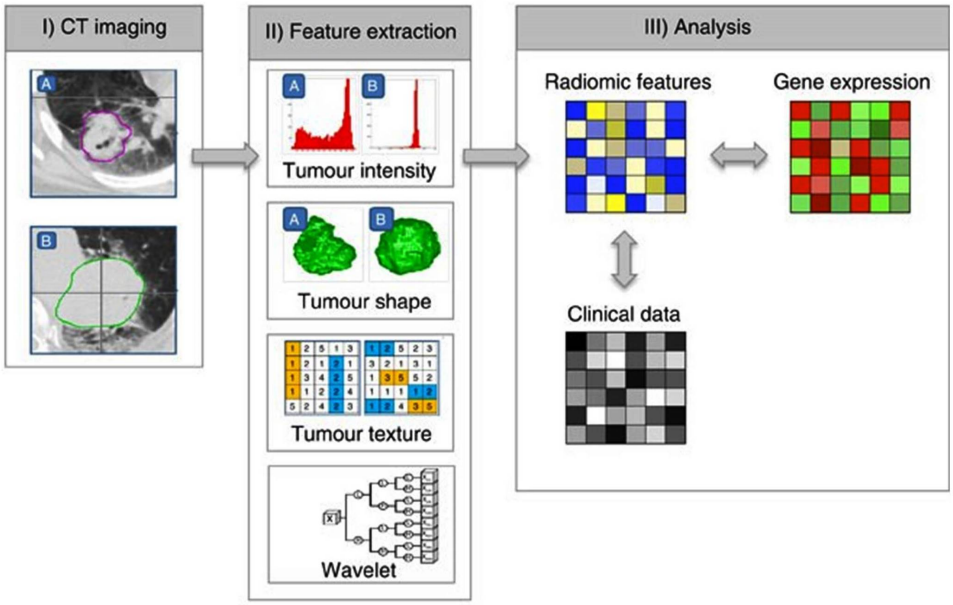

Figure 1.7: Strategy for extracting radiomics analysis of CT images with tumor. (I) Delineation of the tumour areas by human experts, on all CT slices. (II) Radiomics features are extracted from within the defined tumour ROI on the images, quantifying tumour intensity, shape, texture and wavelet texture. (III) For the analysis, the radiomics features are compared with clinical data and gene-expression data. Figure is taken from [58].

Depending on the experimental setting (i.e., type of the software package used for feature extraction, number of filters, number of image transformations) applied during the process, the number of extracted radiomics features to deal with can vary between 100s and 1000s. The higher the number of features in a model and the lower the number of cases in the study, the higher the risk of a model to overfit. For that, reducing the number of features during the feature selection step is of paramount importance to generate valid and generalizable results and to identify an optimal set of radiomics features to be used for model building. Although there are some rules of thumb for deciding the optimal number of features to select regarding the size of data, there is no true evidence for these rules in the literature [57]. Feature selection, or dimensionality reduction, is a multistep process leading to the exclusion of the features that are less informative, redundant, unstable, and provide repetitive information. There are several ways for feature selection. The most commonly followed feature selection pipeline [40, 63–65] starts with the exclusion of non-reproducible, non-informative, and non-robust features then the feature importance is evaluated in a ML framework. Various techniques are available for this step, such as recursive feature elimination [66], random forest al-

gorithms [67], least absolute shrinkage and selection operator (LASSO) [68] and sequential forward or backward feature selection algorithm [69]. The remaining non-correlated and highly relevant features are used, in most of the approaches, to train the model for the respective task.

Analysis of model performance is done with an independent external dataset. This step is crucial for assessment of performance and generalizability of the proposed model. Model performance is thus evaluated using several metrics such as sensitivity, specificity, receiver operating curves and area under the curve (AUC). Additionally, the model performance can also be compared with clinical indices and gene expression, depending on the task. Finally, the main motivation of radiomics analysis is that certain radiomics features are used as a predictor of particular disease states, the identified radiomics features (*radiomics signatures*) may be used to classify clinical outcomes.

Radiomics in cardiology

Although radiomics has been applied most prominently in the field of oncology, recently there has been an interest in applying radiomics in cardiology. One of the earliest attempts in cardiology was done to distinguish certain clinical conditions such as cardiac amyloid [70] and hemochromatosis [71] in echocardiography. Promising results were also obtained with computed tomography imaging studies to analyze coronary artery plaques and perivascular fat [50, 72]. However, limited proof-of-concept studies and difficulties with reproducibility of these aforementioned studies, have demonstrated the potential value and feasibility of CMR radiomics [36, 64, 73–75].

Myocardial tissue characterization in cardiology is a crucial but challenging task to distinguish various cardiovascular diseases. Although shape-based radiomics features have shown good performance in oncology, radiomics analysis in CMR mostly rely on intensity-, texture- and transform-based features such as wavelet transform [76–79]. The application of radiomics analysis to CMR imaging has recently shown great success in order to provide further insights into complex tissue alterations and pathology of cardiovascular diseases [64, 74, 80]. For example, several studies show that cardiomyopathy can be differentiated from healthy cohort using texture-based features [81, 82], as well as textural changes are observed in the myocardium of patients with acute myocarditis [83].

In CMR radiomics any image from CMR study can be used to extract radiomics features, however the short-axis stack is the most convenient to analyze with radiomics, because of existing endocardial and epicardial contours which can be used to define the ROI, avoiding extra segmentation steps [75]. Radiomics anal-

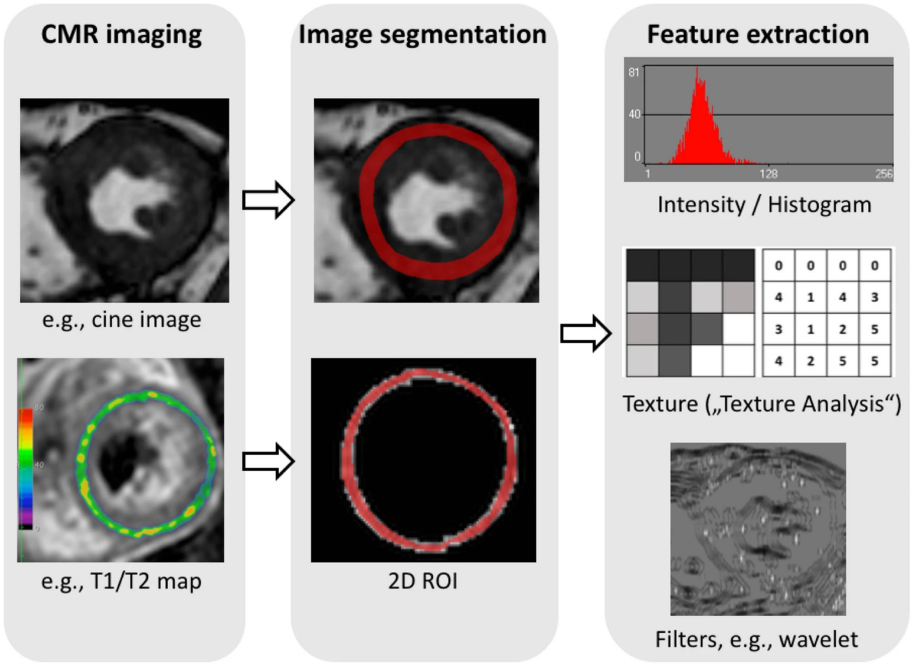

Figure 1.8: Radiomics pipeline in CMR. Radiomic feature extraction can be performed on all types of CMR images, e.g. cine images or T1 / T2 maps. The myocardium is segmented then feature extraction is performed. After extracting a high number of quantitative features, high-level statistical modelling is applied in order to perform either classification or prediction tasks. Figure is taken from [36].

ysis in CMR follows the standard radiomics framework as can be seen in Figure 1.8. The ROI can define a single area (e.g. a ROI with a delineation of either LV or RV) or multiple areas together (e.g. a ROI with delineations of LV and RV). Cardiac motion information may also be captured through analysis of temporal images (analysis of images from ED and ES) or assessment of images from all cardiac cycle phases [84, 85].

Current limitations in radiomics

Although radiomics analysis has shown great potential for diagnostic, prognostic, and predictive purposes in numerous studies, including oncology and cardiology, it faces several challenges.

First and foremost, the large number of potentially available features is one of the primary challenges in radiomics studies where it leads to a risk of overfitting and presents the problem of radiomics model selection [78]. As the performance of

all ML systems depend on the quality of raw data and features use to train these models, it is also important in radiomics analysis to use an accurate dataset with minimal missing values and proper parameterization; yet this is often a challenge in *big data* studies that include a huge amount of cases from different sources [31].

The redundancy in radiomics features needs to be noted and accounted for, and thus proper validation approaches should be applied to minimize this risk [44]. Additionally, radiomics feature values are influenced by patient variabilities in addition to the variations in scanners and settings. For example, geometric structure of the ROI has a big impact on the levels of noise and presence of artifacts in an image which directly affect the intensity- and texture-based features [57].

The reproducibility of radiomics studies is often poor because of insufficient reporting of experimental settings, lack of standardization, or limited open-source code and data [86]. Finally and perhaps most importantly, the lack of interpretability of radiomics features, especially those derived from intensity- and texture-based features, is one of the major challenges that the radiomics studies face and thus often hamper the translation to clinical practice and the use in clinical decision support systems [57, 75, 87].

1.2.3. Deep learning in medical image analysis

Although ML models have shown great promise to fully automate many medical image processing tasks, it is only until introduction of DL models that they are reaching widespread use [88]. DL is a subclass of ML and uses artificial neural networks with hidden layers to make predictions directly from datasets without the need for extracting any discriminative hand-crafted features [89].

The reason behind the success of DL methods is their extraordinary ability to learn complex task-specific imaging features through the stacked layers of non-linear processing which constitutes a neural network [90, 91]. For example, in case of finding the contours of myocardium, DL methods simply learn the image features that are the most useful for predicting the location of the corresponding contours. This new learning paradigm, *end-to-end learning*, has been made possible only in the last decade because of advancements of high-tech central processing units (CPUs), graphics processing units (GPUs), and tensor processing units (TPUs), availability of big data to collect vast amounts of imaging data, developments of learning algorithms, open-source development libraries and freely available working examples [89, 90].

Due to these improvements, DL models have managed to outperform most of ML methods in a variety of medical imaging tasks. In 2017, for example, a challenge

was organized by Bernard et al. [92], the Automated cardiac diagnosis challenge (ACDC), aimed at evaluating the performance of different automatic methods for the classification of 150 subjects into 5 categories (healthy, HCM, DCM, ARV and MINF) as provided by clinical experts. Several approaches were proposed for this problem where the results clearly open the door to highly accurate and fully automatic analysis of CMR, mainly using DL models. Additionally, according to a survey [93], in the last six years to 2017, 300 DL-based approaches have been published in medical image analysis, including CMR, with the numbers growing exponentially.

In medical image analysis, a special type of neural network, called as *Convolutional neural network (CNN)*, is often used. A typical CNN is composed of three different types of layers: *convolution*, *pooling* and *fully connected* layers. Convolution and pooling layers carry out feature extraction employing a set of filters that are applied directly to the data. Fully connected layers, on the other hand, map the features into the final output, such as prediction of a clinical outcome. For the medical image segmentation and image reconstruction tasks, upsampling operations (*transposed convolutional layers*) are used to return the image dimensions back to the original input size. Finally, non-linearity is performed with a *softmax layer* which rescales the components producing a non-negative probability to each pixel class so that outputs sum up to 1 in the output layer.

There are several differences to note between deep CNN networks and radiomics studies. First, CNN does not require hand-crafted feature extraction, and they do not necessarily need segmentation of cardiac substructures by human experts, particularly for the classification tasks. Second, CNN networks involve many millions of weights to optimize and therefore they require a huge amount of data. Although the features outputting from the intermediate convolutional layers contain relevant information to the task under study, because of their black-box nature, it is often difficult to interpret why the network predicts what it predicts or why it fails.

Explainable artificial intelligence (XAI) is, therefore, an emerging research field to answer a question on how a DL network obtains a particular solution. As XAI constitutes the second main part of this thesis, the following section focuses on XAI, while addressing its use in medical image analysis.

1.2.4. Explainable artificial intelligence (XAI)

XAI has experienced significant growth recently due to the broad use of ML applications, specifically DL, that provide highly accurate models but lack of explainability and interpretability. Although the terms, *explainability* and *interpretability* are used interchangeably, there are some distinctions between them [94, 95], such

that the explainable models can be interpretable by default; however, the reverse is not always possible [96]. Interpretation is the extent to which a cause and effect can be detected in a system so that one can predict what is going to happen in the case the input or a parameter in a DL system, is changed. On the other hand, explainability is to expound a set of parameters that contribute to the output model decision [96].

Generally speaking, XAI can be categorized in model-agnostic and model-specific approaches. Model-agnostic approaches are applied in post-hoc analysis and not attached to specific model architecture and rely on a simple surrogate function to explain the predictions, whereas model-specific approaches only benefit from parameters of the individual models and can be used with or without post-hoc analysis [94, 97]. For example, for the complex models like SVM, CNN, RNN and ensemble models, a model specific and post-hoc XAI strategies are specified [98].

Recently model-specific XAI techniques have shown great success to emphasize the importance of different features of the high-dimensional input data providing an explanation of a representative instance. For example, in the case of classification in cancer imaging, an attention module can be applied to explain to the user what image fragments the model focuses on to produce clinical outcome [98–100]. Model-agnostic XAI techniques develop surrogate representations to approximate an interpretable model for the black-box approaches. For example, to evaluate a treatment response on different clinical symptoms, an interpretable decision tree can be employed to approximate more complex DL model [98]. A well-validated model-agnostic XAI tool, local interpretable model-agnostic explanation (LIME), is able to provide local explanations for a complex DL model in the neighborhood of an instance [101].

Moreover, latent representation based models, such as Variational autoencoder (VAE) or Generative adversarial network (GAN), have also become powerful tools in this direction [102–104], as their latent space is able to encode important hidden variables of the input data [105, 106]. In case of interpreting different features (so called *data attributes*) or imaging biomarkers, these approaches provide a way to see how and if these attributes have been encoded in the latent space [107–111].

1.3. Research goals and context

The objectives of the thesis and how the rest of the document is organized are as follows:

Chapter 2 describes the development of a radiomics-based algorithm for automatic quantification of complex cardiovascular conditions. The developed method was used to analyze cardiovascular diseases from ACDC MICCAI17 dataset and hypertensive patients from UK Biobank. The results of this chapter were published in :

- **Cetin I.**, Sanroma G., Petersen S.E., Camara, O., Gonzalez Ballester M.A., Lekadir K., A Radiomics Approach to Computer-Aided Diagnosis with Cardiac Cine-MRI. *Statistical Atlases and Computational Models of the Heart.* STACOM, 82-90 (2018).
- Bernard O., Lalande A., Zotti C., Cervenansky F., Yang X., Heng P-A., **Cetin I.**, et. al., Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?, *IEEE Transactions on Medical Imaging*, 2514-2525 (2018).
- **Cetin I.**, Petersen S.E., Camara, O., Napel, S., Gonzalez Ballester M.A., Lekadir K., A Radiomics Approach to Analyze Cardiac Alterations in Hypertension. *International Symposium on Biomedical Imaging.* ISBI, 640-643 (2019).

Chapter 3 describes the identification of CMR radiomics signatures for the quantification of five different cardiovascular risk factors, using large-scale biomedical dataset, UK Biobank. This chapter was published in:

- **Cetin I.**, Raisi-Estabragh Z., Petersen S.E., Napel, S., K. Piechnik S., Neubauer S., Camara, O., Gonzalez Ballester M.A., Lekadir K., Radiomics Signatures of Cardiovascular Risk Factors in Cardiac MRI: Results From the UK Biobank. *Frontiers in Cardiovascular Medicine*, Volume 7 (2020).

Chapter 4 describes the development of novel attribute-based approach using DL in cardiovascular pathological cases linking both with clinical and interpretable features and also exploring the association with radiomics features. This chapter is adapted from:

- **Cetin I.**, Camara O., Gonzalez Ballester M. A., Attri-VAE: attribute-based, disentangled and interpretable representations of medical images with variational autoencoders. Medical Image Analysis. *Medical Image Analysis*. (2022) [Submitted].

Chapter 5 summarizes the main ideas, contributions, limitations and future directions of the thesis.

CHAPTER

2

Radiomics Approach to Computer-Aided Diagnosis with CMR

2.1. Introduction

Despite continuous progress in clinical research and practice, cardiovascular diseases (CVD) remain the leading cause of mortality and morbidity globally [112]. In this context, cardiac imaging such as cardiovascular magnetic resonance imaging (CMR) is expected to play an essential role due to its ability to quantify structural and functional properties of the heart [113]. However, visual assessment of CVD using CMR remains challenging and labor-intensive due to the complexity of these diseases, mainly when the structural and functional disorders are subtle [17]. Quantitative assessment can be suboptimal for borderline cases through

This chapter is adapted from:

Cetin I., Sanroma G., Petersen S.E., Camara, O., Gonzalez Ballester M.A., Lekadir K., A Radiomics Approach to Computer-Aided Diagnosis with Cardiac Cine-MRI. *Statistical Atlases and Computational Models of the Heart.* STACOM, 82-90 (2018). https://doi.org/10.1007/978-3-319-75541-0_9

Bernard O., Lalande A., Zotti C., Cervenansky F., Yang X., Heng P-A., **Cetin I.**, et. al., Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?, *IEEE Transactions on Medical Imaging*, 2514-2525 (2018). https://doi.org/10.1109/TMI.2018.2837502

Cetin I., Petersen S.E., Camara, O., Napel, S., Gonzalez Ballester M.A., Lekadir K., A Radiomics Approach to Analyze Cardiac Alterations in Hypertension. *International Symposium on Biomedical Imaging.* ISBI, 640-643 (2019). https://doi.org/10.1109/ISBI.2019.8759440

existing clinical indices such as volumetric measures, ejection fraction (EF), and thickening measures.

Additionally, cardiovascular risk factors, particularly hypertension, remain as the major risk factor for developing CVD and cardiac events. Approximately 77% of people who have a first stroke and 70% of people who have a first heart attack have hypertension [114]. While not directly linked to the heart, this condition can induce longitudinal alterations in the heart over a long period, well before symptoms of cardiovascular disease development. Eventually, this can lead to significant cardiac diseases like heart failure and left ventricular hypertrophy. Therefore, it is of paramount importance to identify individuals with a risk of developing hypertension-related diseases at an early stage to apply preventive and corrective measures. Furthermore, there is a significant knowledge gap about which perturbations occur in the heart over time in hypertensive patients, leading to full-blown cardiovascular remodeling and dysfunctions. Consequently, more advanced automated techniques are needed to exploit the richness of the cardiac data to estimate diagnosis and the severity of the phenotype, which is often associated with prognosis.

This work proposes a radiomics approach to the automated image-based diagnosis of complex CVD for imaging phenotyping of cardiovascular alterations due to hypertension. Radiomics is the task of calculating a large number of imaging descriptors from delineated images, which has been developed and exploited mainly in oncology with promising results for tumor classification and treatment planning [115–121]. In CMR, radiomics analysis has been applied to describe changes in image appearance due to CVD only recently [36, 64, 73–75].

The proposed approach estimates a large number of radiomic features, including intensity, shape, and textural descriptors, and assesses their ability to discriminate between different clinical conditions automatically and robustly within a machine learning framework based on support vector machines (SVM). We used two separate datasets in this study: CVD obtained from a database of CMR cases corresponding to five different subclasses from the ACDC challenge of MICCAI 2017 and hypertensive patients acquired from UK Biobank.

The rest of the chapter is organized as follows. Section 2.2 introduces the state-of-the-art that is relevant for this study. Section 2.3 describes the employed databases. Section 2.4 details the methodology followed in this work. Section 2.5 addresses the results obtained, where Section 2.5.1 shows the results from the ACDC challenge, and Section 2.5.2 demonstrates the results from the patients with hypertension identified from UK Biobank. Finally, Section 2.6 discusses the obtained results, their implications on the feasibility of applying this pipeline to other use cases and summarizes the conclusion of this work.

2.2. Related works

2.2.1. Machine learning in cardiovascular analysis

Over the years, machine learning (ML) has been used to address many cardiac imaging issues, including developing fully automatic tools that will directly support clinical experts for decision making, quantification, and visual assessment of cardiac structure and function [36].

Several approaches that use geometrical information have been proposed based on eigendecomposition of the moving cardiac shapes [122–128]. Myocardial tissue characterization is also a crucial task in cardiovascular analysis, and it has also been heavily investigated. Engblom et al., for example, developed an automatic algorithm for quantifying myocardial infarction (MINF) cases based on the expectation-maximization algorithm and weighted intensity [129]. Fahmy et al. developed a novel automated cardiac scar quantification in hypertrophic cardiomyopathy (HCM) patients. They combat one of the problems of using thresholding techniques such that variations in CMR centers result in a model where accuracy and reproducibility remain a preeminent challenge. The same group has also shown the capability to use ML techniques for cardiac relaxometry for tissue characterization [130].

ML has also shown its potential to assess different clinical indices of cardiac function. Winther et al. proposed a deep learning (DL)-based algorithm to assess cardiac mass and functional parameters by automatically segmenting left ventricle (LV) and right ventricle (RV) epicardium and endocardium [131]. They achieved an outcome that is comparable to human experts. However, their small sample size must also be taken into account. In contrast, Bai et al. applied a fully convolutional neural network (CNN) on an extensive database containing 93500 images from 5000 patients to measure LV and RV mass [132]. DL methods have also been used to calculate other functional parameters from cardiac imaging, such as the determination of left ventricle ejection fraction (LVEF), which can subsequently be used to discriminate patients into different CVD using with or without hand-crafted imaging features [133].

Although clinical cardiac function indices are used to diagnose different diseases in today's clinical practice, there are many encouraging ML models for cardiovascular diagnosis. Generally, these techniques either use only the features that DL-based network extracts or use conventional and hand-crafted imaging features together in a machine learning framework. Zhang et al. developed a DL model to extract cardiac motion features from LV and used these features to diagnose MINF patients. They achieved an AUC score of 0.94 out of 299 cases [134]. Ad-

ditionally, Moreno et al. used an approach based on SVM and random forest (RF) to predict MINF and HCM. They obtained a 0.94 AUC score out of a relatively small dataset consisting 45 cases [135]. Puyol-Antón et al. has taken this approach a step further and used a database containing CMR and ultrasound (US) images and added clinical indices into their experimental setting to design an automatic diagnostic algorithm. They analyzed dilated cardiomyopathy (DCM) patients in an SVM framework [136].

Even though ML, particularly DL, methods are powerful tools that revolutionized the cardiac imaging field, they have several limitations that need to be addressed. Most ML methods lack robustness and reproducibility due to different scanners, vendors, sequences, spatial and temporal resolutions, reconstruction algorithms and parameters [75]. Furthermore, one of the major issues is their black-box nature since it is often unclear what information is used to come to a particular outcome.

2.2.2. Radiomics analysis

Radiomics, as defined by Gillies et al. [39], is a process of converting digital medical images into mineable high-dimensional data. Radiomics analysis extracts a large number of hand-crafted imaging features using different mathematical and statistical methods. Definition of the radiomics features and radiomics workflow can be found in Section 1.2.2. Radiomic features have been used so far primarily for cancer image quantification [39, 40], such as for the estimation of patient prognosis and treatment response based on the characteristics of the tumors as encoded by the image data. Its use in cardiology, on the other hand, is only recent.

Prior studies showed that radiomics features could encode different tissue characteristics [137]. For example, Aerts et al. employed radiomics analysis to determine tumor phenotype in lung and head-and-neck cancer patients. They extracted 440 features quantifying tumor shape, intensity, and texture from computed tomography (CT) data of 1019 patients. Their results reveal that prognostic radiomics features encode tumor heterogeneity, and obtained radiomics signature is associated with underlying gene-expression patterns [119]. Vallieres et al. extracted radiomics features from positron emission tomography (PET) and magnetic resonance imaging (MRI) images to predict lung cancer metastases in soft tissue sarcomas. They computed only texture-based radiomics features and used a logistic regression (LR) algorithm, achieving a performance of 0.98 ROC, with only four texture features [138]. Timmeren et al. identified radiomics signatures for survival prediction of non-small cell lung cancer patients. They extracted 1119 radiomics features from data consisting of 194 CT scans. After feature selection, they identified 149 relevant features.

Furthermore, in a study exploring the impact of three ML framework variables in radiomics pipeline, namely feature selection, classification, and the number of selected features, for radiomics based survival prediction, Parmar et al. [139] compared different combinations of 13 feature selection methods and 11 ML classifiers based on 440 radiomics features extracted from CT images of 231 head-and-neck cancer patients. After evaluating each combination with multifactor analysis of variance (ANOVA) on ROC and AUC, they found that the choice of ML classification methods is the primary factor in performance variations (accounted for 29.02% of the total variance). In contrast, the classifier and feature selection interaction results in 14.02% of the total variance. These results indicate that selecting the appropriate combination of feature selection and ML models is extremely important on the performance [120].

Recently radiomics started being exploited in cardiology. Baessler et al. employed texture-based analysis for the diagnosis of subacute and chronic MINF. Five texture features identified to discriminate ischemic scar and normal myocardium achieved an AUC score of 0.92 in combination with LR [140]. Similarly, another texture analysis using CMR images to detect non-viable segments in patients with MINF yielded an AUC of 0.84 [73]. This concept has also been shown to apply to the diagnosis of other CVD. Neisius et al. recently used radiomics analysis to differentiate between hypertensive heart disease (HHD) patients and HCM patients identifying six radiomics features in SVM framework [141].

2.3. Materials

In this study, two datasets were employed, one for CVD diagnosis conducted in the context of the MICCAI 2017 challenge on automated cardiac diagnosis challenge (ACDC) and the other for hypertensive cases obtained from UK Biobank dataset.

The ACDC database consists of 100 cases for training and 50 cases for testing, comprising short-axis CMR data at both end-dyastolic and end-systolic phases, as well as height and weight information for each subject. Five subclasses were included, namely (see examples in Fig. 2.1):

- Normal subjects (NOR).
- Patients with dilated cardiomyopathy (DCM).
- Patients with hypertrophic cardiomyopathy (HCM).
- Patients with abnormal right ventricle (RV)

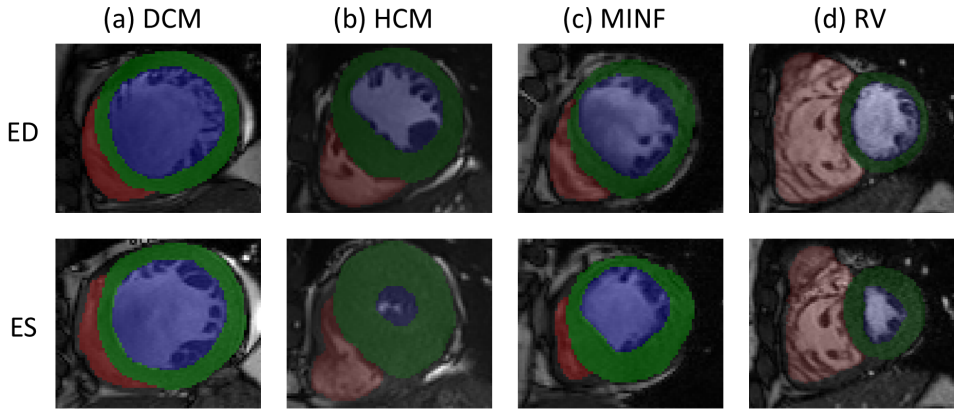

Figure 2.1: Examples of CMR images for the four abnormalities classified in this study (Top: ED, bottom: ES). DCM : dilated cardiomyopathy, HCM : hypertrophic cardiomyopathy, MINF : myocardial infarction, RV: abnormal right ventricle.

- Patients with myocardial infarction (MINF).

The images were acquired at the University Hospital of Dijon (France) by using 1.5 Tesla or 3 Tesla MR scans (Siemens Medical Solutions, Germany) with the following parameters depending on the examination: image sequence = SSFP CMR, slice thickness = 5 *mm* or 8 *mm*, inter-slice gaps = 5 *mm* or 10 *mm*, spatial resolution = 1.37 to 1.68 $mm^2/pixel$, number of frames = 28 to 40. This training dataset was then manually segmented for the LV, MYO and RV by an experienced manual observer at both ED and ES time frames.

On the other hand, hypertensive cases are part of the UK Biobank. The UK Biobank holds an exceptional amount of data (500,000 individuals) which includes biomedical data, physical measures, accelerometry, multimodal imaging including abdominal, brain, and CMR scans as well as whole-body DXA imaging, genome-wide genotyping, and longitudinal follow-up for a wide range of health-related outcomes [142].

For this work, 200 cardiac CMR images were randomly selected, including 100 hypertensive patients and 100 cases without hypertension. Both subgroups have no evidence of CVD. The cardiac images were acquired with 1.5 Tesla scan (MAGNETOM Area, Syngo Platform VD13A, Siemens Healthcare, Erlangen, Germany) and have an in-plane resolution of 1.8 x 1.8 mm^2, a slice thickness of 8.0 *mm* and a slice gap of 2 *mm* [143]. Manual annotation of the images was performed by clinical experts, resulting in a segmentation of the LV, MYO and RV boundaries [143].

2.4. Methodology

2.4.1. Segmentation

To segment the ACDC testing dataset, we employed a semi-automatic atlas-based segmentation approach. Atlas-based segmentation exploits the knowledge acquired from previously labeled training images to segment the target image [144]. For this, we used a publicly available cardiac atlas [145]. To this end, we first define manually six anatomical landmarks on each CMR case, more specifically at:

1. Mid-ventricular slice: RV insertion point next to the liver.
2. Mid-ventricular slice: A point on the RV free wall.
3. Mid-ventricular slice: RV insertion point next to the lung.
4. Mid-ventricular slice: A point on the LV free wall.
5. Apical slice: Apex.
6. Basal slice: Center of the base.

We then use the atlas-based technique described in [146] to extract the cardiac structures of interest, namely the LV, RV, and MYO. This is followed by user-friendly manual correction of the segmented contours to correct for potential errors using the ITK-SNAP tool[1]. Note that this segmentation approach will be only used to segment the testing dataset of the ACDC and that this work focuses only on the classification part of the ACDC challenge.

The hypertensive cases, including training and testing, were already manually segmented, and thus we only employed this segmentation on to ACDC dataset.

2.4.2. Radiomics feature calculation and extraction

Most existing techniques included in clinical practice use shape and motion indices such as EF, left ventricle end-diastolic volume (LVEDV) or left ventricle end-systolic volume (LVESV), and wall thickness (WT) to classify the subjects under investigation. It means that a lot of information produced by the image is lost during this operation, particularly imaging evidence in relation to the tissue

[1] http://www.itksnap.org/pmwiki/pmwiki.php

appearance in the blood pool, MYO and RV, as well as more complex morphological and functional information. But it is unclear which advanced indices could contribute to improved classification of cardiovascular cases. To address these issues, we propose a radiomics approach for computer-aided diagnosis and analysis in CMR.

In this work, we estimate a large pool of radiomic features from the segmented cine-MRI images, which will be then analyzed to extract the most powerful features for classification. In other words, we augment the set of indices to be leveraged for cardiac diagnosis by considering more complex shape and motion radiomic features, as well as advanced textural radiomic features. Specifically, we use 567 features (including height, weight, ED-ES duration, plus 188 features per structure: LV, MYO, RV at ED and ES) from ACDC dataset for cardiovascular diagnosis and 686 radiomics features from hypertensive patients and healthy cases for the purpose of capturing cardiac alterations.

Radiomics features were extracted based on three main categories using the PyRadiomics library [62], namely:

- Shape-based features (Volume, surface area, sphericity, compactness, diameters, elongation, etc.) capture geometrical alterations in the cardiac structures, while size features measure global and localized remodeling or dilation/hypertrophy due to a cardiovascular condition. The main shape/-size radiomics include sphericity, compactness, elongation, ratios, diameters, and main axes. In this study, these estimated shape/size radiomics be used to identify morphological remodeling occurring in hypertensive but not in non-hypertensive individuals.

- Intensity first-order statistics (e.g., mean, standard deviation, energy, entropy, etc.) inform on the distribution of the gray level values in the cardiac tissues without focusing on their spatial relationships. These include simple measures such as the mean intensity in a particular region of the tissue or the standard deviation, as well as more advanced measures such as skewness, uniformity, or entropy.

- Advanced textural features measure changes in the spatial relationships, local contrasts, and tissue homogeneity within the different cardiac structures. These radiomic features can be beneficial, for example, to capture potential alterations in the trabeculae, papillary muscles, and fibrosis in hypertensive vs. non-hypertensive subgroups. Different texture methods are included:

GLCM (autocorrelation, contrast, dissimilarity, homogeneity, inverse difference moment, maximum probability, etc.), GLRLM (short/long run emphasis, gray-level/run-length non-uniformity, etc.), NGTDM (coarseness, busyness, complexity, and strength), and Fractal Dimension.

Note that the first group of radiomics features consists of pure shape information. In contrast, the other remaining groups are intensity-based features, describing the intensity variations inside the cardiac structures and the complexity and repeatability of the tissue texture. In this study, we hypothesize that some of these radiomic values will be modified in the presence of cardiac abnormality in a way that is unique to each subgroup of patients when compared to normal individuals.

2.4.3. Radiomics feature selection

Due to the large number of radiomic features, radiomic-based analysis can easily suffer from overfitting due to the limited number of examples that can be realistically collected for training. As a result, it is of paramount importance to identify a smaller subset of radiomic features optimal for the respective task. In this work, we do this by using sequential forward feature selection (SFFS) [147], through which radiomic features will be added to the final subset one at a time until the classification becomes negatively impacted as a result of adding new radiomic features.

SFFS is a greedy search algorithm that aims to reduce an initial d-dimensional feature space to a k-dimensional feature subspace where $k < d$. In a nutshell, SFFS removes one feature at a time based on performance of the ML classification method, until termination criterion is met [147, 148].

In this concept, within a cross-validation scheme, the SFFS technique will enable to select sequentially, one at a time, the radiomic features that improve the overall classification of CVD vs. healthy cases and hypertensive vs. non-hypertensive individuals. The very first radiomic feature to be selected with this procedure is the one that has the highest predictive performance among all radiomic features. Sequentially, new radiomic features are added to provide complementary evidence for the classification and description of the cardiac phenotypes. In this work, we applied this feature selection algorithm using python-based library, mlxtend [148].

At the end of the procedure, radiomic features that have similar or overlapping distributions between the classes of interest are ignored, while those that contribute to the ML-based classification of pathological and normal hearts are included within the final set of optimal radiomic features, indicating their relevance for describing changes in asymptomatic hearts.

2.4.4. Classification method

To combine the heterogeneous radiomic features within a classification scheme that will learn to discriminate between the different patient subgroups and healthy individuals, we choose to use support vector machines (SVM) [149] due to well-known performance when classifying image data, particularly in small sample size.

An SVM model corresponds to a transformation of the examples to a hyperspace where the hyperplanes achieve a good separation with the largest distance to the nearest training-data point of any class (so-called functional margin). This ensures that the examples belonging to the different classes are separated as clearly as possible. New cases are then mapped onto that same hyperspace and classified based on their location with respect to the hyperplanes separating the different classes [150]. As such, it is suitable for cardiovascular disease classification as the challenge is precisely to identify subtle changes and differences between normal cardiac characteristics and those of pathological cases [151].

2.5. Results

This section demonstrates the results for the two databases described above: Section 2.5.1 shows the results from the ACDC dataset, and Section 2.5.2 explains the results using hypertensive cases from UK Biobank.

2.5.1. Results from the ACDC

For all experiments that were conducted with ACDC data, we used leave-one-out tests to evaluate the proposed method and measured accuracy as a proportion of correct classifications. Firstly, we assessed the accuracy of the CVD classifications by using only intensity-based or only shape-based radiomics features. For intensity-based radiomics, we obtained a maximal accuracy of 0.98 (two misclassifications) when using 13 optimal features. For shape radiomics, we achieved an accuracy of 1.0 (all cases correctly classified) but by using a total of 32 features. Subsequently, we combined intensity, shape, and patient information (height and weight) all together, and the forward feature selection results are provided in Figure 2.2. It can be seen that the best single feature only achieves a 0.62 accuracy. However, after adding three selected features to the classification task, the accuracy is improved beyond the 0.90 accuracy line to reach 0.91 and even 0.94 after five chosen features. A maximum accuracy of 1.0 (all cases correctly classified) is reached in training, and by combining 10 features only when linking intensity,

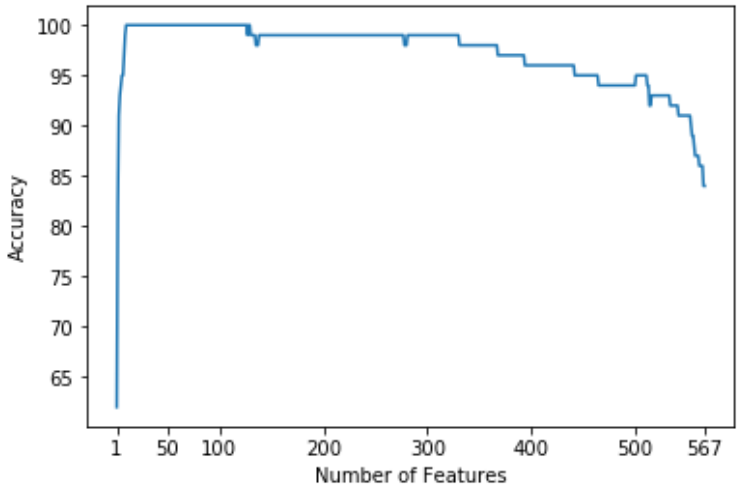

Figure 2.2: Training accuracy of the proposed CVD classification as a function of the number of radiomic features trained in the model.

	NOR	DCM	HCM	MINF	RV
Precision	1	0.85	0.9	0.95	1
Recall	0.87	1	0.86	1	1

Table 2.1: Precision, recall obtained by using the first five optimal radiomic features at accuracy of 0.94 in training.

shape, and patient information. Additionally, we obtained an accuracy of 0.92 in testing with the chosen features.

The shape of the curve in Figure 2.2 indicates the importance of feature selection, as after reaching the maximum accuracy, incorporating additional features leads to model overfitting and reduced accuracy. While these preliminary results are obtained in a small and controlled study, they are encouraging. In comparison, we obtained an accuracy of 0.84 when using all radiomic features and 0.86 when combining conventional clinical indices only, such as ejection fraction, cavity volumes, and body mass index (BMI).

To understand the behavior of the model, we evaluated the precision, recall (see Table 2.1) and confusion matrix (see Table 2.2) after selecting five features, at an accuracy of 0.94 in the training phase. It can be seen that the least accurately detected classes are for the HCM and DCM patients. However, after adding all optimal features, we finally reach a maximal accuracy of 1.0.

The selected list of optimal features is given in Table 2.3, which include one con-

	NOR	DCM	HCM	MINF	RV
NOR	20	0	0	0	0
DCM	0	17	0	3	0
HCM	2	0	18	0	0
MINF	1	0	0	19	0
RV	0	0	0	0	20

Table 2.2: Confusion matrix obtained by using the first five optimal radiomic features at accuracy of 0.94 with training dataset.

Name	Type	Frame	Structure	W/O	Alone
Volume	Conventional shape	ED	MYO	0.92	0.5
Surface Area to Volume	Advanced shape	ES	LV	0.88	0.62
Least Axis	Advanced shape	ES	LV	0.95	0.42
Maximum 2D diameter	Advanced shape	ED	LV	0.95	0.41
Maximum 3D diameter	Advanced shape	ES	RV	0.97	0.36
GLCM Inverse Difference	Intensity/textural	ES	RV	0.96	0.34
Compactness 2	Advanced shape	ES	LV	0.91	0.40
Maximum 3D diameter	Advanced shape	ES	MYO	0.96	0.47
Surface area	Advanced shape	ED	RV	0.97	0.29
Height	Patient Information	-	-	0.91	0.18

Table 2.3: List of 10 selected radiomic features as selected by the proposed technique for CVD classification. W/O: Accuracy without the feature. Alone: Accuracy using only this feature.

ventional shape index (volume), seven advanced shape radiomic features (e.g., compactness, least axis, surface area), one patient information (height), and one textural radiomic feature (GLCM inverse difference). This shows how multiple radiomics of different nature can be complementary to each other, which enables identifying all cases correctly. Also, the table shows that the features are well distributed among the three cardiac structures (LV, MYO, RV), as well as for the ED and ES frames.

To show the relevance of the selected features, we have added to the table the accuracy results by removing each feature from the SVM model (column W/O). It can be seen that the removal of each of these features negatively affects the final accuracy, which is reduced from 1.0 to 0.88 by removing the Surface Area to Volume feature, and to 0.96 by removing the Inverse Difference Intensity (GLCM) feature. This shows how these features can play a role in discriminating some of the challenging and ambiguous cases.

To further show the relevance of combining all of the selected features, we have

also added to the table in the last column the accuracy by using a single radiomic feature. It can be seen that on their own, these features do not enable a satisfactory classification, with the accuracy values varying between 0.18 (Height) and 0.62 (Volume). In particular, the Height variable cannot produce any meaningful classification on its own but contributes to the overall accuracy of the multi-radiomic model by normalizing with respect to size.

2.5.2. Results from the UK Biobank

In this study, 10-fold cross-validation tests are performed to select the optimal radiomic features that best separate hypertensive and non-hypertensive hearts. Note that after applying the proposed feature selection method, the classification accuracy, measured as the number of correct classification divided by the number of cases, reaches 0.8. This confirms the hypothesis that hypertension does alter the values of radiomics features even at the subclinical stage. This is further illustrated in Figure 2.3, which shows that using conventional imaging phenotypes of cardiovascular function (i.e., ejection fractions, stroke volumes, and volumes of left and right ventricles at ED and ES time frames) results in a low classification of the healthy and hypertension subgroups, with an AUC score of 0.62 ± 0.09. This is an expected result as the hypertensive individuals are asymptomatic with normal cardiovascular structure and functions as evaluated by the clinicians. Instead, the proposed radiomics model significantly improves classification with an AUC score of 0.76 ± 0.13.

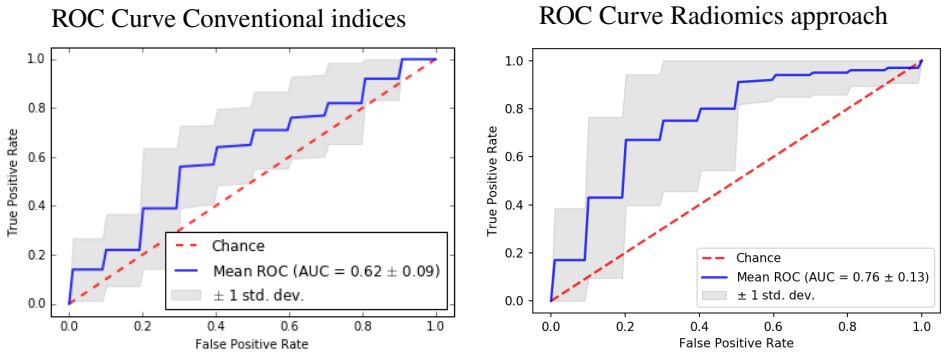

Figure 2.3: ROC curves using the proposed method with selected radiomics features (top) and conventional imaging phenotypes (bottom).

After demonstrating the relevance of the radiomics approach for discriminating hypertension and healthy subgroups, Table 2.4 lists the selected radiomics features, which sum to 11 radiomics features. It can be seen that all selected features are

Name	Frame	Structure	Alone	W/O
Homogeneity 1	ES	LV	0.495	0.77
Inverse variance	ES	LV	0.55	0.685
IDMN	ED	MYO	0.415	0.74
Sum of squares	ED	MYO	0.485	0.765
Large area emphasis	ED	LV	0.455	0.785
Zone entropy	ED	LV	0.56	0.725
LALGLE	ED	RV	0.485	0.765
Short run emphasis	ES	RV	0.505	0.79
Long run emphasis	ED	MYO	0.555	0.795
Coarseness	ED	MYO	0.62	0.76
GLNN	ES	MYO	0.635	0.71

Table 2.4: List of 11 radiomics features selected by the proposed method for discriminating the hearts of hypertensive and normal individuals. Alone: Classification accuracy using only this feature. W/O: Accuracy when removing the feature. ED: end-diastolic. ES: end-systolic. IDMN: Inverse difference moment normalized, GLNN: Gray level non-uniformity; LALGLE: Large area low gray level emphasis.

intensity- and texture-based radiomics, which indicate that the main changes due to hypertension are in the actual tissues rather than the geometry and size of the ventricles or myocardium. This also explains the inability of conventional indices to characterize changes due to hypertension as these typically focus on quantifying cardiac structure and function only. It is important to note that the selected radiomic features cover both ED and ES (second column in Table 2.4), and all three cardiac substructures (LV, MYO, RV - third column in table 2.4). Furthermore, the fourth column of the table lists the classification scores when using each feature alone, showing they do not separate well between the two subgroups, with the accuracy varying between 0.415 (Inverse difference moment normalized) and 0.635 (Gray level non-uniformity). This result confirms that hypertension-related changes are indeed small and subtle. It is the combination of all the features into a radiomic signature, describing multiple co-occurring changes in the heart that is best suited for optimal classification reaching a score of 0.8.

Finally, to further demonstrate the benefit of the radiomics approach, Figure 2.4 shows a comparison between two hearts corresponding to normal (above) and hypertensive (bottom) cases, respectively. Both hearts look visually normal, and furthermore, they have the same left ventricle ejection fraction (LVEF) value of 60%, indicating normal cardiac functions in both cases. In contrast, as shown in Table 2.5, the proposed radiomic signature enables to show apparent differences between the two cases in the radiomic space. Several of the normalized radiomic values (using z-normalization) indicate differences between the values in the normal and hypertensive cases, such as for the large area emphasis (-1.31 vs. 0.15), short-run

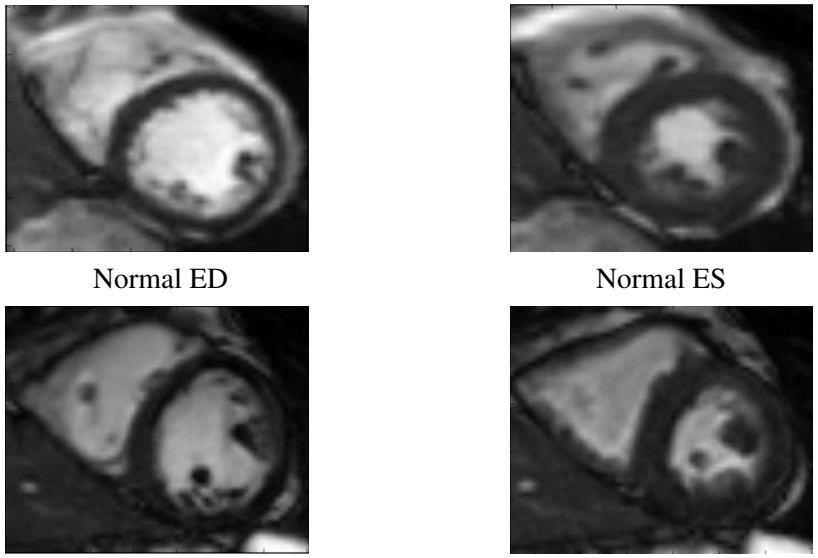

Figure 2.4: Images of hypertensive and normal cases with the same LVEF values and different radiomics signature.

	Original		Normalized	
	Normal	Hypertensive	Normal	Hypertensive
Homogeneity 1	0.543	0.546	-0.199	-0.133
Inverse variance	0.395	0.390	0.838	0.644
IDMN	0.981	0.992	**-1.3880**	**1.386**
Sum of squares	0.706	0.936	-0.467	0.161
Large area emphasis	**196**	**6910**	**-1.315**	**0.151**
Zone entropy	5.79	5.62	0.910	0.089
LALGLE	5	100	-0.885	-0.532
Short run emphasis	0.913	0.831	**2.217**	**-0.325**
Long run emphasis	1.76	2.80	-1.692	-0.401
Coarseness	0.00968	0.00203	**7.761**	**-1.090**
GLNN	**298**	**1670**	**-2.192**	**1.427**

Table 2.5: Original and normalized radiomics values for the two cases of Figure 2.4. IDMN: Inverse difference moment normalized, GLNN: Gray level non-uniformity; LALGLE: Large area low gray level emphasis.

emphasis (-2.21 vs. 0.32), and gray level non-uniformity (-2.19 vs. 1.42). The obtained results show textural differences between hypertensive and normal subgroups that cannot be captured using conventional clinical indices such as EF or visual examination.

2.6. Discussion and conclusions

In this chapter, we proposed the use of large amounts of radiomic features, integrating advanced shape and textural descriptors, to predict cardiovascular disease subgroups and hypertensive patients from two separate datasets. To the best of our knowledge, this is the first radiomics study performed to identify subclinical changes and quantify intermediate phenotypes associated with hypertension in otherwise healthy hearts.

The obtained results from the ACDC dataset suggest that radiomics are indeed capable to encode alterations in the anatomy and tissues of the affected cardiac structures. Furthermore, the feature selection results indicate that shape and intensity descriptors complement each other and their combinations enable to enhance the prediction power of the system, in particular for uncertain cases situated close to the boundary between two disease classes.

The results from hypertensive cases, on the other hand, indicate that the main changes are in the cardiac textures and tissues, which explains the inability of conventional imaging indices, which focus on structural and functional quantification, to identify these alterations. This work shows the promise of the proposed radiomics approach for analyzing subtle and more complex effects of other risk factors of heart disease such as high blood pressure. Future work includes clinical interpretation of the results (e.g. fibrosis formation), as well as application to other risk factors such as diabetes and cholesterol effects.

Finally, the high training accuracy in CVD classification, suggests that further evaluations with additional datasets are required to test this radiomics model in larger and more variable data samples. In particular, inter-subject variability due to semi-automatic segmentation of the boundaries will need to be assessed.

Data Availability

This work was conducted using the UK Biobank resource under Application 2964. UK Biobank will make the data available to all bona fide researchers for all types of health-related research that is in the public interest, without preferential or exclusive access for any person. All researchers will be subject to the same application process and approval criteria as specified by UK Biobank. For the detailed access procedure see `http://www.ukbiobank.ac.uk/register-apply/`. Additionally, the ACDC data is freely available and can be accessed in `https://www.creatis.insa-lyon.fr/Challenge/acdc/`

CHAPTER

3 Radiomics Signatures of Cardiovascular Risk Factors in Cardiac MRI: Results From the UK Biobank

3.1. Introduction

Cardiovascular magnetic resonance imaging (CMR) is the reference standard for assessing cardiac structure and function and is used widely in research and clinical settings. Routine assessment is reliant on visual inspection of CMR images for identifying global and local abnormalities; this is both labor-intensive and reader dependent [123–125, 127]. Existing quantifiers, such as ejection fraction (EF) and chamber volumes, are overly simplistic and often do not capture subtle and complex changes that affect the myocardium at early disease stages [152]. Current approaches are thus suboptimal for early disease detection and outcome prediction. Therefore, there is a need for novel, more advanced quantitative techniques for CMR image analysis to improve clinical diagnosis and risk prediction.

CMR radiomics is a novel image quantification technique whereby pixel-level data is analyzed to derive multiple quantifiers of tissue shape and texture [75]. Technological advancements and the availability of high computational power have

This chapter is adapted from: **Cetin I.**, & Raisi-Estabragh Z., Petersen S.E., Napel, S., Piechnik S. K., Neubauer S., Camara, O., Gonzalez Ballester M.A., Lekadir K., Radiomics Signatures of Cardiovascular Risk Factors in Cardiac MRI: Results From the UK Biobank. *Frontiers in Cardiovascular Medicine*, Volume 7 (2020). https://doi.org/10.3389/fcvm.2020.591368

allowed the deployment of machine learning (ML) methods with radiomics features to discriminate disease or predict outcomes [153]. A distinct advantage of radiomics modeling over unsupervised algorithms is the potential for explainability by identifying the most defining radiomic features in the model. It is thought that radiomics features correspond to alterations at both the morphological and tissue levels, and thus, the most defining characteristics of a particular condition (or its radiomics signature) may provide insights into its pathophysiology [119]. Within oncology, where radiomics is most well-developed, the incremental value of radiomics models for diagnosis and prognosis has been widely reported [39, 40, 119, 154–157]. In cardiology, early studies have shown promising results from CMR radiomics models for the discrimination of important conditions such as myocarditis, hypertrophic cardiomyopathy (HCM), and ischemic heart disease (IHD) [73, 83, 141].

While existing works have primarily focused on image phenotyping of established cardiovascular diseases, CMR radiomics may also provide incremental information to conventional approaches for improved quantification of cardiac alterations related to cardiovascular risk factors at the subclinical stage. This work thus presents the largest and most comprehensive assessment of the performance of CMR radiomics for image phenotyping of important cardiovascular risk factors, including diabetes, hypertension, high cholesterol, and smoking status, by using a large annotated CMR dataset from the UK Biobank.

The rest of the chapter is organized as follows. First, Section 3.2 describes the used dataset and its segmentation protocol. The methodology adopted in this work was explained in Section 3.3. The proposed radiomics workflow is depicted in Figure 3.2. Section 3.4 extensively summarizes the results and compares them with the existing literature. Finally, Section 3.5 contextualizes the obtained results. We have made our code publicly available in `https://github.com/iremcetin/radiomics_cardio_risk_factors`.

3.2. Materials

3.2.1. Population and setting

UK Biobank is a large-scale population health resource aimed at enhancing biomedical research and ultimately improving the prevention, diagnosis, and treatment of a wide range of serious and life-threatening illnesses [158]. Over 500,000 participants aged 40-69 years old were recruited from around the UK between 2006 and 2010. The UK Biobank holds an exceptional amount of data, including detailed lifestyle information, medical history, serum biomarkers, physical mea-

sures, and multi-modal imaging, including magnetic resonance imaging of the abdomen, brain, and heart [142]. Thus, UK Biobank provides the ideal platform for assessing the performance characteristics of novel quantitative biomarkers, such as radiomics, in discriminating common cardiovascular risk factors.

3.2.2. CMR imaging protocol

CMR cine images were acquired using a standardized UK Biobank protocol, detailed in a dedicated publication [143]. In brief, all scans were performed with a 1.5 Tesla scanner (MAGNETOM Area, Syngo Platform VD13A, Siemens Healthcare, Erlangen, Germany), with typical cine parameters as follows: TR / TE (repetition time / echo time) = $2.6/1.1$ *ms*, flip angle 80°, Grappa factor 2, voxel size 1.8 *mm* $\times$ 1.8 *mm* $\times$ 8 *mm*, and a slice gap of 2.0 *mm*. The actual temporal resolution of 32 *ms* was interpolated to 50 phases per cardiac cycle ($\sim$ 20 *ms*). The protocol includes a complete cine short-axis ventricular stack with a base to apex coverage acquired using balanced steady-state free precession (bSSFP) with one breath-hold per image slice.

3.2.3. CMR image segmentation

CMR scans of the first 5,065 UK Biobank participants that completed the imaging study were manually analyzed across two core laboratories (London, Oxford) using a pre-defined standard operating procedure, which is detailed elsewhere [159]. In brief, left and right ventricular (LV, RV) endocardial contours and LV epicardial contours were drawn in end-diastole (ED) and end-systole (ES) on the short axis stack images using the CVI42[1] post-processing software (Version 5.1.1, Circle Cardiovascular Imaging Inc., Calgary, Canada). These contours were used to define three regions of interest (ROIs) for radiomics analysis: RV blood pool, LV blood pool, and LV myocardium (MYO). All acquisitions were ECG gated, and thus ED was defined as the first phase in the sequence. ED was defined as the frame with the smallest LV cavity area by visual assessment detected at the mid-cavity level. Papillary muscles were considered part of the blood pool. Slices with more than 50% circumferential LV myocardium were included in LV contours. RV volume was defined as areas below the pulmonary valve plane identified by visual assessment.

[1] https://www.circlecvi.com/

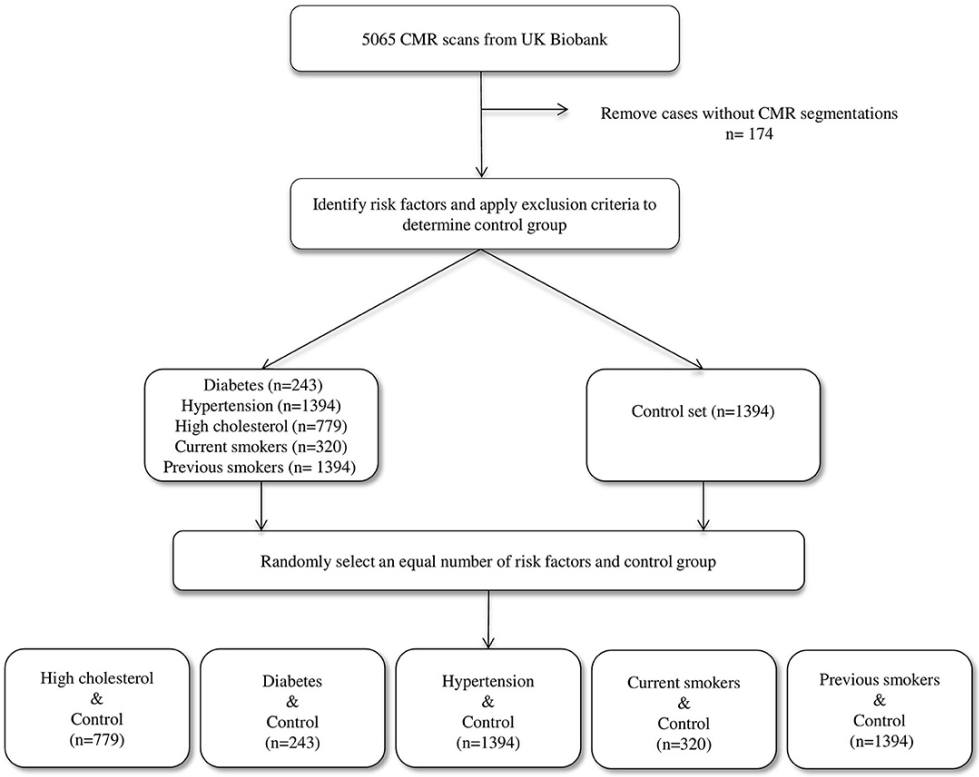

Figure 3.1: The data selection process.

3.3. Methodology and experimental setting

3.3.1. Selection of study sample

We considered the first 5,065 UK Biobank participants to complete CMR imaging. We excluded 174 individuals due to incomplete segmentations (having either one or more cardiac structures missing in the segmentation). From the remaining 4,891 individuals, a healthy cohort (n = 1,394) was defined by considering participants without known cardiovascular disease or risk factors. Diabetes (n = 224), hypertension (n = 1,394) and high cholesterol (n = 779) were taken from self-reported conditions. Smoking status was taken as a self-report of current (n = 320) or previous (n = 1,394) tobacco smoking. Participants positive for each risk factor were compared with an equal number of randomly selected reference healthy subjects to eliminate bias in the machine learning models due to class imbalance (see Figure 3.1).

3.3.2. Conventional CMR indices

For comparison and quantification of the added value of CMR radiomics, conventional CMR indices were also assessed, specifically: left ventricle end-diastolic volume (LVEDV), left ventricle end-systolic volume (LVESV), right ventricle end-diastolic volume (RVEDV), left ventricle end-systolic volume (LVESV), left ventricle stroke volume (LVSV), right ventricle stroke volume (RVSV), left ventricle ejection fraction (LVEF), right ventricle ejection fraction (RVEF), left ventricle mass (LVM).

3.3.3. Radiomics analysis

The overall radiomics workflow is depicted in Figure 3.2. Radiomics shape- and signal intensity-based features were extracted from the three segmented ROIs (LV blood pool: LV, LV myocardium: MYO, RV blood pool: RV) in end-diastole (ED) and end-systole (ES)). The analysis of the radiomics features in the myocardium may enable the identification of tissue-level changes due to cardiovascular risk factors. The inclusion of LV and RV cavities aims to identify changes in the shapes of each ventricle or the patterns of the trabeculation and papillary muscles. Automated extraction of radiomics features was performed using the open-source python-based radiomics library Pyradiomics[2] (version 1.3.0, October 2017) [62].

The customization of image preprocessing and feature extraction was performed with Pyradiomics default settings, including a gray value discretization with a bin width of 25 to extract the intensity-based and texture radiomics features. In total, 684 radiomics features were extracted per study (consisting of 114 radiomics features per cardiac structure: LV, RV, and MYO at two time-points of the cardiac cycle: ED and ES).

Shape-based radiomics features

Sixteen radiomics shape-based features were extracted per ROI at ED and ES (see supplementary material in chapter 6.1.1). Radiomics shape features describe geometrical properties of the defined ROI, and provide incremental value to existing CMR indices as they include conventional shape indices (e.g., cavity volumes) as well as more advanced geometric quantifiers (e.g., sphericity, flatness). They also have the potential to define disease-specific patterns of cardiac alterations beyond those possible with existing CMR indices.

Signal intensity-based radiomics features

[2] https://pyradiomics.readthedocs.io/en/latest/

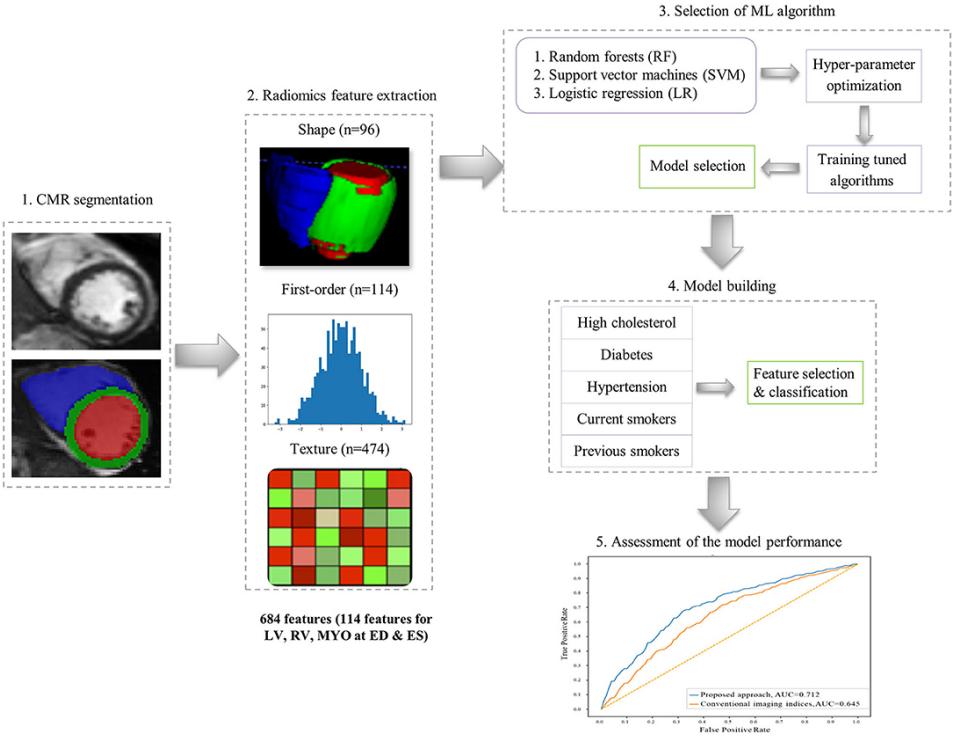

Figure 3.2: The proposed radiomics workflow.

Nineteen signal intensity-based radiomics features were extracted where they may potentially decode variations in cardiac tissue due to abnormalities induced by disease processes. First-order features are histogram-based statistics describing the global distribution of signal intensities within the defined ROI without consideration to their spatial relationships. These include simple measures such as the mean intensity or standard deviation and more advanced measures such as skewness, uniformity, or entropy (the complete list is provided in chapter 6.1.1).

Texture-based radiomics features

In contrast, texture radiomic features allow the quantification of spatial inter-pixel relationships using more advanced matrix analysis methods [81, 160]. Through this, signal intensities patterns within the ROI may be numerically quantified using pre-agreed mathematical definitions. Many texture patterns may be considered to quantify characteristics such as the complexity, heterogeneity, coarseness, or repeatability of the building blocks of the tissue. The idea is that these texture features may reflect myocardial tissue characteristics, which in turn reflect underlying disease processes. In this study, seventy-nine texture features were extracted from

each ROI per cardiac phase.

3.3.4. Identification of optimal radiomic signatures

This study aims to leverage feature selection and ML techniques to identify radiomics signatures that best describe the structural and tissue differences between a risk factor (at-risk) and healthy (no-risk) groups in CMR imaging. To this end, we use the risk factors as "proxy" output variables and build multiple machine learning models by varying the combinations of input radiomic features through systematic feature selection. We obtain various models (and thus multiple candidate radiomic signatures), and through statistical testing, one can select the best model and, therefore, the radiomic signature that best separates the at-risk and no-risk groups. Because these selected radiomics signatures differentiate at-risk from healthy individuals, they can be considered and analyzed as potential descriptors of the cardiac alterations due to the risk factors in question. Importantly, we use machine learning as a more advanced means to combine multiple radiomic features into risk-specific signatures while considering non-linear complementarities between the parameters.

For feature selection, we used the sequential forward feature selection (SFFS) method as it has demonstrated good performance in previous CMR radiomics studies [141], including the work described in previous chapter [161, 162]. The termination criterion was set to 2% in all experiments following literature standards, i.e., the process was stopped if an added feature did not increase model performance beyond the termination criterion. Ten-fold cross-validation was used in the feature selection process, rotating training, and validation folds (80% and 20% of the dataset, respectively), to obtain more robust estimates and improve generalizability.

We combined SFFS with classical ML algorithms [support vector machines (SVM), random forest (RF), logistic regression (LR)] to identify the combination of radiomics features that best define each studied cardiovascular risk/subgroup. For each ML method, hyperparameter optimization was performed to enhance the discrimination between no-risk and at-risk subgroups (see supplementary material in chapter 6.1.2 for the explanation of model selection). Implementation of the SFFS and the ML techniques was based on the mlxtend[3] (version 0.17.0) [148] and scikit-learn[4] (version 0.20.3) [163] python-based libraries, respectively.

[3] http://rasbt.github.io/mlxtend/
[4] https://scikit-learn.org/stable/

The selected radiomics features resulting from the SFFS algorithm and ML techniques were combined to create the radiomics signature that best encodes the changes in CMR induced by the different cardiovascular risk factors. To quantify the added value of the proposed radiomics approach, we built similar ML models/risk signatures using conventional CMR indices as input variables. All radiomics features and cardiac indices were normalized (to a mean of zero and standard deviation of one) to ensure they are equally weighted in all analyses. Note that individuals with multiple risk factors were not excluded. In the ML models, we set the outcome to each risk factor individually, which enabled the identification of the radiomics signatures specific to that risk factor.

In this work, we assess model performance (i.e., the ability of the radiomics signatures to discriminate at-risk vs. no-risk subjects) using receiver operating characteristic (ROC) curve and area under the curve (AUC) scores. We also report model accuracy, defined as a number of correctly discriminated no-risk vs. at-risk cases based on the radiomics signatures, divided by the total number of cases. Additionally, statistical tests were performed to assess the statistical significance of the differences between the various ML models by using McNemar's test for pairwise comparisons, as well as the Cochran's Q test, which is an extension of McNemar's test for the comparison of more than two models [164, 165].

3.4. Results and Discussion

3.4.1. Summary of subgroups and conventional CMR indices

The subjects included in the analysis are summarized in Table 3.1. Across all risk factor groups, there was a higher proportion of male participants (between 52.3% and 60.1% depending on the risk factor), whereas, in the healthy cohort, there were fewer men (42.5%). The average age across the risk groups was between 59 ($\pm$8) and 65 ($\pm$6) years, while it was equal to 60 ($\pm$7) years for the healthy cohort. As expected, there were differences in conventional CMR metrics between the at-risk subgroups and healthy subjects. In particular, on average, all risk groups had greater indexed left ventricle mass (LVMi) compared to the healthy cohort, with the most significant difference in the hypertensive group (50.3 g/m^2 vs. 46.3 g/m^2). All risk factor groups had lower indexed left ventricle stroke volume (LVSVi) and indexed right ventricle stroke volume (RVSVi) in comparison to the healthy cohort. There were also variations in chamber volumes, with different directions of difference depending on the risk category. Finally, it is worth noting that no statistically significant differences (Welch's t-test) in the conventional indices were found between the healthy and each at-risk subgroups, except for LVEF

	Diabetes	Hypertension	High cholesterol	Current smoker	Previous smoker	Healthy
	n=243	n=1,394	n=779	n=320	n=1,394	n=1,394
Male n(%)	146 (60.1%)	786 (56.4%)	460 (59.1%)	172 (53.8%)	729 (52.3%)	592 (42.5%)
Age mean (sd) years	64 (±7)	64 (±7)	65 (±6)	59 (±8)	63 (±7)	60 (±7)
LVEDVi (ml/m2)	73.4 (±13.8)	76.7 (±14.2)	75.0 (±13.9)	77.2 (±15.1)	76.9 (±14.8)	77.9 (±14.7)
LVESVi (ml/m2)	30.8 (±9.2)	31.6 (±9.3)	30.8 (±8.8)	32.5 (±9,4)	31.9 (±10.5)	31.6 (±8.8)
LVMi (g/m2)	49.1 (±9.6)	50.3 (±10.2)	48.6 (±9.7)	49.3 (±9.9)	48.3 (±10.1)	46.3 (±9.7)
LVEF (%)	58.5 (±7.3)*	59.2 (±6.9)	59.3 (±6.7)	58.3 (±6.9)	59.0 (±6.7)	59.7 (±5.9)
LVSVi (ml/m2)	42.7 (±8.3)	45.2 (±8.4)*	44.2 (±8.3)	44.7 (±8.9)*	45.1 (±8.2)	46.3 (±8.8)
RVEDVi (ml/m2)	77.2 (±14.5)	80.1 (±14.9)	79.1 (±14.9)	81.2 (±16.1)	80.8 (±14.8)	83.1 (±16.2)
RVESVi (ml/m2)	34.3 (±9.6)	34.8 (±9.7)	34.7 (±9.7)	36.3 (±10.4)	35.6 (±9.5)	36.8 (±10.5)
RVEF (%)	56.0 (±6.9)	56.9 (±6.7)	56.5 (±6.8)	55.7 (±6.9)	56.3 (±6.4)	56.2 (±6.3)
RVSVi (ml/m2)	42.9 (±8.2)	45.3 (±8.4)	44.4 (±8.5)	44.9 (±8.9)	45.2 (±8.3)	46.3 (±8.5)

Table 3.1: Summary of conventional CMR indices for the risk and healthy groups included in the analysis. LV: left ventricle, RV: right ventricle, EDV: end-diastolic volume, ESV: end-systolic volume, SV: stroke volume, EF: ejection fraction, LVM: left ventricle mass, i: indexed, absolute values divided by body surface area (calculated according to Du Bois formula). Values are given as mean ± standard deviation for continuous variables, and count (%) for categorical variables. *: Indicates statistical differences with respect to the healthy subgroup according to Welch's t-test.

in diabetes and LVSVi values in hypertension and current smokers (see Table 3.1).

3.4.2. Radiomics signatures have superior discriminatory performance over conventional CMR indices

In comparison to conventional indices, radiomics signatures provided better discrimination between healthy and at-risk subjects for diabetes (0.80 AUC for radiomics vs 0.70 for conventional indices), hypertension (0.72 vs. 0.69), high cholesterol (0.71 vs. 0.65), and previous smokers (0.63 vs. 0.60) (see Figure 3.3). The obtained models with radiomics vs. conventional indices were also compared using McNemar's test; the differences were found to be statistically significant for diabetes, hypertension, high cholesterol, and previous smokers but not for current smokers.

3.4.3. Comparison of the degree of discrimination achieved for each subgroup

The degree of discrimination (no-risk vs. at-risk hearts) achieved using radiomics models varied between the different cardiovascular risks, as these have different effects on the heart. The highest degree of discrimination with radiomics models was seen in diabetes (0.78), suggesting that radiomics features are particularly important in distinguishing diabetes-related cardiac changes. The smallest degree of separation was seen in previous smokers (0.61). High cholesterol, hyperten-

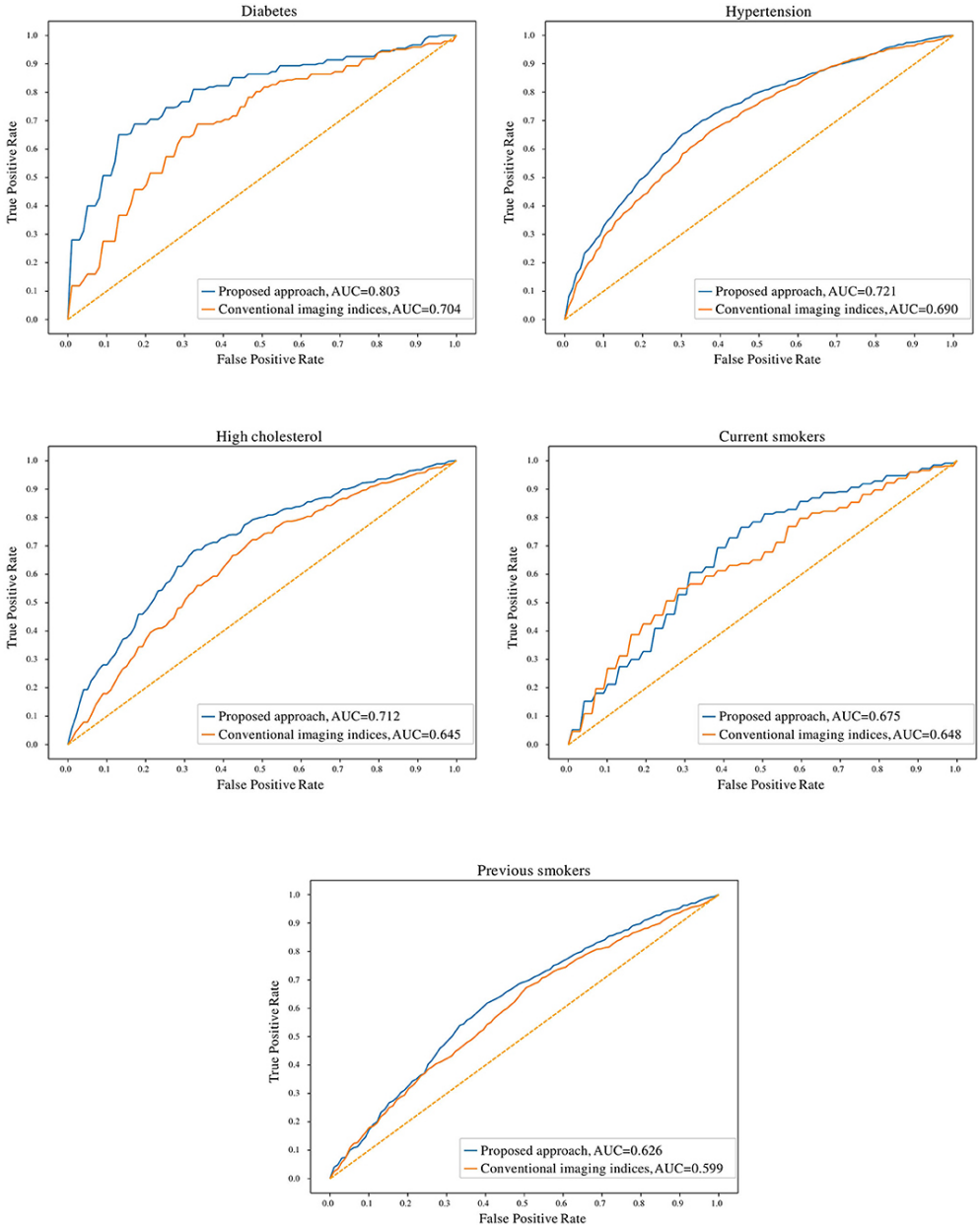

Figure 3.3: Receiver operating characteristic curves for radiomics and conventional CMR indices models for the five cardiovascular risk factor subgroups. AUC: area under the curve.

sion, and current smokers achieved similar degrees of separation by the radiomics

models (i.e., 0.68, 0.68, and 0.67, respectively).

3.4.4. The identified radiomics signatures for each cardiovascular risk factor

The identified radiomics signatures for each risk factor are described in Table 3.2. Overall, there was a more prominent role for shape- and texture-based features than first-order features. For instance, in diabetics, five of the eleven features included in the model were shape-based, and in the hypertension group, no first-order feature was selected. As expected, radiomics features from the LV blood pool and LV myocardium were the most relevant regions, with the RV blood pool having a minor role for the risk factors studied in this work.

In Table 3.3, we consider the most discriminative radiomics feature for each risk factor, i.e., the feature assigned the most important in the model, and compare it with the most discriminative conventional CMR measure, which was LVM for all risk groups. For all the subgroups, the mean value of the most important radiomics features and conventional CMR indices was significantly different in the risk factor vs. healthy cohorts ($p < 0.001$, Table 3.3). In addition, the single best radiomics feature outperformed the conventional CMR indices in its relevance for all risk factors. However, it was the combination of several radiomics features into a radiomic signature (Table 3.4) that provided the best overall discriminative power.

3.4.5. Summary of findings

This study described a methodology based on radiomics, machine learning, and feature selection to discover new discriminatory signatures in CMR. Based on over 5,000 datasets, we presented the largest and most comprehensive study to demonstrate the feasibility and performance of CMR radiomics for identifying new imaging signatures associated with important cardiovascular risk factors such as diabetes, hypertension, cholesterol, and smoking. Over conventional indices, we showed that radiomics enable improved quantification of alterations in both cardiac structure and tissue due to the effects of these risk factors. From the statistical tests performed in Table 3.1, it can be seen that the conventional indices do not capture statistically significant differences between the healthy vs. at-risk subgroups, with very few exceptions (LVEF values in diabetes, LVSVi values in hypertension, and current smokers). In contrast, McNemar's statistical tests comparing the radiomics models and the conventional indices show statistically significant differences between the two approaches for all cardiovascular risk factors, except for current smokers. This indicates that for diabetes, hypertension, high

CV risk factor	Radiomics signature	Type	ROI	Phase	Alone
High cholesterol	SD	Shape	MYO	ED	0.61
	Compactness	Shape	MYO	ED	0.60
	Skewness	First-order	LV	ED	0.59
	IMC	Texture	LV	ES	0.57
	GNN	Texture	RV	ED	0.55
	Contrast	Texture	RV	ES	0.52
	Median	First-order	MYO	ES	0.65
Diabetes	SVR	Shape	MYO	ED	0.61
	Energy	First-order	LV	ED	0.61
	Surface area	Shape	MYO	ES	0.58
	DV	Texture	LV	ED	0.57
	LAHGLE	Texture	MYO	ED	0.57
	Energy	First-order	LV	ES	0.57
	Flatness	Shape	RV	ED	0.56
	Surface area	Shape	LV	ES	0.55
	Max2D	Shape	RV	ED	0.50
	DA	Texture	LV	ES	0.44
Hypertension	SVR	Shape	MYO	ED	0.61
	Percentile 10	First-order	RV	ES	0.58
	IMC	Texture	LV	ES	0.55
	DNN	Texture	LV	ED	0.54
	SZNN	Texture	RV	ED	0.54
	GNN	Texture	MYO	ES	0.60
	DE	Texture	LV	ED	0.57
Current smokers	STD	First-order	MYO	ED	0.53
	Max2D	Shape	RV	ED	0.50
	LDLGLE	Texture	RV	ED	0.45
	SVR	Shape	MYO	ED	0.57
	Busyness	Texture	LV	ES	0.54
Previous smokers	Run entropy	Texture	MYO	ES	0.50
	Skewness	First-order	RV	ES	0.50
	RNN	Texture	RV	ED	0.49
	ZV	Texture	LV	ED	0.49

Table 3.2: Radiomics features selected for each risk factor. Features are presented in order of importance (accuracy using only one feature) in the model for each risk factor. Alone: model performance using each radiomic feature individually, SD: Spherical disproportion, DV: Dependence variance, DA: Difference average, DE; Dependence entropy, STD: Standard deviation, ZV: Zone variance, IMC: Informal measure of correlation, DNN: Dependence non-uniformity normalized, SZNN : Size zone non-uniformity normalized, LAHGLE: Large area high gray level emphasis, LDLGLE: Large dependence low gray level emphasis, GNN: Gray level non-uniformity; SVR: Surface area to volume ratio, Max2D: Max 2D diameter column, RNN: Run length non-uniformity.

CV risk factor	Single most defining feature	CV risk cohort		Healthy cohort		ACC
		Mean	SD	Mean	SD	
High cholesterol	Rad: Spherical disproportion MYO ED (S)	3.631	0.290	3.779	0.311	0.611
	Conv: LVM (g)	93.493	24.199	85.667	24.104	0.576
Diabetes	Rad: Median MYO ES (F)	67.887	9.058	74.652	10.514	0.658
	Conv: LVM (g)	97.856	24.250	85.931	25.024	0.605
Hypertension	Rad: Surface area to volume ratio MYO ED (S)	0.390	0.054	0.425	0.06	0.618
	Conv: LVM (g)	97.131	25.849	85.623	24.101	0.593
Current smokers	Rad: Gray level non uniformity MYO ES (T)	573.448	134.355	515.789	140.307	0.609
	Conv: LVM (g)	93.614	24.804	84.549	25.426	0.564
Previous smokers	Rad: Surface area to volume ratio MYO ED (S)	0.405	0.058	0.425	0.062	0.574
	Conv: LVM (g)	91.902	24.896	85.623	24.101	0.552

Table 3.3: Values of the best radiomics features (Rad) and the conventional CMR indices (Conv). Feature values from risk groups and healthy individuals were statistically significantly different for all selected features (Bonferroni adjusted p-value $< 0.05/684$). S: shape, F: first-order, T: texture, SD: standard deviation, ACC: accuracy, CV: cardiovascular, MYO: LV myocardium, ED/ES: end-diastole/systole, LVM: left ventricular mass (in grams, g).

CV Risk factor	Radiomics features					Clinical indices		
	#	S/F/T	LV/RV/MYO	ED/ES	ACC/AUC	#	LV/RV	ACC/AUC
High cholesterol	6	2/1/3	2/2/2	4/2	0.682/0.712	2	1/1	0.626/0.645
Diabetes	11	5/3/3	5/2/4	6/5	0.782/0.803	4	3/1	0.681/0.704
Hypertension	5	2/0/3	2/2/1	3/2	0.682/0.721	2	1/1	0.646/0.690
Current smokers	5	1/1/3	1/2/2	5/0	0.675/0.675	3	2/1	0.628/0.648
Previous smokers	6	1/1/4	2/2/2	3/3	0.612/0.626	2	1/1	0.579/0.599

Table 3.4: Selected number of radiomic features used for each risk factor and their discriminative accuracy, and results obtained based on conventional imaging indices and size information. #: total selected number of features, S: shape features, F: first-order radiomics, T: texture features, LV: left ventricle, RV: right ventricle, MYO: Myocardium, ED: end-diastole, ES: end-systole, ACC: accuracy (prediction performance), AUC: area under the curve.

cholesterol, and previous smokers, radiomics models provide incremental value in identifying structural and textural differences between healthy and at-risk subgroups.

3.4.6. Clinical interpretation of the radiomics signatures

A distinct advantage of radiomics modeling over black-box techniques such as deep learning is the potential interpretability of the obtained results. Therefore, we can attempt to reason the prominence of certain radiomics features in disease discrimination models. Shape features were highly featured in all models and indicated subtle patterns of ventricular remodeling that are specific to conditions under study. For instance, spherical disproportion (i.e., the inverse of sphericity) of the myocardium at end-diastole was lower in participants with high cholesterol

compared with healthy individuals, indicating that the overall shape of the LV is elliptical and more spherical in this risk factor group. Similarly, for hypertensive individuals and previous smokers, the surface area to volume ratio was smaller in the risk subgroups vs. healthy subjects; this may reflect a pattern of concentric LV hypertrophy in these conditions. For certain risk factors, intensity/texture features seemed more important, such as median intensity for diabetes. As this was a retrospective study, we can only speculate as to the cause of this association. One hypothesis is that diabetes leads to a global alteration of the myocardial tissue and thus of the overall myocardial appearance in CMR images, resulting in higher median intensities compared to non-diabetic subgroups. However, testing this hypothesis is beyond the scope of this study.

As another example of a prominent textural feature, the most important feature identified for current smokers in this study was gray level non-uniformity. In a previous study, [74], the very same radiomic feature was identified as the most important radiomic feature in Hypertrophic cardiomyopathy (HCM). However, as the authors pointed out in their paper, the intensity heterogeneity of myocardial tissue is not unique to HCM, and it might be of importance for other conditions. As smoking is a well-known cause for such cardiovascular diseases [166], there may be some commonality in the patterns of myocardial hypertrophy and tissue fibrosis in these cardiovascular conditions that is being reflected in the observed texture features. Indeed, the increased heterogeneity in grey level intensities for current smokers, as found in our study, supports the potential effects on the myocardium for these subjects.

Thus, radiomics allows more granular distinctions between health and disease in comparison to conventional CMR indices where, rather crudely, the single most discriminatory feature for all risk factors was higher LVM. These findings indicate the potential clinical utility of radiomics in improving understanding of the effects and pathophysiology of important cardiovascular risk factors.

3.4.7. Comparison with the existing literature

Literature in support of the superior diagnostic performance of CMR radiomics models over conventional image analysis is slowly gaining momentum. Several studies have shown the feasibility and clinical utility of CMR radiomics for distinguishing important disease entities. A small study by Baessler et al. [74] demonstrates the superior performance of CMR radiomics in discriminating hypertrophic cardiomyopathy (n=32) from healthy comparators (n=30). The most discriminative feature was grey level non-uniformity, a radiomics texture feature representing heterogeneity. It seems intuitive that this feature would be defining the irregular

myofibrillar architecture of hypertrophic cardiomyopathy. Similar to our observations, in particular with diabetes, it appears that the observed radiomics signatures may reflect clinically meaningful information about significant tissue-level changes.

Furthermore, studies have demonstrated the ability of CMR radiomics to distinguish important conditions that appear morphologically similar to conventional image analysis. For instance, Neisius et al. [141] shown high performance of CMR radiomics models applied to native T1 images to distinguish hypertensive heart disease (n=53), hypertrophic cardiomyopathy (n=108), and healthy volunteers (n=71). There is also emerging work on using CMR radiomics to identify areas of myocardial infarction from non-contrast cine images [73, 140, 167] and to identify acute from chronic myocardial infarction [167].

Our work constitutes the most comprehensive study to assess the relationship between CMR radiomics and cardiovascular risk factors. However, the concept of utilizing information from CMR to obtain more complex geometric information has been addressed previously using atlas-based shape measures. Cardiac atlases produce statistical shape models, giving highly detailed morphometric information [145, 168, 169]. Directly comparable to our findings, Gilbert et al. [170] demonstrate unique morphometric variations associated with individual risk factors (high blood pressure, smoking, high cholesterol, diabetes, angina), which could be quantified and visualized on constructed atlases. The derivation of radiomics shape features is methodologically different from cardiac atlases; however, there are conceptual similarities about the type of information they provide. Both seem to suggest that geometric features not captured by current image analysis approaches may be extracted from existing CMR images and that this information appears to provide additional insight into patterns of cardiac remodeling. CMR radiomics has several advantages over cardiac atlas models. The signal intensity-based radiomics features (first-order, texture) have great potential for not only better disease discrimination and outcome prediction but also gaining deeper insights into disease processes at the tissue level; such information is not provided by cardiac atlas morphometrics. CMR radiomics analysis does not require any dedicated acquisitions or post-processing and the extraction of radiomics features and model building are computationally simpler than atlas models. Therefore, there is real potential for radiomics to enter the clinical workflow as a very high yield and complementary image analysis tool.

Note that in this study, we chose to select a different healthy subsample than in Petersen et al. [159]. This is due to the differences in the objectives of the papers. While Petersen et al. [159] focused on the estimation of normal ranges of cardiac

indices of structure and function and thus used very strict inclusion criteria, we are concerned with the study of cardiovascular risk factors, and therefore we excluded subjects with known cardiovascular risk factor or disease.

3.4.8. Limitations and future work

To the best of our knowledge, this is the largest study to assess the performance of the CMR radiomics model in discriminating several important cardiovascular risk factors. Our findings demonstrate the feasibility of CMR radiomics models to identify cardiac changes related to important cardiovascular risk factors (diabetes, hypertension, high cholesterol, and smoking) with greater accuracy than conventional indices. The UK Biobank provides an excellent platform for this study with a large sample of well-characterized participants with linked CMR imaging. However, the data collection was conducted through a combination of a touchscreen questionnaire and a face-to-face nurse interview, and thus there remain some concerns about the accuracy and objectivity of the self-reported conditions. Studies with consideration of more sophisticated statistical methods to better account for confounding factors, as well as with the inclusion of external validation cohorts, are needed to produce and validate more disease-specific and generalizable models. In particular, there is a need for prospective studies to determine the clinical utility of these models in providing incremental cardiovascular risk information.

As for the pipeline implemented in this paper, alternative approaches may merit exploration, such as testing different methods for feature selection (e.g., LASSO [171], a combination of filter and wrapper-based methods [172]) or applying extensive hyperparameter optimization for each risk group. Also, while cross-validation was performed in the feature selection process to reduce the instability of radiomics features, other strategies have been proposed, such as prior clustering of redundant features [173], or using a concordance correlation coefficient [174]. Additionally, there is a need for proper evaluation of the reproducibility of radiomics features across segmentation protocols and also across imaging acquisitions, which is important due to non-standard pixel values and large variation in signal intensities [175]. Wider use of radiomics quality scores [176] would also enable better quality and more uniform reporting of radiomics studies and foster research reproducibility. Finally, as a common problem of artificial intelligence-based radiomics approaches, we have not assessed the practical value of the present results since there is no comparative gold standard that can be used for comparison.

3.5. Conclusions

CMR radiomics is an emerging technique for deeper and more accurate cardiac phenotyping in comparison to conventional image analysis. Our preliminary results based on a large sample from the UK Biobank indicate the feasibility of CMR radiomics analysis and potential clinical utility in superior image phenotyping of major cardiovascular risk factors, including diabetes, hypertension, high cholesterol, and smoking. The clinical value of these radiomics signatures for the prediction of downstream events warrants further investigation in prospective cohorts.

CHAPTER

4

Attribute-based, disentangled and interpretable representations of medical images with variational autoencoders

4.1. Introduction

Deep learning (DL) methods have recently shown great success in many fields, from computer vision [106, 177, 178] to natural language processing [179, 180], among numerous others. In addition, DL methods have started to dominate the medical imaging field [89], being used in a variety of medical imaging problems, such as segmentation of anatomical structures in the images [92, 181, 182], disease prediction [183], medical image reconstruction [184, 185] and clinical decision support [35]. Despite achieving exceptional results, DL methods face challenges when applied to medical data regarding explainability, interpretability, and reliability because of their underlying black-box nature [94, 95]. Hence, the need for tools that investigate the interpretability in DL is also emerging in healthcare.

Recent reviews of interpretable DL can be found in [94, 186–188]. Some methods have been proposed that employ backpropagation-based attention maps to either generate class activation maps that visualize the regions with high activations in

This chapter is adapted from:
Cetin I., Camara O., Gonzalez Ballester M. A., Attri-VAE: attribute-based, disentangled and interpretable representations of medical images with variational autoencoders. Medical Image Analysis. (2022) [Submitted]

specific units of the network [189] or saliency maps using gradients of the inputs with respect to the outputs [190, 191]. Other methods also proposed creating proxy models that focus on complexity reduction such as local interpretable model-agnostic explanation (LIME) [101] or by approximating a value based on game theory optimal Shapley values to explain the individual predictions of a model [192]. However, it is key to design models that are inherently interpretable, rather than creating post-hoc models to explain the black-box ones [193].

Recently, models based on latent representations, such as variational autoencoder (VAE), have become powerful tools in this direction [102, 103], as their latent space is able to encode important hidden variables of the input data [105]. Especially, when dealing with data that contains different interpretable features (*data attributes*), it is interesting to see how and if these attributes have been encoded in the latent space. Even though the proposed approaches provide promising results, they have some limitations, one of which is that the encoded variables cannot be easily controlled; they mostly show an entangled behavior, meaning each latent factor maps to more than one aspect in the generative process [194].

In order to bypass this limitation, much effort has been done to enforce disentanglement in the latent space [110, 195–198], being the majority of them unsupervised techniques [194, 199]. While many of these methods show good disentanglement performance, they are not only sensitive to inductive biases (e.g., choice of network, hyperparameters, or random seeds), but also some amount of supervision is necessary for learning effective disentanglement [199]. Moreover, since these methods are able to learn a factorized latent representation without attribute specification, they require a post-hoc analysis to determine how different attributes are encoded to different dimensions of the latent space [111].

On the other hand, attribute-based methods aim to establish a correspondence between data attributes of interest and the latent space [107–109, 111]. However, these methods also have their drawbacks: some of them are limited to work only on certain types of data attributes [108]; some impose additional constraints [109]; very few of them are designed to work with continuous variables [107, 111]; some require differentiable computation of the attributes; and they are extremely sensitive to the hyperparameters [107]. However, [111] have recently shown promising results for interpretability with their approach, associating each data attribute to a different regularized dimension of the latent space, which they have applied in the MNIST database for digit number recognition. The same approach was also employed as a post-processing step to generate interpretable and temporally consistent segmentations of echocardiography images [200].

In this work, we propose an attribute-interpreter VAE (Attri-VAE), an approach

based on attribute-based regularization [111] in the latent space, for an enhanced interpretation of clinical and imaging attributes obtained from multi-modal sources. Additionally, the proposed approach also enables classification, e.g., to identify healthy vs. pathological cases. Furthermore, we incorporate gradient-based attention map computation [102] to generate explanations of the attributes that are encoded in the regularized latent space dimensions. The main contributions of this work can be described as follows:

- The proposed approach is able to interpret different data attributes where specific ones are forced to be encoded along specific latent dimensions without the need for any post-hoc analysis, while encouraging attribute disentanglement by employing β-VAE as a backbone [195].
- The structured latent space enables controllable data generation by changing the latent code of the regularized dimension (i.e., following the corresponding attribute), generating new data samples as a result of manipulating these dimensions. For instance, if the attribute represents volume in a region of interest (ROI) and the corresponding regularized dimension is the first one of the latent code, then increasing values of the dimension would result in increasing the ROI volume.
- Attribute-based gradient-based attention maps provide a way to explain how the gradient information of individual attributes flow inside the proposed architecture.
- The classification network provides a way to stratify different cohorts, based on the attributes in the latent space. In this way, the most discriminative features for the classification task are identified by projecting original samples into the latent space.

In this work, we have applied the proposed Attri-VAE approach to study cardiovascular pathological conditions, such as myocardial infarction, using the EMIDEC[1] cardiac imaging dataset [201], including clinical and imaging features, also exploring the association with radiomics descriptors. Additionally, we used ACDC MICCAI17 database[2] as an external testing dataset.

The remainder of this chapter is organized as follows. Firstly, we define the state-of-the-art approaches in Section 4.2. We present the methodology and the

[1] http://emidec.com/

[2] https://acdc.creatis.insa-lyon.fr/description/databases.html

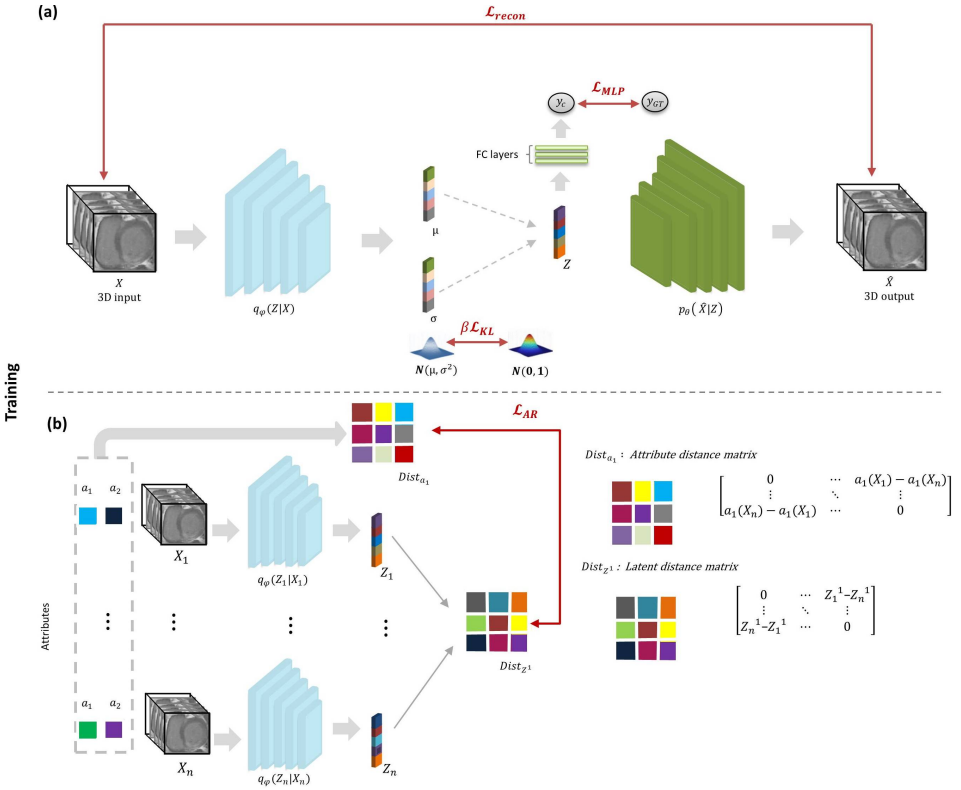

Figure 4.1: Training framework of the proposed approach. Loss functions are shown in red arrows. The total loss function of the model is: $\mathscr{L} = \mathscr{L}_{recon} + \beta\mathscr{L}_{KL} + \mathscr{L}_{MLP} + \gamma\mathscr{L}_{AR}$. (a) Losses computed for each data sample: multilayer perceptron (MLP) loss ($\mathscr{L}_{MLP}$), Kullback-Leibler (KL) loss ($\mathscr{L}_{KL}$), and reconstruction loss ($\mathscr{L}_{recon}$). (b) Attribute-regularization loss ($\mathscr{L}_{AR}$), computed inside a training batch that has n data samples. The input, a 3D image (X), first goes through the 3D convolutional encoder, $q_\varphi(Z|X)$, which learns to map X to the low dimensional space Z by outputting the mean (μ) and variance (σ) of the latent space distributions. The decoder, $p_\theta(\hat{X}|Z)$, then takes Z and outputs the reconstruction of the original input, ($\hat{X}$). The predicted classes of the inputs, y_c, are computed with a MLP module that consists of three fully connected (FC) layers. The corresponding MLP loss function is computed between y_c and the ground truth label y_{GT}. In (b), $\mathscr{L}_{AR}$ is shown to regularize the first dimension of the latent space (Z^1) with the attribute a_1 (a_1 and a_2 represent the first and the second attributes, respectively). $Dist_{Z^1}$ is the distance matrix of the first latent dimension, while $Dist_{a_1}$ represents the distance matrix of the attribute a_1.

details of our architecture in Section 4.3. We then describe the experimental setup and employed dataset in Section 4.4. Section 4.5 demonstrates our results. Section 4.6 discusses the obtained results and proposes future lines of work. Finally,

in Section 4.7 we conclude our findings. We have made our code publicly available in https://github.com/iremcetin/Attri-VAE.

4.2. Related work

4.2.1. Explainable AI in medical imaging

DL-based models have shown great promise in different machine learning tasks. Despite achieving remarkable performance in the medical domain, artificial intelligence (AI) based models still did not significantly deploy in the clinical routine. The main reason is the underlying black-box nature of DL-based networks and their high computational cost. In this line, explainable artificial intelligence (XAI) is an emerging field of research aimed at explaining how AI systems' black-box choices are made [98].

The proxy or shadow model approaches like local interpretable model-agnostic explanation (LIME) [101] and SHapley Additive exPlanations (SHAP) [202] are the simplest way to explain the model's decisions. LIME is a model-agnostic technique that aims to understand the model by perturbating the input of data samples and observing how the predictions change. SHAP is another perturbation-based technique that approximates so-called SHapley values by taking each input feature for a sample number of times. These techniques have been generally used in decision support systems [203]. Du et al. employed the SHAP technique to approximate the interpretability of the radiomics features in assessing patients with non-metastatic nasopharyngeal carcinoma (NPC). They analyzed 277 patients and extracted 525 radiomics features. The obtained SHAP values revealed that tumor shape sphericity, first-order mean absolute deviation, and overall stage were important factors in 3-year disease progression [204]. de Sousa et al. used the LIME approach to generate explanations on how a CNN detects tumor tissue in lymph node metastases [205]. However, these approaches have some drawbacks. SHAP is computationally expensive as the network must run samples $\times$ number of features times. On the other hand, the explanations provided by LIME can be unstable [94, 203].

Backpropagation-based attention generation techniques primarily determine an input feature's contribution to the target neuron, which is usually the output neuron of the correct class for a classification problem [94]. These techniques are also called as gradient-based techniques which can be applied to a black-box CNN without altering the underlying architecture. Some methods, such as DeepTaylor [206], only provide positive evidence and are only suitable for a limited number of tasks [94]. However, some of them, particularly gradcam variations (Grad-CAM

[189], Grad-CAM++ [207]), gained popularity in different medical imaging tasks. In a method proposed by Brinker et al., Grad-CAM was used to assess melanoma images. Their results suggest that proposed CNN-based model outperforms human experts [208]. Joshua et al. presented a study where they employed Grad-CAM++ to explain the decision of their proposed 3D CNN model for classifying lung cancer [209].

The latent representation-based models gained popularity to make deep learning models intrinsically explainable. The main advantage of these models is that their latent space can encode important hidden variables of the input data. Biffi et al. [103, 210] proposed two approaches for classifying heart pathologies (HCM) with cardiac remodeling. The explainable anatomical task-specific shape descriptors were learned directly from 3D segmentations using the latent space of VAE. Other approaches were also proposed to generate explanations in different clinical conditions. Clough et al. [211] introduced a VAE-based method to analyze its latent space to identify meaningful coronary artery disease detection biomarkers. Shakeri et al. [212] employed a VAE approach based on spectral feature representations using hippocampus morphology to classify Alzheimer's disease. Additionally, Puyol et al. [213] used existing clinical knowledge to constrain the latent space of a VAE by employing two task-specific classifiers together for the CRT response prediction of cardiomyopathy patients.

4.2.2. Attribute-based models

Attribute-based models establish a correspondence between data attributes of interest and the latent space by encoding different attributes along different latent space dimensions. This procedure is done in two ways; latent space can be decomposed into different parts representing specific attributes [214], or each data attribute can be encoded along individual dimensions [107, 215].

Recently, the InfoGAN [215] was proposed where it generates these aforementioned encodings maximizing the mutual information between specific dimensions of the latent vector and the generated data points. After training on images, it has been shown that InfoGAN can encode attributes such as rotating, lighting, etc. The significant limitation of this approach is that it is not possible to choose which attributes to encode. As an alternative to this Hadjeres et al. [107] proposed an approach, GLSR-VAE, which introduces a regularization loss to encode a selected attribute along specific latent space dimensions. However, the loss function requires the differentiable computation of the attributes and extensive hyperparameter tuning. Pati et al. [111] proposed AR-VAE where they introduced an attribute-based

regularization loss function to generate a structured latent space in which individual attributes are encoded along specific dimensions of the latent space.

Generative adversarial network (GAN)-based models are also used to encode different attributes, such as Donahue et al. [214] proposed a method to encode the facial identity of a person in the latent space of a GAN model where they decomposed the latent space into two parts: one encoded variation in facial identity and the other encoded variation due to all of the other attribute. Engel et al. trained a generator-discriminator GAN framework on the latent space of a trained VAE to enforce conditional generation. For this, they use conditioning input similar to the conditional GAN framework [216].

There are, however, some limitations of these approaches. Some methods design the work only on certain types of data [108]. Additionally, some techniques impose additional constraints, such as requiring the ability to generate data points by independently varying attributes [217], requiring differentiable computations of attributes [107], or the ability to group data points concerning specific attributes [214]. In addition, just a few of them are designed to work with continuous value attributes [107, 111, 218].

4.3. Methodology

The overall structure of our framework is shown in Figure 4.1 (training) and Figure 4.2 (testing). The proposed Attri-VAE incorporates attribute regularization into a β-VAE framework that was used as a backbone for the interpretation of data attributes. The trained network enables to generate new data samples by manipulating the data attributes, whereas the generated attribute-based attention maps explain how the gradient information of each attribute flows inside the proposed architecture. This section is organized firstly explaining the overall training criterion of the proposed model, with the following subsections describing each of the elements of our methodology and their integration.

4.3.1. Training criterion

Attri-VAE is trained with a loss function, $\mathscr{L}$, which is composed of four terms, as follows:

$$\mathscr{L} = \mathscr{L}_{recon} + \beta \mathscr{L}_{KL} + \mathscr{L}_{MLP} + \gamma \mathscr{L}_{AR}. \tag{4.1}$$

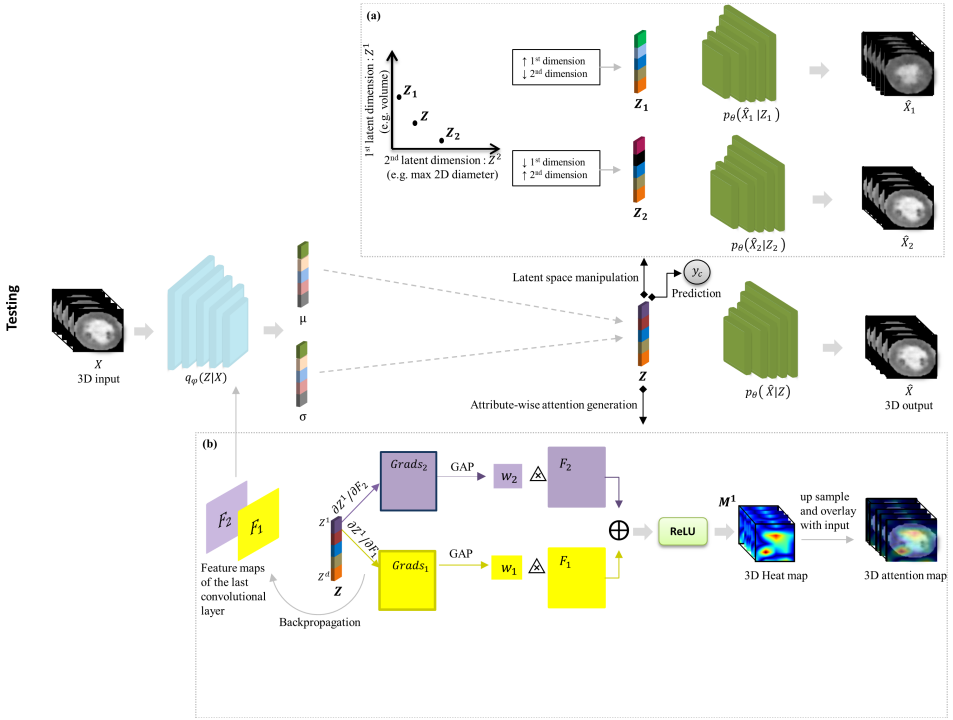

Figure 4.2: The trained network can be used for: (a) latent space manipulation; and (b) generating attribute-based attention maps. For a given 3D data sample, X, the trained 3D convolutional encoder, $q_\varphi(Z|X)$, outputs the mean (μ) and variance (σ) vectors, then Z being sampled with the reparameterization trick. (a) Data generation process by changing only first (Z^1) and second (Z^2) regularized latent dimensions of Z, which correspond to two different data attributes (volume and maximum 2D diameter, respectively). Then, the decoder, $p_\theta(X|Z)$, generates 3D outputs, X_1 and X_2, using the manipulated latent vectors, Z_1 and Z_2, respectively. (b) Attribute-based attention map generation for a given attribute, which is encoded in the first latent dimension (Z^1). First, (Z^1) is backpropagated to the encoder´s last convolutional layer to obtain the gradient maps ($Grads_1$ and $Grads_2$) with respect to the feature maps (F_1 and F_2). The gradient maps of (Z^1) measure the linear effect of each pixel in the corresponding feature map on the latent values. After that, we compute the weights (w_1 and w_2) using global average pooling (GAP) on each gradient map. A heat map is generated by multiplying these values (w_1, w_2) with the corresponding feature map, summing them up and applying an activation unit (ReLU). Finally, the heat map is upsampled and overlaid with the input image to obtain the superimposed image (3D attention map). Additionally, the class score of the input, y_c, is computed with the multilayer perceptron (MLP) that is connected to Z. Note that, in the figure it is assumed that the last convolutional layer of the encoder has 2 feature maps.

The reconstruction loss, $\mathscr{L}_{recon}$, is based on the binary cross-entropy (BCE) between the input X and its reconstruction $\hat{X}$, while the second term, $\mathscr{L}_{KL}$, employs

the Kullback-Leibler (KL) divergence between the learned prior and the posterior distributions, weighted by a hyperparameter (β). An additional term, $\mathscr{L}_{MLP}$, estimates the BCE loss for the classification between the network prediction, y_c, and the ground truth label, y_{GT}. The final loss term, $\mathscr{L}_{AR}$, includes the attribute regularization, with a tunable hyperparameter (γ) that weights its strength. In the following sections, detailed explanations of each loss term in our training criterion can be found (also see Figure 4.1).

4.3.2. Variational autoencoder (VAE) and β-VAE

A variational autoencoder [105] is a generative model that consists of an encoder and a decoder. The encoder, $q_\varphi(Z|X)$, approximates the posterior distribution with parameters φ, taking as input X from a high dimensional space, and learning to map it onto a low dimensional space by outputting the mean and variance (μ and σ, respectively) of a Gaussian probability density. The resulting low dimensional space is referred to as a latent space, with points Z in the latent space being the latent vectors. The decoder, $p_\theta(X|Z)$, parameterized by θ, takes a latent vector Z that is sampled from $p(Z)$ (prior distribution, e.g., unit Gaussian), using the reparameterization trick [105], and outputs $\hat{X}$, which is a reconstructed version of the input X.

A variational autoencoder aims to maximize the marginal likelihood of the reconstructed output, which is written as:

$$log_{p_\theta}(X) \geq \mathbb{E}_{Z\sim q_\varphi(Z|X)}[log p_\theta(X|Z)] - D_{KL}(q_\varphi(Z|X)\|p(Z)) \tag{4.2}$$

In this objective function, the first term is the log likelihood expectation that the input X can be generated by the sampled Z from the inferred distribution, $q_\varphi(Z|X)$. The second term corresponds to the KL divergence between the distribution of Z inferred from X, and the prior distribution of Z. Note that both distributions are assumed to follow a multivariate normal distribution.

In practice, the loss function of the VAE consists of two terms: a first term that penalizes the reconstruction error between the input and output; and a second term forcing the learned distribution, $q_\varphi(Z|X)$, to be as similar as possible to the prior distribution, $p(Z)$. In this case, the overall VAE loss can be written as:

$$\mathscr{L}_{VAE}(\theta,\varphi) = \mathscr{L}_{recon}(\theta,\varphi) + \mathscr{L}_{KL}(\theta,\varphi), \tag{4.3}$$

where the reconstruction loss, $\mathscr{L}_{recon}(\theta,\varphi)$, and the KL loss, $\mathscr{L}_{KL}(\theta,\varphi)$, are computed as follows:

$$\mathscr{L}_{recon}(\theta,\varphi) = \sum_{i=1}^{N} \|\hat{X} - X\|_2^2, \tag{4.4}$$

$$\mathscr{L}_{KL}(\theta,\varphi) = \mathscr{D}_{KL}(q_\varphi(Z|X)\|p(Z)). \tag{4.5}$$

When $q_\varphi(Z|X)$ is a multivariate normal distribution with parameters μ and σ^2, the objective loss function is differentiable with respect to $(\theta,\varphi,\sigma,\mu)$ [105], and the parameters of the VAE can be optimized iteratively with stochastic gradient descent algorithms [219].

A latent representation is disentangled if each dimension in the latent space is sensitive to one generative factor and comparably invariant to the changes in the other factors [220]. Such a disentangled representation is a great asset for interpretability. In this work we chose to use β-VAE as the backbone of our approach to encourage the disentanglement as it is easy to formulate and it has shown good performance based on one or more disentanglement metrics [195, 221].

The β-VAE approach [195] is an extension of the standard VAE that aims to learn a disentangled representation of the encoded variables in a completely unsupervised manner [195, 199] by simply giving more weight to the KL term, compared to the original VAE, with an extra hyperparameter β:

$$\mathscr{L}_{VAE}(\theta,\varphi) = \mathscr{L}_{recon}(\theta,\varphi) + \beta\mathscr{L}_{KL}(\theta,\varphi), \tag{4.6}$$

The main idea here is that adding β restrains the latent representation, forcing it to be more factorized [195, 221]: when $\beta > 1$, it encourages dimensional independence in the latent space, hence leading to a better disentanglement. On the other hand, when $\beta = 1$, it becomes equivalent to the standard VAE. Although, higher values of β have shown promising results to encourage disentangling [222], they often lead to a trade-off between reconstruction accuracy and the disentanglement of the latent space. For this reason, a well chosen β is necessary for both reconstruction accuracy and disentanglement.

4.3.3. Attribute-based regularization

In order to better interpret the data attributes that are encoded in the latent space, we employ an attribute-based regularization loss [111], which aims to encode an attribute a along a dimension d of the latent space (regularized dimension). In this way, as one interpolates along dimension d (in a D-dimensional latent space), the attribute value of the generated data is also monotonically changed. Therefore, our hypothesis is that a model trained with an attribute-based regularization not only improves interpretation but also can be used to generate controllable images by manipulating different dimensions of the latent space, which are corresponding to different data attributes.

In this sense, the attribute regularization loss, $\mathscr{L}_{AR}$, is calculated for the dimension d of the latent space in a training batch containing n training examples for the purpose of forcing the dimension d to have a monotonic relationship with the attribute values of a. The attribute regularization loss is then computed as follows:

$$\mathscr{L}_{AR}(d,a) = MAE(tanh(\delta Dist_{Z^d}) - sgn(Dist_a)), \tag{4.7}$$

where MAE is the mean absolute error, $Dist_a$ is the attribute distance matrix, and $Dist_{Z^d}$ is the distance matrix of the latent dimension d. These matrices are computed for all n data examples in the corresponding training batch, such that:

$$Dist_a = a(X_i) - a(X_j), \tag{4.8}$$

$$Dist_{Z^d} = Z_i^d - Z_j^d, \tag{4.9}$$

where $i, j \in [0,n)$, X_i and X_j are two exemplary samples (Equation 4.8), and each D-dimensional latent vector is represented as $Z = \{Z^d\}$, where $d \in [0,D)$ (Equation 4.9).

In Equation 4.7, $tanh$ and sgn refer to hyperbolic tangent function and sign function, respectively, whereas δ is the hyperparameter that modulates the spread of the posterior distribution. As we are interested in whether a certain sample´s attribute value is higher or lower than the others inside the corresponding mini-batch, the sgn function is used. Additionally, a $tanh$ function was chosen for the regularized dimension's distance matrix, $Dist_d$, because it has the same range as $sgn(Dist_a)$, and it is a differentiable function (i.e., the loss is also differentiable with respect to the latent vectors and the encoder's parameters). Consequently, the objective function tries to minimize the MAE between $tanh(\delta Dist_{Z^d})$ and $sgn(Dist_a)$ so that the regularized dimension has a monotonic relationship with the attribute values.

While the above procedure gives an objective function for one attribute, for multiple selected attributes of interest to be encoded in the latent space, the overall loss

function can be computed by summing all the corresponding objective functions together. Specifically, when the attribute set is $A : \{a_k\}$, where $k \in [0, K)$ contains K attributes ($K \leq D$, being D the latent size), then the overall loss function is computed as:

$$\mathscr{L}_{AR} = \sum_{k=0}^{K-1} \mathscr{L}_{d_k, a_k}, \tag{4.10}$$

where d_k represents the index of the regularized dimension for the attribute k. This process is represented in Figure 4.1 (b).

4.3.4. Classification network

Recently, performing a classification task using VAEs has been proposed to learn and separate different cohorts in the latent space. For example, Biffi et. al. [103] classified heart pathologies with cardiac remodelling using explainable task-specific shape descriptors learned directly with a VAE architecture from the input segmentations. Additionally, other approaches based on VAE have also been applied to analyse coronary artery diseases [211], Alzheimer's disease [212] or to predict the response of cardiomyopathy patients to cardiac resynchronization therapy [213].

In this line, to enforce class separation to the Attri-VAE, a multilayer perceptron (MLP) prediction network was connected to the latent vector, $p(y_c|Z)$ (see Figure 4.1). The corresponding objective function can be computed as the binary cross-entropy (BCE) between the network prediction y_c and the ground truth label y_{GT}, such that:

$$\mathscr{L}_{MLP} = BCE(y_c, y_{GT}) \tag{4.11}$$

4.3.5. Attribute-based attention generation

The Attri-VAE facilitates data interpretation by generating new data samples as a result of scanning the regularized latent dimensions. Furthermore, it also provides a way to obtain attention maps from these dimensions (attribute-based attention map generation) for a better understanding on how gradient information of these attributes flows inside the proposed architecture (as can be seen in Figure 4.2).

Attribute-based visual attention maps were generated by means of gradient-based computation (Grad-CAM) [189], as proposed by [102]. Basically, a score is calculated from the latent space that is then used to estimate the gradients and attention maps. Specifically, given the posterior distribution inferred by the trained network

for a data sample X, $q_\varphi(Z|X)$, the corresponding D-dimensional latent vector Z is sampled using the reparameterization trick [105]. Subsequently, for a given attribute set $A : \{a_k\}$, where $k \in [0, K)$ contains K attributes, attribute-based attention maps, M^{d_k}, are generated for each regularized latent dimension Z^{d_k} by backpropagating the gradients to the encoder´s last convolutional feature maps ($F : \{F_i\}$ where $i \in [0, n)$):

$$M^{d_k} = ReLU(\sum_{i=1}^{n} w_i F_i), \tag{4.12}$$

where d_k is index of the regularized latent dimension for a given attribute k. The weights, w_i, are computed using global average pooling (GAP), which allows us to obtain a scalar value, as follows:

$$w_i = GAP(\frac{\partial Z^{d_k}}{\partial F_i}) = \frac{1}{T} \sum_{p=1}^{j} \sum_{q=1}^{l} (\frac{\partial Z^{d_k}}{\partial F_i^{pq}}), \tag{4.13}$$

where $T = j \times l$, (i.e., $width \times height$), and F_i^{pq} is the pixel value at location (p, q) of the $j \times l$ matrix F_i. This process is visually summarized in Figure 4.2.

4.4. Application for interpretable cardiology

4.4.1. Datasets

Initially, the EMIDEC dataset [201] was used in our experiments. It is a publicly available database with of delay-enhancement magnetic resonance images (DE-MRI) of 150 cases (100 and 50 cases for training and testing, respectively), with the corresponding clinical information. Each case includes a DE-MRI acquisition of the left ventricle (LV), covering from base to apex. The training set, with ground-truth segmentations, includes 67 myocardial infarction (MINF) cases and 33 healthy subjects. The testing set includes 33 MINF and 17 healthy subjects. Some clinical parameters were also provided along with the MRI: sex, age, tobacco (yes, no, and former), overweight, arterial hypertension, diabetes, family history of coronary artery disease, electrocardiography (ECG), killip max[3], troponin[4], LV ejection fraction (EF), and NTproBNP[5]. Furthermore, we also used an

[3] A score based on physical examination and the development of the heart failure to predict the risk of mortality.

[4] A parameter that shows the level of the protein that is released into the blood stream.

[5] A parameter that shows a level of a peptide, which is an indicator for the diagnosis of heart failure.

additional external testing dataset for a more robust assessment of the classification performance, the ACDC MICCAI17 challenge training dataset [6] (end-diastole (ED), and end-systole (ES), cine-MRI from 20 healthy volunteers and 20 MINF cases). The ACDC dataset includes ground-truth segmentations of the left ventricle, myocardium and right ventricle by an experienced manual observer at both ED and ES timepoints [92]. The reader is referred to [92, 201] for more details on the MRI acquisition protocol.

As a pre-processing step, the intensities of the left ventricle in all images were scaled between 0 and 1. Additionally, each image was cropped and padded ($x = 80$; $y = 80$; $z = 80$; $t = 1$).

4.4.2. Cardiac attributes

Three different types of attributes were studied in our experiments. Initially, the Attri-VAE was trained with cardiac shape descriptors (e.g., wall thickness, LV and myocardial volumes, ejection fraction), extracted from ground-truth segmentations, which can easily be visually interpreted. In addition, attributes available from clinical information with the highest discriminative performance were identified using recursive feature elimination (RFE) with a support vector machines (SVM) classification model (linear kernel, regularization parameter $C = 10$) since this approach has already shown good performance for feature selection tasks [223–225]. The most discriminative attributes were then included in our analysis (e.g., gender, age, tobacco). The feature selection pipeline was done using the python-based machine learning library scikit-learn (version 1.0.2).[7]

Finally, the Attri-VAE was also trained with radiomics features. Radiomics analysis was originally proposed to capture alterations at both the morphological and tissue levels in oncology applications[40, 119], deriving multiple quantifiable features from pixel-level data. More recently, radiomics approaches have provided promising results on cardiac MRI data, for discriminating different cardiac conditions [73, 83, 141, 162], and to study cardiovascular risk factors in large databases [226]. Radiomics analysis represents a step towards interpretability compared to other black-box approaches since some features can be related to pathophysiological mechanisms [226]. However, there is a need for improving robustness and reproducibility of radiomics outcomes across different feature selection strategies and imaging protocols, which would lead to enhanced explainability. For this reason, radiomics features were employed in our experiments to benefit from the

[6] https://www.creatis.insa-lyon.fr/Challenge/acdc/
[7] https://scikit-learn.org/stable/

proposed network's ability to explain the encoded attributes. The open source library PyRadiomics (version 3.0.1) [8] was used to derive 114 features per analysed cardiac structure. Subsequently, radiomics features with the highest discriminative performance were identified using the above-mentioned feature selection approach as this strategy has also demonstrated good performance with previous radiomics studies [66, 227, 228]. The top performing features of this process were then selected to train the Attri-VAE.

4.4.3. Architectural details

The 3D convolutional encoder of the proposed Attri-VAE framework compresses the input into a 250 dimensional embedding through a series of 5 3D convolutional layers with kernel size 3 and stride 2, except the last convolutional layer that has stride 1. The prediction network was constructed with a shallow 3-layer MLP to be able to discriminate between the healthy and infarct subjects, using a ReLU activation function as a non-linearity after the first two layers. The upsampling and convolutional layers used in the encoder and the decoder were followed by batch normalization and ReLU non-linearity, except the decoder's last convolutional layer (Attri-VAE output) where a sigmoid function was applied. All the network weights were randomly initialized with xavier initialization [229]. The tunable parameters of the loss function (Equation 4.1) were fixed as follows: KL weight $\beta = 2$; and regularization weight $\gamma = 200$. Additionally, δ (Equation 4.7) was set to 10. The model architecture and other details are provided in our GitHub repository [9].

The Attri-VAE was trained on a NVIDIA Tesla T4 GPU using Adam optimizer with learning rate equals to 0.0001 and batch size of 16 for 10000 epochs. The dataset was splitted into 70/30 training (47 pathological, 23 healthy) and testing (20 pathological and 10 healthy subjects) sets. Subsequently, random oversampling of the normal subjects was employed in the training set as a strategy to treat the unbalanced behavior of the dataset; however, testing set was kept unchanged. Note that the proposed model is implemented using python programming language and PyTorch library (version 1.10.0) [10]. Image pre-processing and transformations were done using the python-based MONAI library (version 0.8.0) [11].

[8] https://pyradiomics.readthedocs.io/
[9] https://github.com/iremcetin/Attri-VAE
[10] https://pytorch.org/
[11] https://monai.io/

4.4.4. Experimental setting and evaluation criteria

The performance of the proposed Attri-VAE, both qualitatively and quantitatively, was compared with baseline VAE and β-VAE models in several experiments. First of all, the degree of disentanglement of the proposed latent space was evaluated with respect to different data attributes, using the following metrics available in the literature: the modularity metric, to analyse the dependence of each dimension of the latent space on only one attribute [230]; the mutual information gap (MIG), to evaluate the MI difference between a given attribute and the top two dimensions of the latent space that share maximum MI with the corresponding attribute [198]; the separated attribute predictability (SAP), to measure the difference in the prediction error of the two most predictive dimensions of the latent space for a given attribute [231]; and the spearman correlation coefficient (SCC) score, to compute its maximum value between an attribute and each dimension of the latent space.

In parallel, the interpretability metric introduced in [232] was used to measure the ability to predict a given attribute using only one dimension of the latent space. As for the β-VAE mode, dimensions having a high MI with the corresponding data attribute were chosen for the interpretability estimation. The reconstruction fidelity performance was also evaluated, employing the maximum mean discrepancy (MMD) score [233], which measures the distance between the distributions of real and reconstructed data examples, as well as their mutual information (MI) as an image similarity metric. The interpretability and MI metrics were then used to identify the optimal values of the most relevant hyperparameters in Equation 4.10 and Equation 4.7 (i.e., β, γ and δ), evaluating the influence of the KL divergence (β) and attribute regularization (γ) loss terms, as well as the weight of the distance matrix between two samples in a latent dimension. As a proof-of-concept, the hyperparameter sensitivity analysis was performed with only the four cardiac shape-based interpretable attributes.

Another set of experiments was carried out to explore the potential of the latent space generated by the Attri-VAE approach to create synthetically realistic samples. First, two samples in the Attri-VAE latent space, corresponding to input data with distinct cardiac characteristics (e.g., thin vs. thick myocardium, absence vs. presence of myocardium infarct), were chosen as references to synthetically generate interpolated images through their trajectory. Secondly, we qualitatively evaluated the control over individual data attributes during the generation process of the Attri-VAE model. Given a sample with a latent code z, a given attribute (e.g., LV volume) can be scanned from low to high values changing the latent code of the corresponding regularized dimension, due to their monotonic relationship. The attribute scanning creates synthetically generated samples in a latent space trajec-

tory where only the chosen attribute is changing, facilitating its interpretation. In order to further facilitate the identification of each attribute's visual influence in the synthetically generated images, gradient-based attention maps were also estimated.

Finally, the performance of the Attri-VAE model for classifying healthy and pathological hearts was assessed using the area under the curve (AUC) and accuracy (ACC) metrics, using both the EMIDEC and the ACDC17 challenge datasets. The Attri-VAE results were benchmarked against other VAE-type approaches (VAE+MLP, β-VAE+MLP), as well as to classical radiomics analysis (with SVM). The latent space projections of the Attri-VAE model, regularized by different attributes, were also qualitatively analysed to identify the attributes better differentiating healthy and pathological clusters of samples.

4.5. Results

4.5.1. Hyperparameter sensitivity analysis

Figure 4.3 shows the effect of several hyperparameters on the interpretability and the reconstruction fidelity of the Attri-VAE scheme. For comparison, the performance of β-VAE ($\beta = 3$) is also represented. A visual inspection of the figure suggests that γ, i.e., the hyperparameter controlling the attribute regularization, was the key to obtain good interpretability values while keeping reasonable reconstruction fidelity (mutual information ≥ 0.88), with values of $\gamma \geq 100$. Additionally, values of $\delta \leq 10$ (e.g., hyperparameter on the attribute regularization controlling the weight of the distance matrix between two samples) also ensured a good trade-off between interpretability and reconstruction fidelity. On the other hand, the β hyperparameter was not as relevant as the other two. As expected, the β-VAE approach without attribute regularization, provided acceptable reconstruction fidelity results but low values of interpretability. We need to point out that the same results were obtained when using radiomics features instead of shape-based attributes.

4.5.2. Disentanglement and interpretability

The proposed Attri-VAE approach outperformed β-VAE across all tested disentanglement metrics using shape and clinical attributes, implying a more disentangled latent space. Firstly, both Attri-VAE and β-VAE provided high modularity values (Attri-VAE: 0.98 vs. β-VAE: 0.97), signalling that each dimension of the latent spaces in both models only depended on one data attribute. The Attri-VAE also resulted in higher MIG/SAP scores than β-VAE (Attri-VAE: 0.60/0.63 vs.

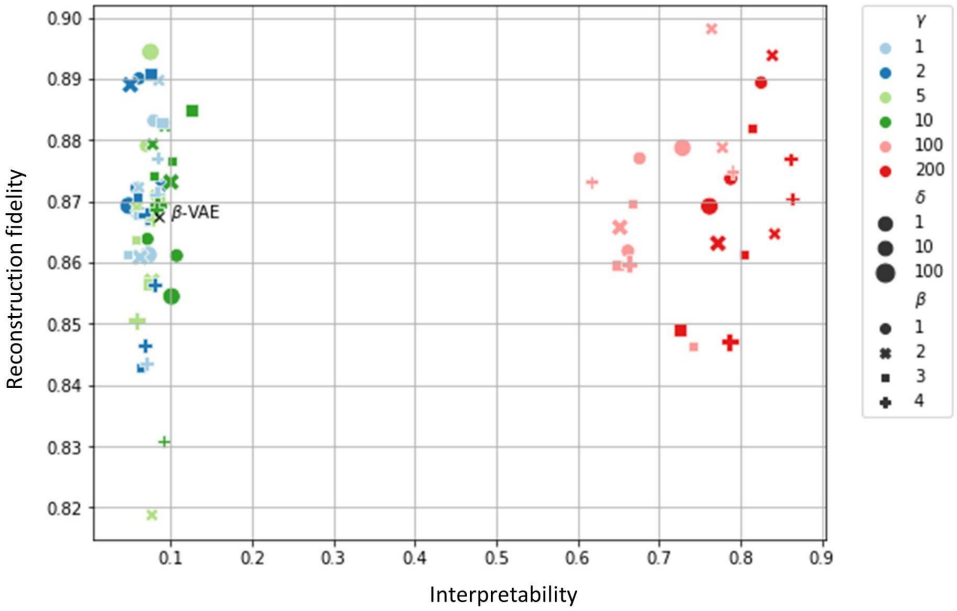

Figure 4.3: Effect of hyperparameters on the interpretability and reconstruction fidelity of the Attri-VAE approach. The hyperparameters β and γ of the Attri-VAE model control the influence of the loss terms for the Kullback-Leibler divergence between learned prior and posterior distributions, and attribute regularization, respectively. In its turn, δ weights the contribution of the distance matrix between two samples in a latent dimension in the attribute regularization scheme. Each marker represents a unique combination of the hyperparameters β, γ and δ, which is indicated by color, size and marker type, respectively. For comparison, the performance of β-VAE ($\beta = 3$) is also represented. Best performance combinations are located in the top right corner of the graph.

β-VAE: 0.02/0.05). In its turn, the SCC metric estimated for Attri-VAE was substantially higher than the corresponding β-VAE one (Attri-VAE: 0.97 vs. β-VAE: 0.46) due to the monotonic relationship between a given attribute and the regularized latent dimension enforced by the former. When using radiomics features, the same trend was observed, with some Attri-VAE disentanglement metrics (MIG and SAP) slightly lower than when using shape and clinical attributes (Attri-VAE / β-VAE): modularity, 0.98/0.98; MIG, 0.49/0.01; SAP, 0.51/0.06; SCC, 0.98/0.42).

Table 4.1 shows the interpretability scores for both Attri-VAE and β-VAE obtained with shape, clinical and radiomics attributes. The radiomics feature selection identified seven of them having the most discriminative power: four shape-based, being the sphericity of the left ventricle, the maximum 2D diameter of the myocardium, as well as left ventricle and myocardial volumes; three texture-based, being the correlation of the left ventricle, the difference entropy of the myocardium and the

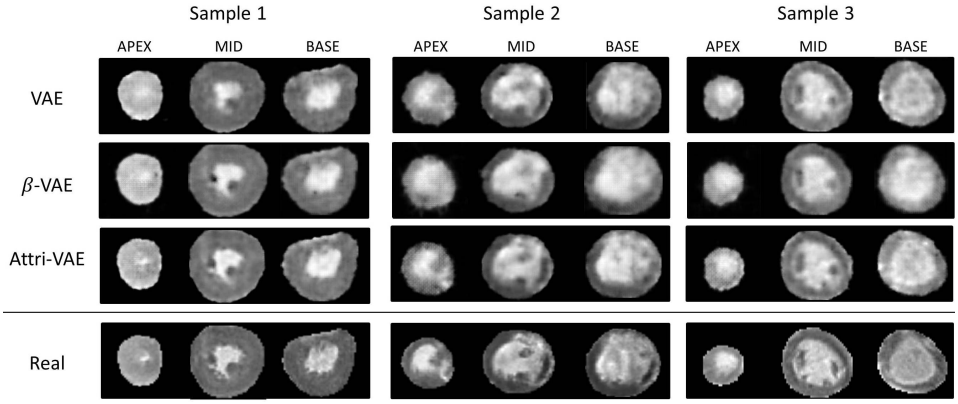

Figure 4.4: Three examples of real and reconstructed images using the VAE, β-VAE and Attri-VAE approaches. Three slices are shown in every example: apical (APEX), mid-ventricle (MID) and basal (BASE) slices. Sample 1 and 3 correspond to healthy hearts while Sample 2 shows an infarcted myocardium.

	Attri-VAE	β-VAE
LV volume	0.89	0.14
MYO volume	0.93	0.02
Wall thickness	0.95	0.10
EF	0.94	0.03
Gender	0.98	0.19
Age	0.93	0.12
Tobacco	0.70	0.19
Radiomics	0.91	0.06

Table 4.1: Interpretability score [232] of most relevant shape, clinical and radiomics attributes, as encoded in the latent space, with the Attri-VAE and β-VAE approaches. LV: left ventricle, MYO: myocardium, EF: ejection fraction. Maximum interpretability is 1.0.

inverse variance of the left ventricle.

It can easily be observed that the Attri-VAE provided a high degree of interpretability (i.e., close to 1.0) for all attributes, with the exception of tobacco (0.70). Among shape and clinical features, gender was the attribute with a higher interpretability (0.98), followed by the wall thickness (0.95), meaning that they could be predicted with only one dimension of the latent space. As for radiomics features, the average interpretability metric value was of 0.91, with shape-based ones showing slightly larger values than texture features (0.93 and 0.89, respectively); the maximum 2D diameter of the myocardium presented the highest value (0.97). On the other hand, the β-VAE clearly resulted in lower interpretability values (average of 0.11 for shape/clinical attributes and 0.06 for radiomics features).

4.5.3. Reconstruction fidelity

Table 4.2 summarizes the results of the reconstruction fidelity metrics (MMD and MI) for the VAE, β-VAE and Attri-VAE models. The proposed Attri-VAE approach obtained the lowest MMD values, representing a lower distance between input and reconstructed images. However, the VAE approach had the (slightly) best MI (0.91 and 0.89 for VAE and Attri-VAE, respectively), since its latent space was less constrained, compared to the other models.

Figure 4.4 shows the reconstructions of three data examples from the EMIDEC dataset using the VAE, β-VAE and Attri-VAE approaches. Even though the three models achieved similar qualitative reconstruction results, the Attri-VAE model generated images better preserving the heart shape and details than the other models: see the papillary muscles in mid-myocardium slices (dark regions in the blood pool) or the left ventricular cavity in apical slices of Sample 2 and Sample 3 in Figure 4.4. We can also observe in the figure that apical slides were more difficult to reconstruct that mid-ventricle and basal ones for the three tested models.

4.5.4. Latent space interpolation and attribute scanning

Figure 4.6 shows three examples of interpolation between two distinct and well-separated samples in the learned latent space of the Attri-VAE model. As it can be appreciated in the figure, the proposed approach generates synthetic interpolated images that have a realistic appearance, gradually changing the main sample characteristics in the trajectory between the chosen samples. The first row of Figure 4.6 clearly demonstrate the Attri-VAE model's ability to create smooth transitions between hearts having largely different characteristics such as (thin to thick) wall thickness. The other two rows of the figure demonstrate a similar behaviour from non-infarcted/scar to infarcted/scar patients.

Figure 4.5 illustrates the effect of scanning an individual attribute along its corresponding regularized dimension in the Attri-VAE model, where all the remaining attributes remain fixed. The first three rows of the figure exemplify the attribute

	MMD $\times 10^2$	MI
VAE	1.86 ± 0.06	0.91
β-VAE	1.38 ± 0.04	0.87
Attri-VAE	1.18 ± 0.03	0.89

Table 4.2: Reconstruction accuracy on the EMIDEC dataset of the VAE, β-VAE and Attri-VAE approaches, quantified with the maximum mean discrepancy (MMD) and mutual information (MI) metrics. The MMD results are given as $\pm$ standard deviation.

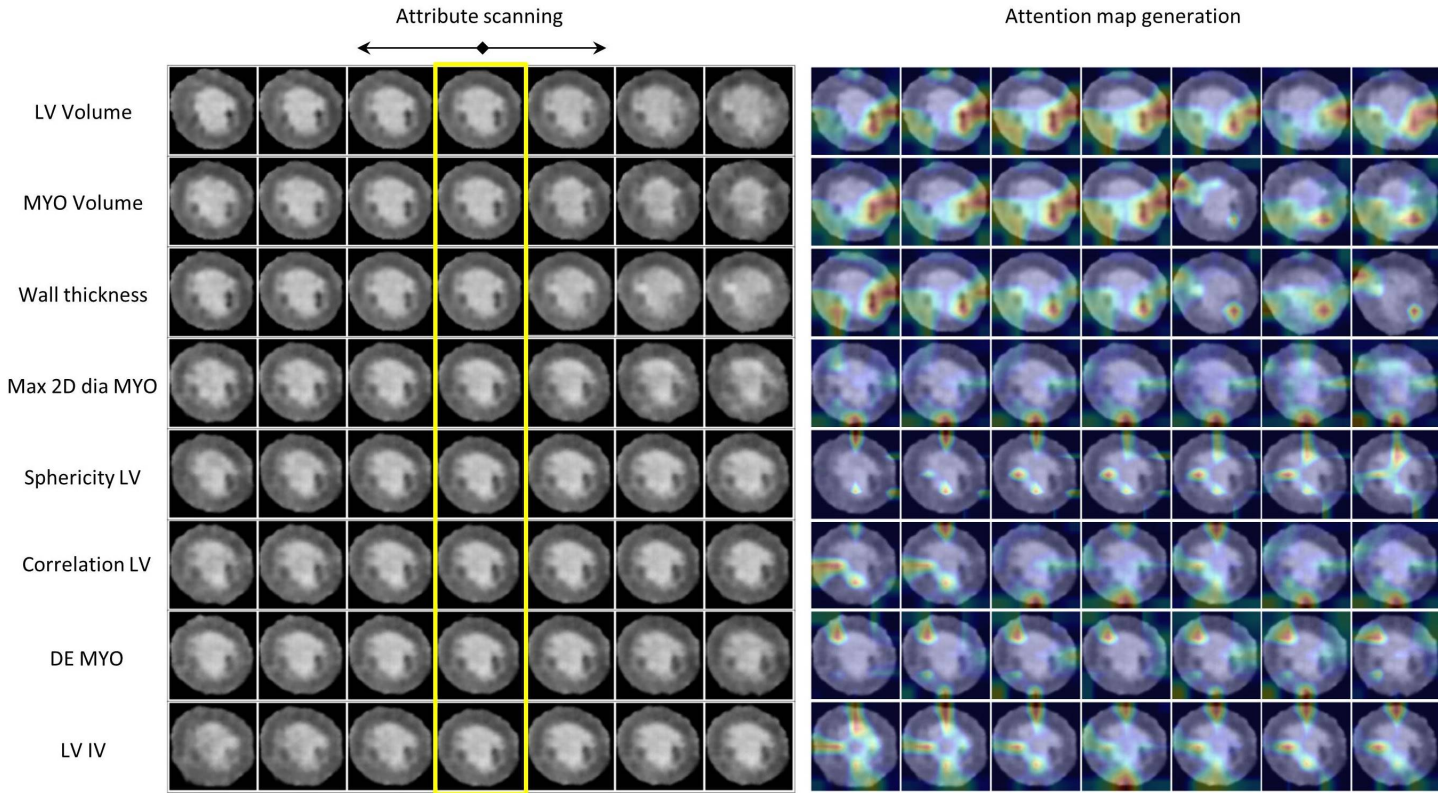

Figure 4.5: Scanning of attributes and corresponding gradient-based attention maps for shape and radiomics features. The image in the middle (4th column, in yellow frame) shows the original reconstructed image. DE: difference entropy, IV: inverse variance, Max 2D dia: maximum 2-dimensional diameter, LV: left-ventricle, MYO: myocardium. Note that the first three rows demonstrate the attribute scanning that was done on the latent space of Attri-VAE, which was trained with clinical and shape features. The remaining rows represent the attribute scanning on the latent space of Attri-VAE trained with selected radiomics features.

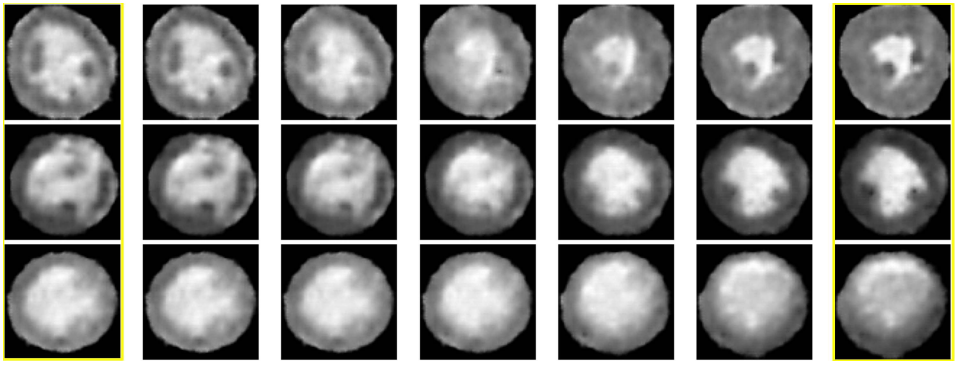

Figure 4.6: Linear latent space interpolation between two data samples (extremes of each row in yellow frames) from the EMIDEC dataset. Each row depicts the interpolation from the left to the right latent vector dimension. Top: from thin to thick myocardium. Middle: from a myocardium with scar to one without. Bottom: from healthy subject to a patient with a myocardial infarct.

scanning that was done on the latent space of Attri-VAE, which was trained with clinical plus shape features. The rest of the rows represent the attribute scanning on the latent space of Attri-VAE trained with selected radiomics features. For shape-based attributes, the changes in the attribute when moving along different values of the regularized dimension are clearly seen. For instance, from the left to the right in Figure 4.5, how LV and myocardial volumes are increasing in the first and two rows, respectively, or how the LV becomes more spherical. More subtle changes are observed with texture-based radiomics but they can still be identified with a careful inspection of the generated images. For example, moving along the latent space dimension corresponding to the correlation LV, we find more or less intensity homogeneity in the LV. The LV inverse variance (LV-IV) and the difference entropy of the myocardium (DE-MYO) only produced small changes that consisted in slightly thicker myocardium with lower values of LV-IV (left samples in Figure 4.5) and some more darker patches and heterogeneous texture in the myocardium for higher values of DE-MYO (right in Figure 4.5). It needs to be pointed out that attribute scanning for clinical attributes such as age, gender and tobacco is not shown since the images do not visually change along the corresponding regularized dimensions.

Additionally, the right side of Figure 4.5 shows the attention maps associated with the changes in each regularized dimension of the Attri-VAE model, as a way to better understand the effect of each studied attribute. We can see in the figure that more attention (i.e., higher response) is paid to more varying regions for shape-based attributes (e.g., right side of the slide for LV volume, where LV is increasing from the left to the right in the regularized dimension). In general, at-

tention maps for texture-based features have less high-response regions than for shape-based attributes. However, in some texture-based features such as the difference entropy of the myocardium, higher response can still be localized (in this example, darker regions in the top left part of the slice). On the other hand, interpretation and validation of the resulting attention maps for other attributes such as for LV-IV are more challenging.

4.5.5. Classification

	EMIDEC	ACDC
Attri-VAE (Clinical+Shape)	0.96 / 0.94	0.58 / 0.54
Attri-VAE (Radiomics)	0.98/ 0.96	0.59 / 0.52
β-VAE+MLP	0.91 / 0.90	0.45 / 0.31
VAE+MLP	0.87 / 0.80	0.54 / 0.35
Radiomics analysis (SVM)	0.77 / 0.75	0.60 / 0.61

Table 4.3: Classification performance of EMIDEC and ACDC datasets (healthy vs myocardial infarction) with different models. The results are reported as accuracy / AUC score. SVM: support vector machine.

Table 4.3 shows that the Attri-VAE approach, besides increasing interpretability, it also achieves a better classification performance comparing to state-of-the-art models. The best result was obtained in both EMIDEC and ACDC datasets with the Attri-VAE trained with radiomics features (accuracy of 0.98 and 0.59 for both datasets), while the standard radiomics+SVM analysis was the worst for EMIDEC (accuracy of 0.77) and the β-VAE+MLP for ACDC. There were only minor differences in the accuracy of the Attri-VAE method when trained with clinical and shape attributes or radiomics features. All the evaluated models, trained with the EMIDEC data, substantially dropped their performance when tested on the external ACDC dataset, specially the VAE-based approaches.

Finally, the latent space projections of different regularized latent dimensions are visualized in Figure 4.7, with plot axes representing the encoded data attributes. As it can be observed in the figure, our model is able to build several reduced dimensionality spaces, based on different attributes, where healthy and pathological cases (red and blue in the figure, respectively) can easily be clustered. For instance, the maximum 2D diameter of the myocardium and the LV volume attributes correctly separate most samples into two clusters. Interestingly, despite Attri-VAE having poor control over clinical attributes such as age or gender, they also facilitate the construction of the latent spaces and sample discrimination, as can be seen in the gender-age plot of Figure 4.7.

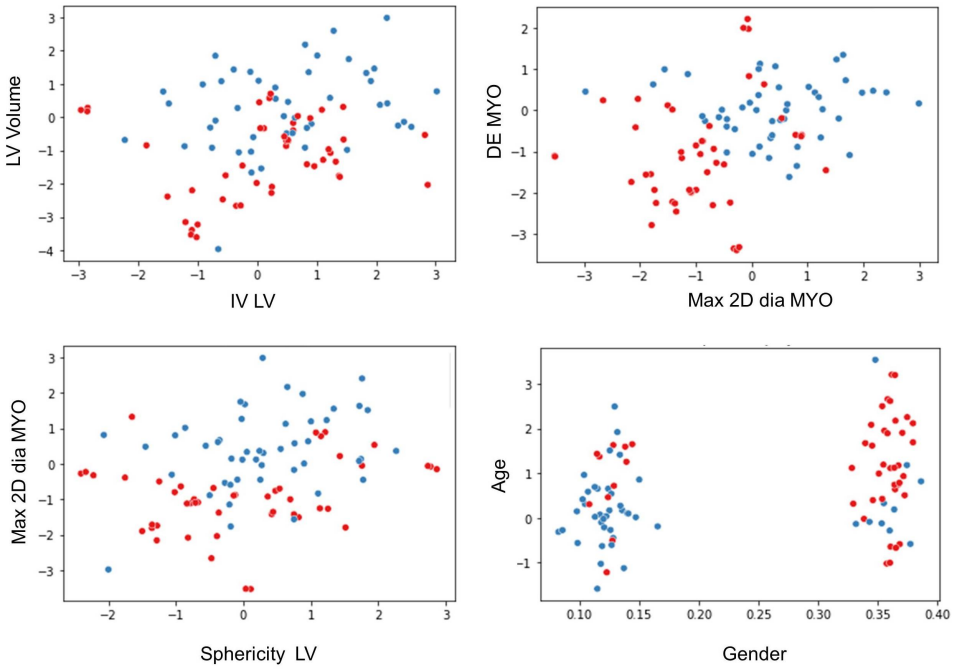

Figure 4.7: Latent space projections of regularized dimensions for different clinical, shape and radiomics attributes. Each point in the graphs represent a healthy or a myocardial infarction patient (red and blue, respectively), LV: left-ventricle, MYO: myocardium, IV: inverse variance, DE: difference entropy, Max 2D dia: maximum 2-dimensional diameter.

4.6. Discussion

The analysis of medical data demands for interpretable methods. However, the majority of deep learning methods do not fulfill the minimum level of interpretability to be used in reasoning medical decisions [35], being difficult to relate clinically and physiologically meaningful attributes with model parameters and outcomes. Fortunately, interpretable and explainable deep learning methods are starting to emerge. Models creating latent space representations, such as variational autoencoders, are promising but attributes are usually entangled in the resulting reduced dimensionality space, hampering its interpretation. In this work, we have presented the Attri-VAE approach that generates disentangled and interpretable representations where different types of attributes (e.g., clinical, shape, radiomics) are individually encoded into a given dimension of the resulting latent space.

The results obtained by the proposed Attri-VAE model based on disentanglement and interpretability metrics clearly outperformed the state-of-the-art β-VAE approach, indicating a high degree of disentanglement and a monotonic relationship

between a given attribute and the corresponding regularized dimension. However, Attri-VAE values for some metrics such as the MIG and SAP, although substantially better than those of β-VAE, were far from the maximum (e.g., 1.0). The same trend was observed by [111] in the MNIST (i.e., for digit number identification) dataset, suggesting that other latent dimensions, beyond the regularized ones, share a high MI with different attributes.

Hyperparameter selection was a key step to find the optimal Attri-VAE configuration providing an excellent trade-off between reconstruction fidelity, at the level of state-of-the-art alternatives, and interpretability; even though the Attri-VAE approach had a more constrained latent space, it generated reconstructions that are less smooth than other VAE models and more similar to the original input images. The most critical parameter to enforce interpretability was the weight of the attribute regularization loss term (γ in Equation 4.1), together with the influence of the distance matrix between two samples in a latent dimension (δ in Equation 4.7). The Attri-VAE plot of reconstruction fidelity vs interpretability, shown in Figure 4.3 had the same pattern as the one obtained by [111]. Interestingly, their optimal γ values were lower than ours ([5.0, 10.0] vs $\geq$ 100), likely due to the higher complexity of the cardiac MRI data and corresponding latent space compared to the MNIST dataset. On the other hand, the best δ values were the same in the two studies ([1.0, 10.0]).

One of the most interesting characteristics of the Attri-VAE approach is the ability of creating realistic synthetic data by sampling the created latent space and interpolating between different original reconstructed inputs, which can be very useful for controllable and attribute-based data augmentation of training datasets in machine learning applications. Scanning a regularized dimension of the latent space creates synthetic images where the corresponding attribute changes its values, as can easily be observed for shape descriptors (e.g., LV and myocardial volumes, wall thickness) in Figure 4.5. In addition, the proposed approach allows a better understanding of some (texture-based) radiomics features, which are often difficult to interpret. However, clinical attributes such as age, gender or tobacco consumption, despite obtaining good interpretability scores, did not create visually different interpolated samples over the regularized dimensions. One potential reason is the difficulty of the attribute regularization to control binary attributes, as suggested by [111]. Furthermore, the studied clinical attributes cannot be disassociated from shape and image intensity variations (e.g., morphological changes of the heart with age), thus it is too restrictive to keep all attributes fixed except a clinical one. In consequence, more work is needed to better construct latent spaces where clinical information can be disentangled from other attributes.

The generated gradient-based attention maps contributed to locally identify the cardiac regions where the attributes were influencing, which was particularly useful for global attributes and for complex features such as the texture ones. However, we only employed the well-known Grad-CAM method, which could be complemented with additional interpretability methods (e.g., LIME and its variations [101]) to better understand the attribute effects on the latent space. Additionally, the reliability of attention maps still requires further investigation to assess its robustness and reliability with respect to data input and model parameter perturbations [234]. In parallel, enhanced 3D visualizations of the generated samples are needed to have an overall perspective of the cardiac differences, beyond 2D slice views of the resulting images.

The proposed Attri-VAE model also achieved excellent classification performance (healthy vs. myocardial infarction), outperforming the other VAE-based approaches, with slightly better results when trained with radiomics. When evaluated in the EMIDEC training dataset with ground-truth labels, the Attri-VAE approach provided accuracy results (0.98) equivalent to the best challenge participants reporting their performance on the same dataset (1.0 [235], 0.95 [236], 0.94 [237] and 0.90 [238]). For the testing EMIDEC dataset [239], the best participant method obtained a decreased accuracy (0.82, [235, 240]), increasing to 0.92 for the challenge organizers [236]. As for the ACDC dataset, which was tested as an external database (i.e., without considering it in training), classification accuracy was substantially reduced (0.59), being worst than results reported by challenge participants [92] (0.96) to classify between the different pathologies (not only between healthy and myocardial infarction). Therefore, further work is required to improve the generalization of the Attri-VAE model to unseen data, being more robust to different quality and imaging acquisition protocols, through domain adaptation techniques, using databases such as the M&Ms challenge [241].

One limitation of the Attri-VAE approach, also acknowledged by Pati and Lerch [111], is the dependence on the selection of the data attributes to train the model. An incorrect attribute selection could lead to undesired strong correlations of several attributes that will not ensure a monotonic relationship with the corresponding regularized dimension, leading to less attribute interpretation and reconstruction quality. However, the projection of original samples in latent spaces with regularized dimensions for different attributes (see Figure 4.7) could be used as an interpretable attribute selection, identifying the ones better separating the analyzed classes such as the maximum 2D diameter of the myocardium and the LV volume attributes in our experiments. Further work will focus on fully integrating advanced feature selection techniques with the Attri-VAE model, as well as exploring alternative interpretability methods (see the recent review of Salahuddin

et al. [242]) to better understand the role of clinical and imaging attributes on medical decisions in cardiovascular applications.

4.7. Conclusion

We have presented a novel approach, referred to as Attri-VAE, which implements attribute-based regularization in a β-VAE scheme with a classification module for the purpose of attribute-specific interpretation, synthetic data generation and classification of cardiovascular images. The basis of the proposed Attri-VAE model is to structure its latent space for encoding individual data attributes to specific latent dimensions, being guided by an attribute regularization loss term. The resulting constrained latent space can be easily manipulated along its regularized dimensions for an enhanced interpretation of different attributes. Additionally, the proposed approach improves the current state-of-the-art for classifying cardiovascular images and allows the visualization of the most discriminative attributes by projecting the trained latent space. Future work will be focused on improving the generalization of the trained Attri-VAE models to images with different acquisition characteristics.

4.8. Availability of data and materials

This research was conducted using the publicly available EMIDEC and ACDC datasets. These datasets can be accessed in `http://emidec.com/dataset` and `https://www.creatis.insa-lyon.fr/Challenge/acdc/`. We have also made our code publicly available and can be found in `https://github.com/iremcetin/Attri-VAE`

CHAPTER

5

General discussion and conclusions

Cardiovascular magnetic resonance imaging (CMR) is the reference gold standard for analyzing cardiac structure and function and is used widely in research and clinical practice. The current understanding of cardiac alterations due to different clinical conditions has relied mainly on visual inspection of CMR images to identify global and local abnormalities; this is both labor-intensive and observer-dependent [123–125, 127]. Existing functional parameters such as ejection fraction (EF) and ventricle volumes are overly simplistic and mainly insensitive to subtle and complex modifications that affect the myocardium at the earliest disease stages [152]. Machine learning (ML) approaches have gained tremendous success in addressing these challenges in cardiovascular research tasks. It is mainly due to their excellent capability at feature selection and at learning highly-complex functions modeling a specific task under study [33–35, 88, 150]. Yet, ML approaches infrequently result in image features/biomarkers that a clinical expert can readily understand, and explaining why a model has made a specific prediction is often tricky. This poses the need for the development of novel methodologies that can successfully employ learning-based strategies while, at the same time, being explainable.

Therefore, this thesis has delved deeply into developing ML models to identify changes at both morphological and tissue levels in CMR, while also focusing on developing explainable learning strategies to identify the most relevant and interpretable biomarkers for clinical decisions. In chapter 2, a radiomics model was

developed and applied to different cardiovascular pathologies. Chapter 3 presents one of the largest CMR radiomics studies for image phenotyping of important cardiovascular risk factors. In chapter 4, a DL-based algorithm based on attribute regularization was developed, which is able to provide explanations of imaging biomarkers, and we also explored the association with radiomics descriptors.

In Chapter 2, the development of a machine learning-based radiomics approach, aiming at deeper imaging phenotyping of cardiovascular alterations and diagnosis of complex cardiovascular diseases was presented. The results suggest that radiomics are capable to encode alterations in the anatomy and tissues of the affected cardiac structures. The results from CVD reveal the importance of how shape-based and intensity-based features complement each other, and their combinations enhance the prediction power of the proposed model, particularly for the cases situated close to the boundary between two diseases. The results from hypertensive patients indicate the main alterations are in the cardiac textures and tissues, which explains the inability of conventional clinical images. Functional parameters of the heart focus on structural and functional quantification to identify these alterations. This work showed the great potential of using radiomics analysis in cardiac imaging and opened a way to apply this method to the study of other risk factors.

Chapter 3 presents the most extensive application of CMR radiomics analysis to discover new discriminatory signatures associated with important cardiovascular risk factors such as diabetes, hypertension, cholesterol, and smoking status, by using a large annotated CMR dataset from the UK Biobank. The results reveal that radiomics features lead to improved quantification of cardiac structures and tissue alterations due to the effects of underlying risk factors over conventional indices. Additionally, it has also been shown that the standard clinical indices do not capture the statistical differences between healthy and at-risk cohorts, with very few exceptions. In contrast, the proposed radiomics models outperformed the conventional indices according to the statistical tests employed during the study.

The explainability of different imaging biomarkers was studied in Chapter 4. An attribute-based regularization in the latent space of the proposed DL-based network was developed for an enhanced interpretation of clinical and imaging biomarkers obtained from multi-modal sources. Furthermore, a gradient-based attention calculation was also incorporated to explain the attributes encoded in the proposed model's latent space. The experimental results obtained from several quantitative and qualitative experiments suggest that the proposed approach generates disentangled and interpretable latent space representations, that can be used to develop explanations by manipulating interpretable data attributes while also

classifying various clinical conditions.

Limitations and future work

This work, however, faces several challenges which can open up avenues.

As for the pipelines implemented in chapter 2 and chapter 3, alternative approaches may merit exploration, such as testing the effects of different feature selection methods, for example, LASSO [171], a combination of filter and wrapper methods [172], or applying extensive hyperparameter sensitivity analysis for each cardiovascular cohort. Exploring the effects of different varieties of ML-based classifiers and combinations of different feature selection algorithms for each cardiovascular subgroup under the study is also needed, as Parmar et al. studied [139] this for the analysis of head-and-neck cancer patients. Additionally, as cross-validation was performed for feature selection in the proposed radiomics pipelines to identify the most relevant features (Chapter 2 and Chapter 3), other strategies can also be studied, including prior clustering of redundant features or using a concordance correlation coefficient [173, 174].

As the performance of the machine learning approaches is highly influenced by the selection of data to train the network, it is also vital to identify an accurate dataset with minimal missing values. In chapter 3, for example, data collection was conducted through a questionnaire and a face-to-face nurse interview. Thus, there are some concerns about the accuracy and objectivity of the self-reported conditions. For this reason, studies with more sophisticated statistical methods to better account for confounding factors and the inclusion of external validation cohorts are needed to produce and validate disease-specific generalizable models.

The approach proposed in Chapter 4 is able to create realistic synthetic data by sampling the designed latent space and interpolating between different reconstructed inputs. This characteristic can be extremely beneficial for attribute-based controllable data augmentation in ML-based medical imaging applications. Attribute scanning generated synthetic images changing the values of the corresponding attributes as can be observed in Chapter 4, for example, with shape descriptors (e.g., volumes of left ventricle and myocardium). However, some attributes such as age, gender, and smoking status did not create a visually noticeable difference in the images. One of the reasons for this, as suggested by [111], is that the formulation of attribute-regularization loss is not suitable for binary attributes. The other reason is that the studied binary attributes are clinical and primarily associated with variations in shape and image (e.g., aging results in morphological changes in the heart). For these reasons, our future work will include better understanding of the latent space so that clinical information can be disentangled from other at-

tributes. Moreover, further research will be needed to assess the potential of the approach by combining different attributes to visualize their joint effect.

Hyperparameter selection in Chapter 4 demonstrated the importance of finding optimal configurations for the proposed Attri-VAE. Although Attri-VAE has more constrained latent space, it provided an excellent trade-off between interpretability and reconstruction. The most critical parameter that made the significant change was the weight of attribute-regularization loss. The Attri-VAE plot of reconstruction fidelity vs. interpretability, as shown in Chapter 4 had the same pattern as the one obtained by [111]. However, their optimal values are slightly lower than ours. One potential reason is that the higher complexity of the employed CMR data results in a more complex latent space than the MNIST dataset. For this reason, future work will include studying the effects of different datasets on the optimal parameter selection.

The attention map generation identified the local cardiac regions most affected by changes in attributes. This approach was beneficial for global shape descriptors and some texture features. However, we only employed the well-known Grad-CAM approach [189]. For this reason, further work is needed in this direction by complementing attention map generation with additional interpretability methods, such as local interpretable model-agnostic explanation (LIME) and its variations [101] to better understand the impact of different attributes on the latent space. Additionally, the reliability of attention maps still requires further research to evaluate its robustness concerning other datasets and model parameter perturbations [234]. Additionally, the results in Chapter 4 presented in 2D, and yet, enhanced 3D visualizations are needed to have an overall idea of the cardiac differences. Further work will include 3D visualizations of the changes in different attribute.

The proposed Attri-VAE showed good classification performance where it outperformed the baseline models. It demonstrated similar performance to the best challenge participants on the same training dataset (1.0 [235], 0.95 [236], 0.94 [237] and 0.90 [238]). However, we observed worse results with the ACDC dataset, which was used as an external testing dataset (0.96 [92] vs. 0.59). Therefore, future work is needed to improve the generalization and the robustness of the proposed Attri-VAE on different datasets with different imaging protocols.

The selection of attributes plays a vital role on the performance of the network proposed in Chapter 4. Strongly correlated features result in a latent space that does not have a monotonic relation with the corresponding regularized dimension, leading to poor control and reconstruction quality. Future work should consider selecting appropriate attributes to encode into latent space. Finally, the prediction property of the proposed network indicates that the trained latent space could

constitute an interesting tool for feature selection by simply projecting different attributes as encoded in the latent space. However, the projection of data examples in latent space with regularized dimensions for different attributes could be used as an interpretable feature selection, identifying the ones better separating different clinical conditions. Our further work will continue to fully integrate advanced feature selection techniques using the Attri-VAE model, and other alternative interpretability approaches to assess the effect of different attributes (e.g., clinical and imaging attributes) on cardiovascular disease diagnosis.

CHAPTER

6 Appendix

6.1. Supplementary materials for chapter 3

6.1.1. Supplementary material 1

The purpose of creating this supplementary material is to demonstrate the potential and to explain the meaning of each radiomics feature used in the radiomics pipeline in Chapter 3. In the following sections you will find the description of radiomics features generated from the segmented region of interest (ROI) of the image. We extracted 684 radiomics features which encode two phases: end-diastolic and end-systolic information of left ventricle, right ventricle and myocardium using 114 unique radiomics plus demographic information (height, weight) and fractals (117 unique features total).

Shape Features	
Feature name	Interpretation
Volume	The volume of the ROI is approximated by multiplying the number of voxels in the ROI by the volume of a single voxel.
Surface Area	Surface Area is an approximation of the ROI surface based on triangulation interpretation.

Surface Area to Volume ratio	For details refer to preceding 2 features. Lower values of this parameter indicate a sphere-like shape of the ROI.
Sphericity	A measure of the roundness of the ROI relative to a sphere.
Compactness1	A measure of how compact the shape of the ROI is relative to a sphere.
Compactness2	A measure of how compact the shape of the ROI is relative to a sphere.
Spherical Disproportion	The inverse of Sphericity. Measures the ratio of the surface area of the ROI to the surface area of a sphere with the same volume as the ROI.
Maximum 3D diameter	The largest pairwise Euclidean distance between ROI surface voxels.
Maximum 2D diameter (Slice)	The largest pairwise Euclidean distance between ROI surface voxels of specific axial slice.
Maximum 2D diameter (Column)	The largest pairwise Euclidean distance between ROI surface voxels of specific coronal slice.
Maximum 2D diameter (Row)	The largest pairwise Euclidean distance between ROI surface voxels of specific sagittal slice.
Major Axis	A feature derived from the principal component analysis proportional to the square root of length of the largest principal component axes
Minor Axis	A feature derived from the principal component analysis proportional to the square root of length of the second largest principal component axes.
Least Axis	A feature derived from the principal component analysis proportional to the square root of length of the second largest principal component axes.
Elongation	A feature derived from the principal component analysis proportional to the ratio of lengths of the second largest and the largest principal component axes.
Flatness	A feature derived from the principal component analysis proportional to the ratio of lengths of the smallest and the largest principal component axes.

First Order Features	
Energy	Energy is a measure of the magnitude of voxel values in an image.
Total Energy	Total Energy is the value of Energy feature scaled by the volume of the voxel in cubic mm.
Entropy	Entropy specifies the uncertainty or randomness in the image values. It measures the average amount of information required to encode the image values.
Minimum	Minimum intensity value present in the ROI.
10th percentile	Value below which 10% of the intensities may be found in the histogram of the ROI.
90th percentile	Value below which 90% of the intensities may be found in the histogram of the ROI.
Maximum	Maximum grey level intensity found in the ROI.
Mean	Mean gray level intensity found in the ROI.
Median	Median grey level intensity found in the ROI.
Interquartile Range	The difference between the 25th and 75th percentile of ROI.
Range	A difference between the maximum and minimum gray tone present in the ROI.
Mean Absolute Deviation	MAD is the mean distance of all intensity values from the Mean Value present in the ROI.
Robust Mean Absolute Deviation	Robust MAD is a modification of MAD which takes into account only ROI intensities present in between 10th and 90th percentile which helps to avoid noise impact.
Root Mean Squared	RMS is the square-root of the mean of all the intensity values squared. Characterizes the magnitude of the image gray tone.
Standard Deviation	Measures the amount of variation from the mean intensity value.
Skewness	Skewness measures the asymmetry of the distribution of values around the Mean value

Kurtosis	Kurtosis measures the 'peakedness' of the values distribution in the image ROI.
Variance	Variance is the the mean of the squared distances of each intensity value from the Mean value.
Uniformity	Uniformity is a measure of the sum of the squares of each intensity value. This is a measure of the heterogeneity of the ROI.
Texture Features Gray level co-occurence matrix (GLCM)	
Autocorrelation	Autocorrelation detects repetitive patterns present in the ROI. Intends to measure the magnitude of the fineness and coarseness of texture .
Joint Average	Returns the mean gray level intensity of the i distribution.
Cluster Prominence	Cluster Prominence is a measure of the skewness and asymmetry of the GLCM.
Cluster Shade	Cluster Shade is a measure of the skewness and uniformity of the GLCM.
Cluster Tendency	Cluster Tendency is a measure of groupings of voxels within the ROI with similar gray-level values.
Contrast	Contrast is a measure of the local intensity variation, favoring values away from the diagonal of the GLCM.
Correlation	Correlation is a value between 0 (uncorrelated) and 1 (perfectly correlated) showing the linear dependency of gray level values to their respective voxels in the GLCM.
Difference Average	Difference Average measures the relationship between occurrences of pairs with similar intensity values and occurrences of pairs with differing intensity values in GLCM.
Difference Entropy	Difference Entropy is a measure of the randomness/variability in neighborhood intensity value differences.

Difference Variance	Difference Variance is a measure of heterogeneity that places higher weights on differing intensity level pairs that deviate more from the mean.
Dissimilarity	Mathematically equal to Difference Average
Joint Energy	Energy is a measure of how homogeneous are the patterns in the ROI.
Joint Entropy	Joint entropy is a measure of the randomness/variability in neighborhood intensity values.
Correlation1	Alternative definition of Correlation based on ratio of entropy dependencies to the maximum entropy.
Correlation2	Alternative definition of Correlation based on entropy dependencies. Uses square root of entropies difference instead of the max.
Inverse Difference Moment (IDM)	IDM is a measure of the local homogeneity of an image.
Inverse Difference Moment Normalized (IDMN)	Normalization of IDM. IDMN normalizes the square of the difference between neighboring intensity values by dividing over the square of the total number of discrete intensity values.
Inverse Difference (ID)	ID is another measure of the local homogeneity of an image.
Inverse Difference Normalized (IDN)	IDN normalizes the difference between the neighboring intensity values by dividing over the total number of discrete intensity values.
Inverse Variance	Inverse of the variance. Sums up the elements of the GLCM matrix while decreasing the values which lay further from the diagonal proportional to the distance.
Maximum Probability	Maximum Probability is the occurrence of the most predominant pair of neighboring intensity values.
Sum Average	Sum Average measures the relationship between occurrences of pairs with lower intensity values and occurrences of pairs with higher intensity values.

Sum Entropy	Sum Entropy is a sum of neighborhood intensity value differences.
Sum of Squares	Sum of Squares is a measure in the distribution of neighboring intensity level pairs about the mean intensity level in the GLCM.
Homogeneity1	An alternative measure of the local homogeneity of an image.
Homogeneity2	An alternative measure of the local homogeneity of an image.
Gray level size zone matrix (GLSZM)	
Small area emphasis (SAE)	SAE measures how many small regions with the same intensity value(fine texture) are present in the ROI opposed to big regions with same intensity value(homogeneous texture).
Large Area emphasis(LAE)	LAE measures how many big regions with same intensity value(homogeneous texture) are present in the ROI opposed to the small regions with the same intensity value(fine texture).
Gray Level Non-Uniformity (GLN)	GLN measures the variability of gray-level intensity values in the image, with a lower value indicating more homogeneity in intensity values and higher value indicating the presence of fine texture texture.
Gray Level Non-Uniformity Normalized (GLNN)	Normalized version of GLN which takes into account the number of zones with the same intensity present within the ROI.
Size-Zone Non-Uniformity (SZN)	SZN measures the variability of the size zone volumes(regions with the same intensity) in the image, with a lower value indicating that ROI has even size zones volumes.
Size-Zone Non-Uniformity Normalized (SZNN)	Normalized SZN which takes into account the number of zones with the same intensity present within the ROI.
Zone Percentage (ZP)	ZP measures the coarseness of the texture by taking the ratio of number of zones with the same intensity and number of voxels in the ROI.

Gray Level Variance (GLV)	GLV measures the variance in gray level intensities for the zones (regions with same intensity).
Zone Variance (ZV)	ZV measures the variance in zone(region with the same intensity) size .
Zone Entropy (ZE)	ZE measures the uncertainty/randomness in the distribution of zone sizes and gray levels.
Low Gray Level Zone Emphasis (LGLZE)	LGLZE measures the distribution of lower gray-level size zones, with a higher value indicating a greater proportion of lower gray-level values and size zones in the image.
High Gray Level Zone Emphasis (HGLZE)	HGLZE measures the distribution of the higher gray-level values, with a higher value indicating a greater proportion of both higher gray-level values and size zones in the image.
Small area low gray level emphasis (SALGLE)	SALGLE measures the proportion in the image of the joint distribution of smaller size zones with lower gray-level values.
Small area high gray level emphasis (SAHGLE)	SAHGLE measures the proportion in the image of the joint distribution of smaller size zones with higher gray-level values.
Large area low gray level emphasis (LALGLE)	LALGLE measures the proportion in the image of the joint distribution of larger size zones with lower gray-level values.
Large area high gray level emphasis (LAHGLE)	LAHGLE measures the proportion in the image of the joint distribution of larger size zones with higher gray-level values.
Gray level run length matrix (GLRLM)	
Short run emphasis (SRE)	SRE is a measure of the distribution of short run lengths, with a greater value indicative of shorter run lengths and more fine textural textures.
Long run emphasis (LRE)	LRE is a measure of the distribution of long run lengths, with a greater value indicative of longer run lengths and more coarse structural textures.

Gray level non-uniformity (GLN)	GLN measures the similarity of gray-level intensity values in the image, where a lower GLN value correlates with a greater similarity in intensity values.
Gray level non-uniformity normalized (GLNN)	GLNN measures the similarity of gray-level intensity values in the image, where a lower GLNN value correlates with a greater similarity in intensity values. This is the normalized version of the GLN formula.
Run length non-uniformity (RLN)	RLN measures the similarity of run lengths throughout the image, with a lower value indicating more homogeneity among run lengths in the image.
Run length non-uniformity normalized (RLNN)	RLNN measures the similarity of run lengths throughout the image, with a lower value indicating more homogeneity among run lengths in the image. This is the normalized version of the RLN formula.
Run percentage (RP)	RP measures the coarseness of the texture by taking the ratio of number of runs and number of voxels in the ROI.
Gray level variance (GLV)	GLV measures the variance in gray level intensity for the runs.
Run variance (RV)	RV is a measure of the variance in runs for the run lengths.
Run entropy (RE)	RE measures the uncertainty/randomness in the distribution of run lengths and gray levels. A higher value indicates more heterogeneity in the texture patterns.
Low gray level run emphasis (LGLRE)	LGLRE measures the distribution of low gray-level values, with a higher value indicating a greater concentration of low gray-level values in the image.
High gray level run emphasis (HGLRE)	HGLRE measures the distribution of the higher gray-level values, with a higher value indicating a greater concentration of high gray-level values in the image.

Short run low gray level emphasis (SRLGLE)	SRLGLE measures the joint distribution of shorter run lengths with lower gray-level values.
Short run high gray level emphasis (SRHGLE)	SRHGLE measures the joint distribution of shorter run lengths with higher gray-level values.
Long run low gray level emphasis (LRLGLE)	LRLGLE measures the joint distribution of long run lengths with higher gray-level values.
Long run high gray level emphasis (LRHGLE)	LRHGLE measures the proportion in the image of the joint distribution of larger size zones with higher gray-level values.
Neigbouring Gray Tone Difference Matrix (NGTDM)	
Coarseness	Coarseness is a measure of average difference between the center voxel and its neighbourhood and is an indication of the spatial rate of change. A higher value indicates a lower spatial change rate and a locally more uniform texture.
Contrast	Contrast is a measure of the spatial intensity change, but is also dependent on the overall gray level dynamic range. Contrast is high when both the dynamic range and the spatial change rate are high, i.e. an image with a large range of gray levels, with large changes between voxels and their neighborhood.
Busyness	A measure of the change from a pixel to its neighbor. A high value for busyness indicates a busy image, with rapid changes of intensity between pixels and its neighborhood.
Complexity	An image is considered complex when there are many primitive components in the image, i.e. the image is non-uniform and there are many rapid changes in gray level intensity.
Strength	Strength is a measure of the primitives in an image. Its value is high when the primitives are easily defined and visible, i.e. an image with slow change in intensity but more large coarse differences in gray level intensities.
Gray level dependence matrix (GLDM)	

Small dependence emphasis (SDE)	Measures how many small dependencies are present in ROI. Greater values represents smaller dependence and less homogeneous texture
Large dependence emphasis (LDE)	Measures how many large dependencies are present in ROI. Greater value indicates larger dependence and more homogeneous texture.
Gray level non-uniformity (GLN)	Measures the similarity of gray-level intensity values in the image. Higher value indicates smaller similarity whereas lower value indicates higher similarity in gray level intensity values.
Gray level non-uniformity normalized (GLNN)	GLNN measures the similarity of gray-level intensity values in the image, where a lower GLNN value correlates with a greater similarity in intensity values. This is the normalized version of the GLN formula.
Dependence non-uniformity (DN)	Measures the similarity of dependence throughout the image, with a lower value indicating more homogeneity among dependencies in the image.
Dependence non-uniformity normalized (DNN)	Measures the similarity of dependence in the image, with a lower value indicating more homogeneity among dependencies in the image. This is the normalized version of the DLN formula.
Gray level variance (GLV)	Measures the variance in grey level in the image.
Dependence variance (DV)	Measures the variance in gray level dependence size in the image.
Dependence entropy (DE)	DE measures the randomness in the gray level dependencies and gray levels.
Dependence percentage (DP)	DP is the ratio between voxels with a dependence zone and the total number of voxels in the image.
Low gray level emphasis (LGLE)	Measures the distribution of low gray-level values, with a higher value indicating a greater concentration of low gray-level values in the image.
High gray level emphasis (HGLE)	Measures the distribution of the higher gray-level values, with a higher value indicating a greater concentration of high gray-level values in the image.

Small dependence low gray level emphasis (SDLGLE)	Measures the joint distribution of small dependence with lower gray-level values.
Small dependence high gray level emphasis (SDHGLE)	Measures the joint distribution of small dependence with higher gray-level values.
Large Dependence Low Gray Level Emphasis (LDLGLE)	Measures the joint distribution of large dependence with lower gray-level values.
Large Dependence High Gray Level Emphasis (LDHGLE)	Measures the joint distribution of large dependence with higher gray-level values.

6.1.2. Supplementary material 2

This supplementary material explains the strategy to select the best model.

Radiomics feature selection and model building

After radiomics feature extraction, the next step consisted of identifying the combination of features that best discriminate the no-risk vs. at-risk subgroups for all risk factors. In this case, the selected radiomics features would encode alterations due to the risk factors under investigation. For this purpose, ML techniques (support vector machines, SVM; random forests, RF; logistic regression, LR) were implemented in combination with a feature selection algorithm.

Implementation of the SFFS and the ML techniques was based on the mlxtend (version 0.17.0) and scikit-learn (version 0.20.3) python-based libraries, respectively. An optimization process was performed by tuning the hyper-parameters of the ML techniques to find the optimal approach for the discrimination tasks. In total 33 combinations of ML methods and hyper-parameter values were tested:

- SVM (15 configurations): linear vs radial-basis function (RBF) kernel, gamma parameter of the RBF kernel (values of 0.1, 1 and 10) and regularization parameter (C, with values 0.1, 1 and 10);

Classifier	S	F	T	W	ACC	AUC
SVM	1	2	4	1	0.763	0.770
LR	5	3	3	0	0.782	0.803
RF	2	2	5	1	0.761	0.791

Table 6.2: Selected radiomics features and prediction performance for the optimal machine learning technique configurations. SVM: Support vector machines, LR: logistic regression, RF: random forests, S: shape, F: first-order, T: texture, W: size, ACC: accuracy, AUC: area under the curve

- LR (6 configurations): l1 (liblinear library [243]) vs l2 (lbfgs library [244]) penalty regularization and regularization parameter (C, with values 0.1, 1 and 10);
- RF (12 configurations): number of trees/estimators in the forest (nest with values of 10 and 100), maximum number of features in the best split (maxfeat = none, i.e. taking all features; maxfeat = sqrt, i.e. taking the square root of the number of features; maxfeat = $log2$, i.e. taking $log2$ of the number of features) and split quality criterion (gini impurity vs entropy). The selected radiomics features resulted from the SFFS algorithm and ML techniques were combined to create the radiomics signature that best encode the changes in CMR induced by the different cardiovascular risk factors.

Hyperparameter optimization on a subset of the data

To illustrate the process of hyperparameter optimization, we compared variants of the three studied ML techniques (SVM, RF, LR) on the subset of the data composed of diabetes vs. healthy controls (on 243x2 cases), generating a total of 33 different combinations of methods and hyper-parameter values.

The best discriminative performances for each ML technique were of 0.763 (SVM), 0.782 (LR) and 0.761 (RF), as can be seen in Table 6.2. These results were obtained with different amount (8, 11 and 10 features, respectively) and distribution of radiomics features. Notably, the best prediction performance in this data subset was provided by the LR technique, which selected 5 shape, 3 first-order and 3 texture based radiomic features.

Table 6.3 shows the results of two phases of the Cochran's Q statistical tests, aiming at first identifying the best hyper-parameter combinations within each ML technique separately and secondly comparing the different ML techniques among them. In a first step, statistically significant differences were found for the different combinations of the LR and RF techniques but the null hypothesis was accepted for SVM.

Subsequently, 9 classifiers were implemented with different ML techniques and hyperparameters for the next test; 2 SVM (C1, C2), 2 LR (C3, C4) and 5 RF (C5, C6, C7, C8, C9) classifiers. A statistical test was performed on all the classifiers and a p-value less than 0.5 was obtained, showing that there was a statistical difference among them. Afterwards a Bonferroni corrected post-hoc test was employed, with the new p-value equal to 0.0014, to perform pairwise comparisons. As a result of this test, statistical significant differences were found when comparing five different ML techniques (C1, C2, C3, C4 and C9), as illustrated in Table 6.3. After considering the overall prediction performances of these selected classifiers and the pairwise comparison results in Table 6.3, the optimal LR classifier, i.e. C4, was selected as the best method overall.

CLF	Cochran's Q test results			Post-hoc	
	Q	p-value	Result	BC	Selected classifiers
SVM	11.97	p=0.6	H0 accepted	-	Best AUC: C1: SVM (RBF, gamma = 0.1, C = 10) Best ACC : C2: SVM (linear, C = 1)
LR	19.37	p<0.05	H0 rejected	0.03	C3: LR (l1, C = 0.1) Best AUC and ACC: C4: LR (l1, C = 10)
RF	45.09	p<0.05	H0 rejected	0.0008	C5: RF (nest = 100, maxfeat = sqrt, gini) (best AUC and ACC) C6: RF (nest = 100, maxfeat = log2, gini) C7: RF (nest = 100, maxfeat= none, gini) C8: RF (nest = 100, maxfeat = none, entropy) C9: RF (nest = 10, maxfeat = none, entropy)
Second test					
	38.32	p<0.05	H0 rejected	0.0015	C1, C2, C3,C4 and C9
Identified pairwise comparisons					
1. C1 vs C9: C1 is better (with 39:85 ratio) 2. C2 vs C9: C2 is better (with 46:93 ratio) 3. C3 vs C4: C4 is better (with 29:65 ratio) 4. C4 vs C9: C4 is better (with 44:100 ratio)					

Table 6.3: Results of the Cochran's Q test and Bonferroni corrected McNemar post-hoc analysis. The results of the pair-wise tests show the misclassified ratios of the respective machine learning techniques. SVM: support vector machines, LR: logistic regression, RF: random forest, C (in SVM): regularization parameter, RBF: radial basis functions kernel, nest: number of estimators in RF, maxfeat: maximum number of features, AUC: area under the curve, ACC: accuracy, CLF: Classifier, BC: Bonferroni corrected p-value.

6.2. Supplementary material for Chapter 4

This supplementary material provides the architectural details of the proposed Attri-VAE (explained in Chapter 4).

The input is compressed into a 250-dimensional embedding through a 5-layer convolutional encoder. The prediction network (MLP linear stack) is constructed with a shallow 3-layer linear stack. The 3D convolutional decoder consists of 3 convolutional and 4 transposed convolutional layers. The proposed architecture can be seen in Figure 6.1. The details of configurations of the network are provided in Table 6.4.

Model	**Architectural Details**
Encoder (5-layer convolutional network)	Conv3D(input = 1, output = 8, ks = 3, s = 2, pad = 1) + BN + ReLU
	Conv3D(input = 8, output = 16, ks = 3, s = 2, pad = 1) + BN + ReLU
	Conv3D(input = 16, output = 32, ks = 3, s = 2, pad = 1) + BN + ReLU
	Conv3D(input = 32, output = 64, ks = 3, s = 1, pad = 1) + BN + ReLU
	Conv3D(input = 64, output = 2, ks = 3, s = 2, pad = 1) + BN + ReLU
Encoder linear stack	fc1: Linear(input = 250, output = 128) + ReLU + Dropout(d = 0.25)
	fc2: Linear(input = 128, output = 96) + ReLU
	fc3: Linear(input = 96, output = 64) (x2 in parallel)
Bottleneck	Z (latent dimension = 64)
MLP linear stack	fc5: Linear(input = 64, output = 32) + BN + ReLU
	fc6: Linear(input = 32, output = 16) + BN + ReLU
	fc7: Linear(input = 16, output = 1)
Decoder linear stack	fc8: Linear(input = 64, output = 250) + ReLU
Decoder (7-layer convolutional network)	Conv3D(input = 2, output = 64, ks = 3, s = 1, pad = 1) + BN + ReLU
	Trconv3D(input = 64, output = 32, ks = 3, s = 2, pad = 1) + BN + ReLU
	Trconv3D(input = 32, output = 16, ks = 3, s = 2, pad = 1) + BN + ReLU
	Trconv3D(input = 16, output = 8, ks = 3, s = 2, pad = 1) + BN + ReLU
	Trconv3D(input = 8, output = 4, ks = 3, s = 1, pad = 1) + BN + ReLU
	Conv3D(input = 4, output = 2, ks = 3, s = 1, pad = 1) + BN + ReLU
	Conv3D(input = 2, output = 1, ks = 3, s = 1, pad = 1)

Table 6.4: Configurations of the proposed approach as visualized in Figure 6.1. Conv3D: 3-dimensional convolutional layer, Trconv3D: 3-dimensional transposed convolutional layer, input: input channels, output: output channels, ks: kernel size, s: stride, pad: padding, BN: batch normalization, d: dropout probability, ReLU: Rectified linear unit.

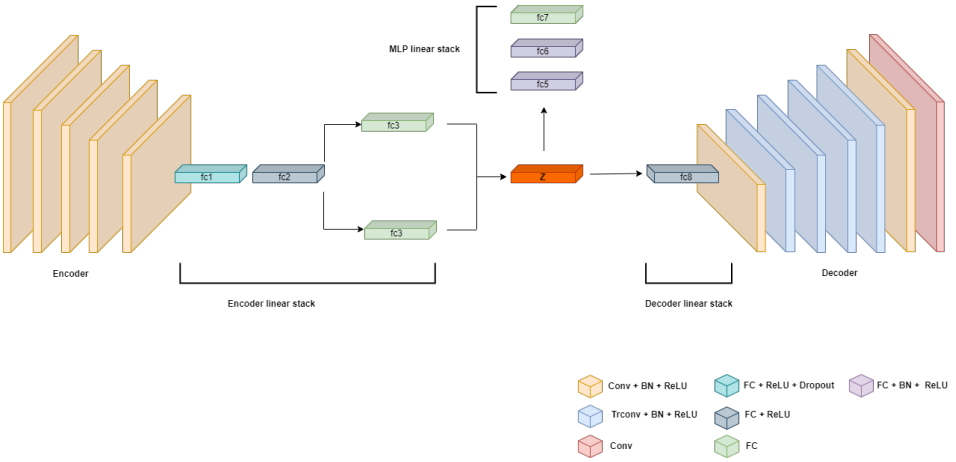

Figure 6.1: Architectural details of the proposed Attri-VAE. Conv: convolutional layer, Trconv : transposed convolutional layer, BN: batch normalization, fc: fully connected layer, ReLU: Rectified linear unit. Details of the configurations were provided in Table 6.4.

6.3. Additional radiomics experiments

This section explains additional experiments that were conducted during the thesis work. Section 6.3.1 briefly demonstrates the radiomics analysis to study in-depth the changes due to atrial fibrillation. Section 6.3.2, on the other hand, shows the results of a radiomics analysis in abdominal aortic aneurysm.

6.3.1. Identifying alterations in the cardiac ventricles in atrial fibrillation: a radiomics approach

Introduction

Atrial fibrillation (AF) is the most common cardiac arrhythmia, increasing the risk of stroke, heart failure, and other cardiovascular diseases. Furthermore, remodeling of the cardiac ventricles may occur due to AF. However, the exact nature of these changes remains unclear. Conventional imaging studies using imaging indices of cardiac structure and function might not be able to identify the subtle and complex changes that occur in the ventricular muscles due to AF.

This section is adapted from: **Cetin I.**, Petersen S.E., Camara, O., Gonzalez Ballester M.A., Lekadir K., Identifying alterations in the cardiac ventricles in atrial fibrillation: a radiomics approach **CARS**, 82-90 (2019).

In this work, we propose a radiomics approach to study in-depth the changes that occur in AF, integrating a comprehensive set of size, shape, intensity and texture radiomic descriptors in the analysis. The method combines then feature selection and machine learning to discriminate AF subgroups as compared to healthy individuals.

The obtained results demonstrate that the proposed radiomics model is capable of detecting intensity and textural changes well beyond the capabilities of conventional imaging phenotypes, indicating its potential for improved understanding of the longitudinal effects of atrial fibrillation on cardiovascular health and disease.

Methods

In this work cardiovascular magnetic resonance imaging (CMR) images from the first 5065 UK Biobank participants were assessed. The CMR parameters are: scanner = 1.5 Tesla Siemens, in-plane resolution = 1.8 x 1.8, slice thickness = 8.0 , slice gap = 2. Manual annotation of the images was undertaken by our clinical collaborator, resulting in a segmentation of left ventricle (LV), right ventricle (RV) and myocardium (MYO).

To develop and validate the proposed method, 60 AF patients were identified from UK Biobank. Furthermore, to develop a multi-classifier approach with 10 different classifiers, 600 normal subjects were added to our sample. Each classifier (AF vs. healthy) is built by analyzing a large pool of diverse radiomics features such that relevant ventricular alterations due to AF can be captured. Concretely, we calculated a total of 686 radiomic features using pyRadiomics library [62], describing a range of shape, intensity and textural characteristics of the cardiac substructures and tissues. A radiomic feature selection is thus necessary to select only those features that specifically deviate from normality in the presence of AF.

For each classifier, we implemented a sequential forward feature selection (SFFS) to identify the most relevant radiomic features as those that best discriminate AF and healthy hearts in a classification setting. Support vector machines (SVM) was chosen as the underlying classification model. The ventricular radiomic features that have similar or overlapping distributions between AF and healthy classes are ignored, while those that contribute to the SVM classification of AF and normal hearts are included within the final set of optimal radiomic features. This process was employed 10 times to obtain 10 different classifiers from 45 AF cases and 450 normal cases (by varying randomly the reference set of normal cases). The remaining 15 AF patients and 150 normal cases were used for testing. Note that for each test, majority voting was used to fuse the individual predictions from each of the 10 base classifiers.

Results

Table 6.5 lists the list of most common radiomic features within the 10 trained classifiers. The proposed approach achieved a classification accuracy of 0.83, indicating the relevance of radiomics features for describing AF-specific changes in the ventricles. In Table 6.5, the frequently identified features are shape-, intensity- and texture-based radiomics features,

Name	Frame	Structure	Type
Maximum	ES	LV	Intensity
Maximum	ED	RV	Intensity
Large area emphasis	ES	RV	Texture
Kurtosis	ED	RV	Intensity
Surface area to volume ratio	ED	MYO	Shape
Strength	ED	RV	Texture
Low gray level zone emphasis	ED	MYO	Texture
Least axis	ED	LV	Shape
Energy	ES	LV	Intensity
Least axis	ED	LV	Shape

Table 6.5: List of the most frequently identified radiomics features in 10 classifiers, for healthy/AF classification. ED: End-diastole. ES: End-systole.

suggesting that AF induced changes concern shape and also the tissues of the ventricles. Note that the selected radiomics cover all three ventricular substructures (LV, MYO and RV). This indicates that the changes are multi-form and that these radiomics features form a multi-variate signature to describe AF related remodeling in the ventricles.

Finally, to compare the obtained results, conventional imaging indices of cardiovascular structure and function (e.g ejection fraction (EF), stroke volume, LV and RV volumes, volume of left atrium (LA)) were tested as an alternative to the radiomics using the same machine learning approach. The most common indices are: stroke volume of RV and LV, maximum volume of LA in 2- and 4-chamber views, LV ejection fraction, volume of LV at ED and ES. They result in a classification accuracy of 0.73 (versus 0.83 for the radiomics model), which suggests that the radiomics carry additional information on the tissue changes that take place in AF individuals.

Conclusion

This work shows the promise of cardiac radiomics for analyzing changes in the ventricles caused by AF. Future work includes the clinical interpretation and applicability of the results.

6.3.2. 3D radiomics analysis to predict patient evolution after endovascular aneurysm repair

An abdominal aortic aneurysm (AAA) is a dilation of the aorta which, if not treated, tends to grow and rupture with a high risk of mortality [245]. In the last decade, the treatment of

AAA has shifted from open surgery to a minimally invasive technique known as Endovascular Aneurysm Repair (EVAR). In EVAR, a stent is deployed inside the aorta to isolate the damaged aneurysm wall from the blood flow. After a successful EVAR intervention, the excluded aneurysm is thrombosed and reabsorbed. However, in a large percentage of patients the aneurysm continues growing after the intervention due to EVAR-specific complications known as endoleaks [246]. These endoleaks refer to a persistent blood flow entering into the excluded aneurysm, which increases the risk of rupture and may lead to a re-intervention.

There are different types of endoleaks depending on its source, among which the following are considered in the current work:

- Type I: there is a gap between the stent and the aneurysm wall at the sealing zones, which requires urgent attention due to a high risk of aneurysm rupture.
- Type II: there is a retrograde flow inside the aneurysm coming from side branches; it is the most common type, considered to have a good prognosis, but are sometimes unpredictable.
- Type V-Endotension: it is the least understood endoleak, referring to the case where the aneurysm grows but there is no visible leak in the image.

Thus, patients treated with EVAR undergo lifelong surveillance, based on yearly Computed Tomography Angiography (CTA) scans, to detect possible complications and evaluate the risk of aneurysm rupture. Currently, the clinical procedure to evaluate the progress and prognosis of an AAA postoperatively consists in a qualitative evaluation of the CTA images to detect endoleaks and the measurement of the 2D maximum diameter in manually selected CTA slices. Hereby, our goal is to study the feasibility of using radiomics to assist the clinician during this evaluation.

Radiomics methods refer to the estimation and mining of a large number of advanced imaging features that describe shape, size, intensity and textural properties of anatomical structures and tissues. While they have gained great popularity in oncology to predict tumor progression and treatment response [247, 248], their use in other clinical domains such as cardiology is only recent [162, 249].

In this study, we combine 3D radiomics and machine learning for analyzing the first two postoperative CTA volumes of 12 patients with the aim to predict their evolution during follow-up. In contrast to previous works analyzing postoperative data [250, 251], our method performs the analysis in 3D, combining shape, texture and histogram features and taking into consideration all the information within the thrombus without the need to select certain slices that could influence the outcome of the algorithm. Furthermore, analyzing always the first two postoperative scans increases the complexity since in some cases the leak is still not visible in the image. Our hypothesis is that a multi-scale quantification of complex, as well as subtle changes in morphology, function or appearance within the AAA may provide new clinically-useful information for predicting unfavorable evolution after EVAR and identify at-risk individuals.

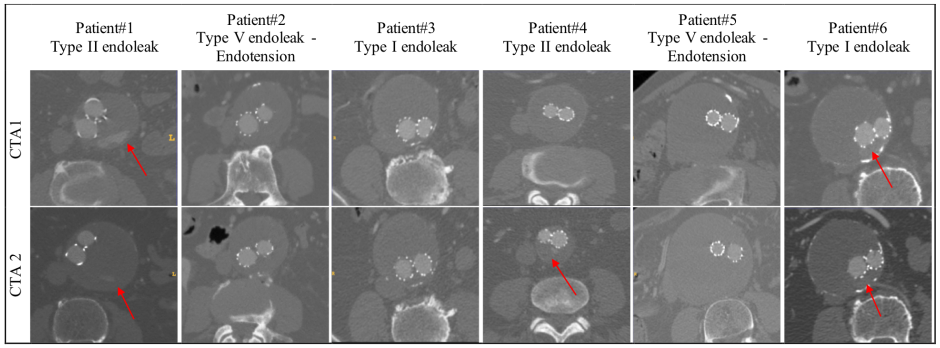

Figure 6.2: Sample slices of the two postoperative CTA series of the patients with unfavorable evolution, where the arrows point to the endoleak when it is visible.

Materials and Methods

Dataset description

Our experiments are run on postoperative CTA datasets of 12 different patients treated with EVAR in Donostia University Hospital (Spain). For each patient, we employ the first two follow-up scans, since our goal is to predict their evolution shortly after the intervention. The datasets have been obtained with scanners of different manufacturers and have a spatial resolution ranging from 0.725 mm to 0.977 mm in x and y, and 0.625 mm-1 mm in z. They also have varying contrast agent doses.

Six patients in our database have a favorable evolution, i.e. the aneurysm shrank. The other 6 patients present complications: two of them have a type I endoleak; another two a type II endoleak; and another two are endotension cases. Only in some cases is the endoleak visible in the image, as shown in Fig 6.2.

Aneurysm segmentation

Radiomic features are extracted from the delineated 3D aneurysm region. The segmentation of the AAA is obtained with the algorithm described in [182], which combines a convolutional neural network and a k-means based post-processing approach to isolate the thrombus in postoperative CTA images. The resultant segmentations are further refined by an expert vascular surgeon with an in-house semi-automatic software to include all image information that could be relevant to characterize the aneurysm.

Radiomic features

In this work, an extensive pool of imaging features are estimated to characterize a range of geometric, functional and appearance properties of the aneurysms that may predict favorable vs. unfavorable patient evolution after EVAR intervention. A total of 104 radiomic features are calculated, which include:

- Shape features, such as sphericity, elongation and diameters.

- First-order statistics, such as intensity variance and skewness, which may identify asymmetries in the intensity distribution.
- Well-established textural features extracted with different methods, such as the Gray Level Co-occurrence Matrix (GLCM) or the Gray Level Run Length Matrix (GLRLM).
- Advanced textural features, such as those computed from the Neighboring Gray Tone Difference Matrix (NGTDM), to identify more localized contrast changes.
- Other descriptors, such as the Fractal Dimension, to characterize the complexity of the AAA appearance.

Note that all these radiomic features are estimated in 3D to obtain an anatomically meaningful analysis of the aneurysm, while existing 2D approaches may lead to information loss.

Classification method

The next step is to combine the heterogeneous radiomic features within a classification scheme that will learn to discriminate patients according to their evolution. We employ the support vector machines (SVM) classifier due to its well-known performance when classifying image data, in particular with small sample sizes. An SVM model finds a hyperplane in the feature space that induces the largest distance to the nearest training data point of any class (so-called functional margin). This ensures that samples belonging to different classes are separated as clearly as possible. New cases are then mapped onto that hyperspace and classified based on their location with respect to the decision boundary.

Feature selection

Due to the large number of extracted radiomic features and the limited number of samples used for training, the classification can easily suffer from overfitting. Thus, it is of paramount importance that we select a smaller and optimized subset of radiomics for the classification task. This is achieved using the sequential forward feature selection (SFFS) method [148, 252], through which features are added to the final subset one at a time until the classification becomes negatively impacted when adding a new feature. Due to their heterogeneity, all radiomic features are normalized to a mean of zero and standard deviation of one to eliminate a potential bias.

Experiments and results

In this work, two different experiments are conducted:

- Experiment 1: The radiomics features are analyzed from the first and the second follow-up CTA scans separately
- Experiment 2: The radiomics features are combined from both postoperative CTA scans

Leave-one-out (LOO) tests are carried out to evaluate the proposed method and the accuracy is measured as the number of correct classifications divided by the test sample size.

Radiomics analysis of the two CTA scans separately

Using only the first postoperative CTA scan, the proposed method yields a classification accuracy of 0.92, corresponding to one misclassification. As shown in Fig. 6.3, the predictive power of the 104 radiomics features in this case varies greatly, which indicates the importance optimal feature selection. Table 6.6 lists the classification results obtained by including into the SVM classifier all the radiomic features for each radiomic type separately (shape, first-order, texture), showing that the texture features provide more predictive evidence than shape and first order features. Table 6.6 also shows the classification accuracy by using the best feature from each type, with the best accuracy achieved by using the gray level variance as the classification variable.

Finally, Fig. 6.4 plots the classification accuracy as a function of the number of radiomic features sequentially added to the SVM classification. It can be seen that no improvement is achieved when adding new features to the classifier or when combining features of different types due to model overfitting. Note that by using the first CTA scan for radiomics-based classification of AAA, the only misclassification corresponds to patient 4 with a Type II endoleak. This is an expected result since type II endoleaks usually have a good prognosis and they can dissappear, although they are considered unpredictable [253].

Subsequently, we tested the classification performance by using the second postoperative CTA scan instead of the first one. In this case, the best classification accuracy is reduced to 0.83, as shown in Table 6.7. In this case, two cases are misclassified, i.e. patient 3 and patient 4. Patient 3 present a very subtle endoleak, which can probably lead the model to misclassifying it.

Type	Accuracy using all features	Best Feature	Accuracy of the Best Feature
Shape	0.58	Maximum 2D diameter	0.83
Histogram/ First-order	0.5	Entropy	0.83
Texture	0.66	Gray level variance (GLCM)	0.92

Table 6.6: Accuracies using 3 types of radiomic features separately for the first CTA series.

Type	Accuracy using all features	Best Feature	Accuracy of the Best Feature
Shape	0.33	Minor axis	0.75
Histogram/ First-order	0.66	Mean	0.75
Texture	0.75	Large dependence emphasis (GLDM)	0.83

Table 6.7: Accuracies using 3 types of radiomic features separately for the second CTA series.

Radiomics analysis of the first and second CTA scans simultaneously

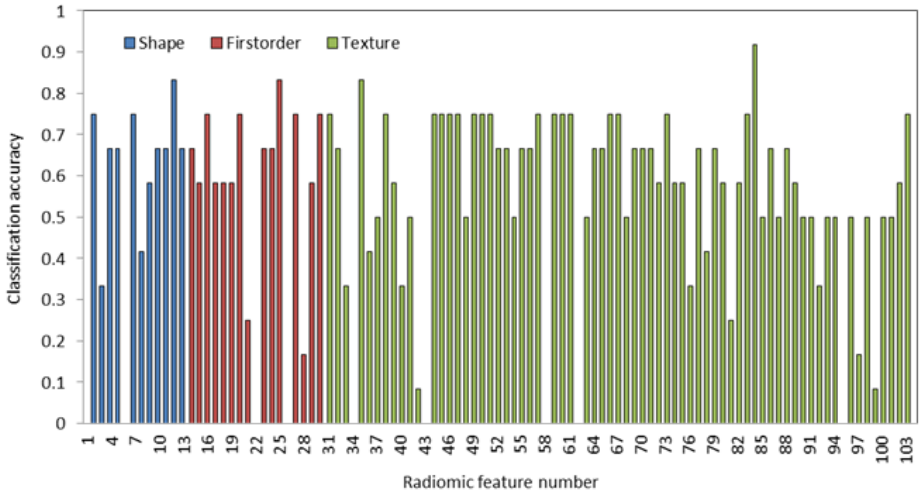

Figure 6.3: Classification accuracy by using each radiomic feature estimated from the first CTA scan.

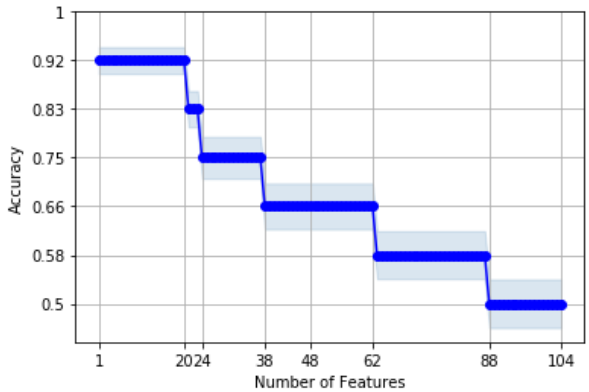

Figure 6.4: Classification accuracy as a function of the number of radiomic features added to the classification model.

In this experiment, all tests are performed by enabling the proposed technique to combine radiomic features from both postoperative scans. However, the obtained classification accuracy is the same as when using only the first CTA scan, i.e. 0.92. Subsequently, we also build a classification model based on features computed as the difference between the radiomics values of the first and second CTA scans. Again, the achieved accuracy remains 0.92. This result suggests that a single CTA scan taken just after the EVAR intervention is sufficient to predict the longer term evolution of a patient. In all these experiments, the misclassified case remains being patient 4 with a type II endoleak.

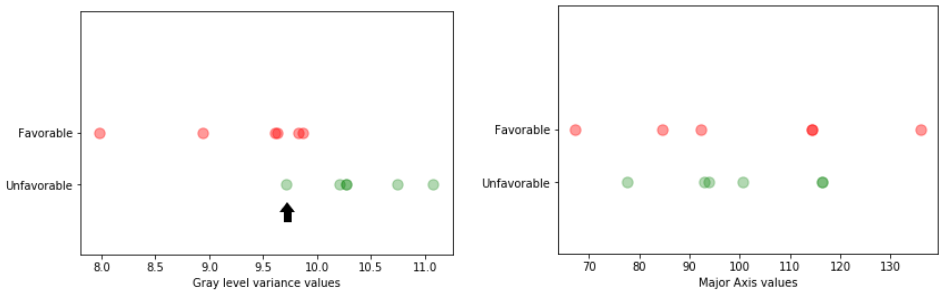

Figure 6.5: Comparative distribution of the radiomic values for the favorable and unfavorable patient groups for selected radiomic features: left image illustrates a radiomic feature with a good predictive power (gray level variance from scan 1), right image illustrates a shape feature that induces overlap between the two subgroups (major axis from scan 2).

Type	Accuracy using all features	Best Feature(s)	Accuracy of the Best Feature(s)
Shape	0.16	Surface to volume ratio	0.75
Histogram/ First-order	0.66	Entropy	0.83
Texture	0.83	Inverse difference moment normalized (GLCM) Correlation (GLCM)	0.92

Table 6.8: Accuracies using 3 types of radiomic features separately computed from the differences between the radiomics values of the first and the second CTA scans.

Experiment	Selected Features	Feature Type	Accuracy
Only first CTA series	Gray level variance (GLCM)	Texture	0.92
Only second CTA series	Large dependence emphasis (GLDM)	Texture	0.83
First and second CTA series	Gray level variance (GLCM)	Texture	0.92
Difference between the CTA series	Inverse difference moment normalized (GLCM) Surface area to volume ratio	Texture Shape	0.92

Table 6.9: Summary of classification results of EVAR patients using different radiomics strategies

Conclusions

This work presents several experiments testing the feasibility of a 3D radiomics for predicting patient evolution after EVAR from postoperative CTA scans. The results, summarized in Table 6.9, show that texture features describing alterations in the 3D aneurysm's appearance are best suited for predicting patient evolution, as they provide information on the presence and type of endoleak. Furthermore, radiomics from the first CTA scans appear to be sufficient identify patients at risk, since a classification accuracy of 0.92 is obtained. However, further research in a larger clinical dataset is required to confirm and further interpret these results, which will be the subject of our future work.

Bibliography

[1] WHO. Cardiovascular diseases (CVDs). *https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)*, (2019), [Online; Accessed 21-February-2022].

[2] P. Peng, K. Lekadir, A. Gooya, L. Shao, S. E. Petersen, and A. F. Frangi. A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging. *MAGMA*, 29(2):155–195, Apr 2016.

[3] Charles Steenbergen and Nikolaos G. Frangogiannis. *Ischemic heart disease*, volume 1, pages 495–521. Elsevier Inc., 2012.

[4] T. J. Pollard. The acute myocardial infarction. *Prim Care*, 27(3):631–649, Sep 2000.

[5] Regina. Bailey. The 3 Layers of the Heart Wall." ThoughtCo. (accessed March 7, 2022).

[6] M. S. Figueroa and J. I. Peters. Congestive heart failure: Diagnosis, pathophysiology, therapy, and implications for respiratory care. *Respir Care*, 51(4):403–412, Apr 2006.

[7] J. I. Hoffman and S. Kaplan. The incidence of congenital heart disease. *J Am Coll Cardiol*, 39(12):1890–1900, Jun 2002.

[8] B. J. Maron, J. A. Towbin, G. Thiene, C. Antzelevitch, D. Corrado, D. Arnett, A. J. Moss, C. E. Seidman, and J. 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, Apr 2006.

[9] B. J. Maron. Hypertrophic cardiomyopathy: a systematic review. *JAMA*, 287(10):1308–1320, Mar 2002.

[10] C. W. Tsao, A. W. Aday, Z. I. Almarzooq, A. Alonso, A. Z. Beaton, M. S. Bittencourt, A. K. Boehme, A. E. Buxton, A. P. Carson, Y. Commodore-Mensah, M. S. V. Elkind, K. R. Evenson, C. Eze-Nliam, J. F. Ferguson, G. Generoso, J. E. Ho, R. Kalani, S. S. Khan, B. M. Kissela, K. L. Knutson, D. A. Levine, T. T. Lewis, J. Liu, M. S. Loop, J. Ma, M. E. Mussolino, S. D. Navaneethan, A. M. Perak, R. Poudel, M. Rezk-Hanna, G. A. Roth, E. B. Schroeder, S. H. Shah, E. L. Thacker, L. B. VanWagner, S. S. Virani, J. H. Voecks, N. Y. Wang, K. Yaffe, and S. S. Martin. Heart Disease and Stroke Statistics-2022 Update: A Report From the American Heart Association. *Circulation*, page CIR0000000000001052, Jan 2022.

[11] R. Blankstein. Cardiology patient page. Introduction to noninvasive cardiac imaging. *Circulation*, 125(3):e267–271, Jan 2012.

[12] Jose Luis Zamorano, Jeroen Bax, Juhani Knuuti, Patrizio Lancellotti, Fausto Pinto, Bogdan A. Popescu, and Udo Sechtem. *The ESC Textbook of Cardiovascular Imaging*. Oxford University Press, 2015.

[13] Raymond J. Gibbons and Philip A. Araoz. The year in cardiac imaging. *Journal of the American College of Cardiology*, 44(10):1937–1944, 2004.

[14] John Pierre Greenwood, Neil Maredia, John F. Younger, Julia M. Brown, Jane Nixon, Colin C. Everett, Petra Bijsterveld, John P. Ridgway, Aleksandra Radjenovic, Catherine J Dickinson, Stephen G. Ball, and Sven Plein. Cardiovascular magnetic resonance and single-photon emission computed tomography for diagnosis of coronary heart disease (CE-MARC): a prospective trial. *Lancet*, 379:453 – 460, 2012.

[15] W. Gregory Hundley, David A. Bluemke J. Paul Finn, Scott D. Flamm, Mark A. Fogel, Matthias G. Friedrich, Vincent B. Ho, Michael Jerosch-Herold, Christopher M. Kramerand Warren J. Manning, Manesh Patel, Gerald M. Pohost, Arthur E. Stillman, Richard D. White, and Pamela K. Woodard. 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, Jun 2010.

[16] 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, Alistair A. Young, Steve Garratt, Tim Peakman, Jonathan Sellors, Rory Collins, and Stefan Neubauer. Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of UK Biobank - rationale, challenges and approaches. *J Cardiovasc Magn Reson*, 15:46, May 2013.

[17] D. J. Pennell. Cardiovascular magnetic resonance. *Circulation*, 121(5):692–705, Feb 2010.

[18] Sang Eun Lee, Christopher Nguyen, Yibin Xie, Zixin Deng, Zhengwei Zhou, Debiao Li, and Hyuk Jae Chang. Recent Advances in Cardiac Magnetic Resonance Imaging. *Korean Circ J*, 49(2):146–159, Feb 2019.

[19] Dudley J. Pennell, Udo P. Sechtem, Charles B. Higgins, Warren J. Manning, Gerald M. Pohost, Frank E. Rademakers, Albert C. van Rossum, Leslee J. Shaw, and E. Kent Yucel. Clinical indications for cardiovascular magnetic resonance (CMR): Consensus Panel report. *Eur Heart J*, 25(21):1940–1965, Nov 2004.

[20] Theresa A McDonagh, Marco Metra, Marianna Adamo, Roy S Gardner, Andreas Baumbach, Michael Böhm, Haran Burri, Javed Butler, Jelena Čelutkienė, Ovidiu Chioncel, John G F Cleland, Andrew J S Coats, Maria G Crespo-Leiro, Dimitrios Farmakis, Martine Gilard, Stephane Heymans, Arno W Hoes, Tiny Jaarsma, Ewa A Jankowska, Mitja Lainscak, Carolyn S P Lam, Alexander R Lyon, John J V McMurray, Alexandre Mebazaa, Richard Mindham, Claudio Muneretto, Massimo Francesco Piepoli, Susanna Price, Giuseppe M C Rosano, Frank Ruschitzka, Anne Kathrine Skibelund, and ESC Scientific Document Group. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: Developed by the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) With the special contribution of the Heart Failure Association (HFA) of the ESC. *European Heart Journal*, 42(36):3599–3726, 08 2021.

[21] 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. *AJR Am J Roentgenol*, 197(4):808–815, Oct 2011.

[22] G. S. Gulsin, A. Singh, and G. P. McCann. Cardiovascular magnetic resonance in the evaluation of heart valve disease. *BMC Med Imaging*, 17(1):67, 12 2017.

[23] Ahmed L. Fathala. Cardiac magnetic resonance imaging: A teaching atlas with emphasizing current clinical indications. *J Saudi Heart Assoc*, 23(4):255–266, Oct 2011.

[24] Michael Salerno, Behzad Sharif, Håkan Arheden, Andreas Kumar, Leon Axel, Debiao Li, and Stefan Neubauer. Recent Advances in Cardiovascular Magnetic Resonance: Techniques and Applications. *Circ Cardiovasc Imaging*, 10(6):e003951, Jun 2017.

[25] Sandra Pujadas, Gautham P Reddy, Oliver Weber, Jennifer J Lee, and Charles B Higgins. MR imaging assessment of cardiac function. *J Magn Reson Imaging*, 19(6):789–799, Jun 2004.

[26] F. Grothues, G. C. Smith, J. C. Moon, N. G. Bellenger, P. Collins, H. U. Klein, and D. J. Pennell. Comparison of interstudy reproducibility of cardiovascular magnetic resonance with two-dimensional echocardiography in normal subjects and in patients with heart failure or left ventricular hypertrophy. *Am J Cardiol*, 90(1):29–34, Jul 2002.

[27] B. H. Lorell, B. A. Carabello, and M. Schneider. Left ventricular hypertrophy: pathogenesis, detection, and prognosis. *Circulation*, 102(4):470–479, Jul 2000.

[28] M. Carlsson, R. Andersson, K. M. Bloch, K. Steding-Ehrenborg, H. Mosén, F. Stahlberg, B. Ekmehag, and H. Arheden. Cardiac output and cardiac index measured with cardiovascular magnetic resonance in healthy subjects, elite athletes and patients with congestive heart failure. *J Cardiovasc Magn Reson*, 14:51, Jul 2012.

[29] M. G. Sutton and N. Sharpe. Left ventricular remodeling after myocardial infarction: pathophysiology and therapy. *Circulation*, 101(25):2981–2988, Jun 2000.

[30] Bradley J. Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L. Kline. Machine learning for medical imaging. *RadioGraphics*, 37(2):505–515, 2017.

[31] D. Dey, P. J. Slomka, P. Leeson, D. Comaniciu, S. Shrestha, P. P. Sengupta, and T. H. Marwick. Artificial Intelligence in Cardiovascular Imaging: JACC State-of-the-Art Review. *J Am Coll Cardiol*, 73(11):1317–1335, 03 2019.

[32] H. Seo, M. Badiei Khuzani, V. Vasudevan, C. Huang, H. Ren, R. Xiao, X. Jia, and L. Xing. Machine learning techniques for biomedical image segmentation: An overview of technical aspects and introduction to state-of-art applications. *Med Phys*, 47(5):e148–e167, Jun 2020.

[33] Marleen de Bruijne. Machine learning approaches in medical image analysis: From detection to diagnosis. *Medical Image Analysis*, 33:94–97, 2016. 20th anniversary of the Medical Image Analysis journal (MedIA).

[34] Muneer Ahmad Dedmari, Sailesh Conjeti, Santiago Estrada, Phillip Ehses, Tony Stöcker, and Martin Reuter. Complex fully convolutional neural networks for mr image reconstruction. In Florian Knoll, Andreas Maier, and Daniel Rueckert, editors, *Machine Learning for Medical Image Reconstruction*, pages 30–38, Cham, 2018. Springer International Publishing.

[35] S. Sanchez-Martinez, O. Camara, G. Piella, M. Cikes, M. A. González Ballester, Vellido A. Miron, M., E. Gomez, A. Fraser, and B. Bijnens. Machine learning for clinical decision-making: Challenges and opportunities in Cardiovascular imaging. *Frontiers in Cardiovascular Medicine*, 8, 2022.

[36] T. Leiner, D. Rueckert, A. Suinesiaputra, B. Baeßler, R. Nezafat, I. Išgum, and A. A. Young. Machine learning in cardiovascular magnetic resonance: basic concepts and applications. *J Cardiovasc Magn Reson*, 21(1):61, 10 2019.

[37] Y. Xu and R. Goodacre. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *J Anal Test*, 2(3):249–262, 2018.

[38] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol*, 19(1):64, 03 2019.

[39] R. J. Gillies, P. E. Kinahan, and H. Hricak. Radiomics: Images Are More than Pictures, They Are Data. *Radiology*, 278(2):563–577, Feb 2016.

[40] P. Lambin, R. T. H. Leijenaar, T. M. Deist, J. Peerlings, E. E. C. de Jong, J. van Timmeren, S. Sanduleanu, R. T. H. M. Larue, A. J. G. Even, A. Jochems, Y. van Wijk, H. Woodruff, J. van Soest, T. Lustberg, E. Roelofs, W. van Elmpt, A. Dekker, F. M. Mottaghy, J. E. Wildberger, and S. Walsh. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*, 14(12):749–762, Dec 2017.

[41] Marius E. Mayerhoefer, Andrzej Materka, Georg Langs, Ida Häggström, Piotr Szczypiński, Peter Gibbs, and Gary Cook. Introduction to radiomics. *Journal*

of Nuclear Medicine, 61(4):488–495, 2020.

[42] Alex Zwanenburg, Martin Vallières, Mahmoud A. Abdalah, Hugo J. W. L. Aerts, Vincent Andrearczyk, Aditya Apte, Saeed Ashrafinia, Spyridon Bakas, Roelof J. Beukinga, Ronald Boellaard, Marta Bogowicz, Luca Boldrini, Irène Buvat, Gary J. R. Cook, Christos Davatzikos, Adrien Depeursinge, Marie-Charlotte Desseroit, Nicola Dinapoli, Cuong Viet Dinh, Sebastian Echegaray, Issam El Naqa, Andriy Y. Fedorov, Roberto Gatta, Robert J. Gillies, Vicky Goh, Michael Götz, Matthias Guckenberger, Sung Min Ha, Mathieu Hatt, Fabian Isensee, Philippe Lambin, Stefan Leger, Ralph T.H. Leijenaar, Jacopo Lenkowicz, Fiona Lippert, Are Losnegård, Klaus H. Maier-Hein, Olivier Morin, Henning Müller, Sandy Napel, Christophe Nioche, Fanny Orlhac, Sarthak Pati, Elisabeth A.G. Pfaehler, Arman Rahmim, Arvind U.K. Rao, Jonas Scherer, Muhammad Musib Siddique, Nanna M. Sijtsema, Jairo Socarras Fernandez, Emiliano Spezi, Roel J.H.M. Steenbakkers, Stephanie Tanadini-Lang, Daniela Thorwarth, Esther G.C. Troost, Taman Upadhaya, Vincenzo Valentini, Lisanne V. van Dijk, Joost van Griethuysen, Floris H.P. van Velden, Philip Whybra, Christian Richter, and Steffen Löck. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, 295(2):328–338, 2020. PMID: 32154773.

[43] H. Y. Kim. Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restor Dent Endod*, 38(1):52–54, Feb 2013.

[44] Evangelos K Oikonomou, Musib Siddique, and Charalambos Antoniades. Artificial intelligence in medical imaging: A radiomic guide to precision phenotyping of cardiovascular disease. *Cardiovascular Research*, 116(13):2040–2054, 02 2020.

[45] H. Yu, K. Buch, B. Li, M. O'Brien, J. Soto, H. Jara, and S. W. Anderson. Utility of texture analysis for quantifying hepatic fibrosis on proton density MRI. *J Magn Reson Imaging*, 42(5):1259–1265, Nov 2015.

[46] G. Thibault, J. Angulo, and F. Meyer. Advanced statistical matrices for texture characterization: application to cell classification. *IEEE Trans Biomed Eng*, 61(3):630–637, Mar 2014.

[47] Mary M. Galloway. Texture analysis using gray level run lengths. *Computer Graphics and Image Processing*, 4(2):172–179, 1975.

[48] Chengjun Sun and William G. Wee. Neighboring gray level dependence matrix for texture classification. *Comput. Graph. Image Process.*, 20:297, 1982.

[49] M. Amadasun and R. King. Textural features corresponding to textural properties. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(5):1264–1274, 1989.

[50] M. Kolossváry, M. Kellermayer, B. Merkely, and P. Maurovich-Horvat. Cardiac Computed Tomography Radiomics: A Comprehensive Review on Radiomic Techniques. *J Thorac Imaging*, 33(1):26–34, Jan 2018.

[51] L Cuthbert and V M Huynh. Statistical analysis of optical fourier transform patterns for surface texture assessment. *Measurement Science and Technology*, 3(8):740–745, aug 1992.

[52] J. Yao, P. Krolak, and C. Steele. The generalized Gabor transform. *IEEE Trans Image Process*, 4(7):978–988, 1995.

[53] E. Florez, A. Fatemi, P. P. Claudio, and C. M. Howard. Emergence of Radiomics: Novel Methodology Identifying Imaging Biomarkers of Disease in Diagnosis, Response, and Progression. *SM J Clin Med Imaging*, 4(1), 2018.

[54] J. Zhou, J. Lu, C. Gao, J. Zeng, C. Zhou, X. Lai, W. Cai, and M. Xu. Predicting the response to neoadjuvant chemotherapy for breast cancer: wavelet transforming radiomics in MRI. *BMC Cancer*, 20(1):100, Feb 2020.

[55] A. Materka. Texture analysis methodologies for magnetic resonance imaging. *Dialogues Clin Neurosci*, 6(2):243–250, Jun 2004.

[56] A. Laine and J. Fan. Texture classification by wavelet packet signatures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1186–1191, 1993.

[57] J. E. van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler. Radiomics in medical imaging-"how-to" guide and critical reflection. *Insights Imaging*, 11(1):91, Aug 2020.

[58] Zhenyu Liu, Shuo Wang, Di Dong, Jingwei Wei, Cheng Fang, Xuezhi Zhou, Kai Sun, Longfei Li, Bo Li, Meiyun Wang, and Jie Tian. The applications of radiomics in precision diagnosis and treatment of oncology: Opportunities and challenges. *Theranostics*, 9(5):1303—1322, 2019.

[59] R. Cattell, S. Chen, and C. Huang. Robustness of radiomic features in magnetic resonance imaging: review and a phantom study. *Vis Comput Ind Biomed Art*, 2(1):19, Nov 2019.

[60] C. Bailly, C. Bodet-Milin, S. Couespel, H. Necib, F. Kraeber-Bodéré, C. Ansquer, and T. Carlier. Revisiting the Robustness of PET-Based Textural Features in the Context of Multi-Centric Trials. *PLoS One*, 11(7):e0159984, 2016.

[61] Ralph T. H. Leijenaar, Georgi I. Nalbantov, Sara Carvalho, Wouter J. C. van Elmpt, Esther G. C. Troost, Ronald Boellaard, Hugo J.W.L. Aerts, Robert James Gillies, and Philippe Lambin. The effect of suv discretization in quantitative fdg-pet radiomics: the need for standardized methodology in tumor texture analysis. *Scientific Reports*, 5, 2015.

[62] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J. C. Fillion-Robin, S. Pieper, and H. J. W. L. Aerts. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res*, 77(21):e104–e107, 11 2017.

[63] M. Mannil, J. von Spiczak, R. Manka, and H. Alkadhi. Texture Analysis and Machine Learning for Detecting Myocardial Infarction in Noncontrast Low-Dose Computed Tomography: Unveiling the Invisible. *Invest Radiol*, 53(6):338–343, 06 2018.

[64] B. Baessler, M. Mannil, S. Oebel, D. Maintz, H. Alkadhi, and R. Manka. Subacute and Chronic Left Ventricular Myocardial Scar: Accuracy of Texture Analysis on Nonenhanced Cine MR Images. *Radiology*, 286(1):103–112, 01 2018.

[65] P. Yin, N. Mao, C. Zhao, J. Wu, C. Sun, L. Chen, and N. Hong. Comparison of radiomics machine-learning classifiers and feature selection for differentiation of

sacral chordoma and sacral giant cell tumour based on 3D computed tomography features. *Eur Radiol*, 29(4):1841–1847, Apr 2019.

[66] X. Zhang, X. Xu, Q. Tian, B. Li, Y. Wu, Z. Yang, Z. Liang, Y. Liu, G. Cui, and H. Lu. Radiomics assessment of bladder cancer grade using texture features from diffusion-weighted imaging. *J Magn Reson Imaging*, 46(5):1281–1288, 11 2017.

[67] B. Zhang, X. He, F. Ouyang, D. Gu, Y. Dong, L. Zhang, X. Mo, W. Huang, J. Tian, and S. Zhang. Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma. *Cancer Lett*, 403:21–27, 09 2017.

[68] E. Thornton, K. M. Krajewski, K. N. O'Regan, A. A. Giardino, J. P. Jagannathan, and N. Ramaiya. Imaging features of primary and secondary malignant tumours of the sacrum. *Br J Radiol*, 85(1011):279–286, Mar 2012.

[69] Darcie A. P. Delzell, Sara Magnuson, Tabitha Peter, Michelle Smith, and Brian J. Smith. Machine learning and feature selection methods for disease classification with application to lung cancer screening image data. *Frontiers in Oncology*, 9, 2019.

[70] B. Pinamonti, E. Picano, E. M. Ferdeghini, F. Lattanzi, G. Slavich, L. Landini, F. Camerini, A. Benassi, A. Distante, and A. L'Abbate. Quantitative texture analysis in two-dimensional echocardiography: application to the diagnosis of myocardial amyloidosis. *J Am Coll Cardiol*, 14(3):666–671, Sep 1989.

[71] F. Lattanzi, P. Bellotti, E. Picano, F. Chiarella, M. Paterni, G. Forni, L. Landini, A. Distante, and C. Vecchio. Quantitative Texture Analysis in Two-Dimensional Echocardiography: Application to the Diagnosis of Myocardial Hemochromatosis. *Echocardiography*, 13(1):9–20, Jan 1996.

[72] E. K. Oikonomou, M. C. Williams, C. P. Kotanidis, M. Y. Desai, M. Marwan, A. S. Antonopoulos, K. E. Thomas, S. Thomas, I. Akoumianakis, L. M. Fan, S. Kesavan, L. Herdman, A. Alashi, E. H. Centeno, M. Lyasheva, B. P. Griffin, S. D. Flamm, C. Shirodaria, N. Sabharwal, A. Kelion, M. R. Dweck, E. J. R. Van Beek, J. Deanfield, J. C. Hopewell, S. Neubauer, K. M. Channon, S. Achenbach, D. E. Newby, and C. Antoniades. A novel machine learning-derived radiotranscriptomic signature of perivascular fat improves cardiac risk prediction using coronary CT angiography. *Eur Heart J*, 40(43):3529–3543, 11 2019.

[73] A. Larroza, M. P. López-Lereu, J. V. Monmeneu, J. Gavara, F. J. Chorro, V. Bodí, and D. Moratal. Texture analysis of cardiac cine magnetic resonance imaging to detect nonviable segments in patients with chronic myocardial infarction. *Med Phys*, 45(4):1471–1480, Apr 2018.

[74] B. Baeßler, M. Mannil, D. Maintz, H. Alkadhi, and R. Manka. Texture analysis and machine learning of non-contrast T1-weighted MR images in patients with hypertrophic cardiomyopathy-Preliminary results. *Eur J Radiol*, 102:61–67, May 2018.

[75] Zahra Raisi-Estabragh, Cristian Izquierdo, Victor M Campello, Carlos Martin-Isla, Akshay Jaggi, Nicholas C Harvey, Karim Lekadir, and Steffen E Petersen. Cardiac magnetic resonance radiomics: basic principles and clinical perspectives. *European Heart Journal - Cardiovascular Imaging*, 21(4):349–356, 03 2020.

[76] Tommaso Di Noto, Jochen von Spiczak, Manoj Mannil, Elena Gantert, Paolo Soda, Robert Manka, and Hatem Alkadhi. Radiomics for distinguishing myocardial infarction from myocarditis at late gadolinium enhancement at mri: Comparison with subjective visual analysis. *Radiology: Cardiothoracic Imaging*, 1(5):e180026, 2019.

[77] Fei Peng, Tian Zheng, Xiaoping Tang, Qiao Liu, Zijing Sun, Zhaofeng Feng, Heng Zhao, and Lianggeng Gong. Magnetic resonance texture analysis in myocardial infarction. *Frontiers in Cardiovascular Medicine*, 8, 2021.

[78] Cameron Hassani, Farhood Saremi, Bino A. Varghese, and Vinay Duddalwar. Myocardial radiomics in cardiac mri. *American Journal of Roentgenology*, 214(3):536–545, 2020.

[79] A. S. Antonopoulos, M. Boutsikou, S. Simantiris, A. Angelopoulos, G. Lazaros, I. Panagiotopoulos, E. Oikonomou, M. Kanoupaki, D. Tousoulis, R. H. Mohiaddin, K. Tsioufis, and C. Vlachopoulos. Machine learning of native T1 mapping radiomics for classification of hypertrophic cardiomyopathy phenotypes. *Sci Rep*, 11(1):23596, 12 2021.

[80] Zahra Raisi-Estabragh, Akshay Jaggi, Polyxeni Gkontra, Celeste McCracken, Nay Aung, Patricia B. Munroe, Stefan Neubauer, Nicholas C. Harvey, Karim Lekadir, and Steffen E. Petersen. Cardiac magnetic resonance radiomics reveal differential impact of sex, age, and vascular risk factors on cardiac structure and myocardial tissue. *Frontiers in Cardiovascular Medicine*, 8, 2021.

[81] R. Schofield, B. Ganeshan, M. Fontana, A. Nasis, S. Castelletti, S. Rosmini, T. A. Treibel, C. Manisty, R. Endozo, A. Groves, and J. C. Moon. Texture analysis of cardiovascular magnetic resonance cine images differentiates aetiologies of left ventricular hypertrophy. *Clin Radiol*, 74(2):140–149, 02 2019.

[82] Rebecca E. Thornhill, Myra S. Cocker, Girish Dwivedi, Carole Dennie, Lyanne Fuller, Alexander Dick, Terrence D. Ruddy, and Elena Peña. Quantitative texture features as objective metrics of enhancement heterogeneity in hypertrophic cardiomyopathy. *Journal of Cardiovascular Magnetic Resonance*, 16:P351 – P351, 2014.

[83] B. Baessler, C. Luecke, J. Lurz, K. Klingel, M. von Roeder, S. de Waha, C. Besler, D. Maintz, M. Gutberlet, H. Thiele, and P. Lurz. Cardiac MRI Texture Analysis of T1 and T2 Maps in Patients with Infarctlike Acute Myocarditis. *Radiology*, 289(2):357–365, 11 2018.

[84] Lennart Tautz, Hannu Zhang, Markus Hüllebrand, Matthias Ivantsits, Sebastian Kelle, Titus Kuehne, Volkmar Falk, and Anja Hennemuth. Cardiac radiomics: an interactive approach for 4d data exploration. *Current Directions in Biomedical Engineering*, 6(1):20200008, 2020.

[85] Matthias Ivantsits, Markus Huellebrand, Sebastian Kelle, Stefan O. Schönberg, Titus Kuehne, and Anja Hennemuth. Deep-learning-based myocardial pathology detection. In Esther Puyol Anton, Mihaela Pop, Maxime Sermesant, Victor Campello, Alain Lalande, Karim Lekadir, Avan Suinesiaputra, Oscar Camara, and Alistair Young, editors, *Statistical Atlases and Computational Models of the Heart. M&Ms*

and EMIDEC Challenges, pages 369–377, Cham, 2021. Springer International Publishing.

[86] J. E. van Timmeren, R. T. H. Leijenaar, W. van Elmpt, J. Wang, Z. Zhang, A. Dekker, and P. Lambin. Test-Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? *Tomography*, 2(4):361–365, Dec 2016.

[87] A. Chalkidou, M. J. O'Doherty, and P. K. Marsden. False Discovery Rates in PET and CT Studies with Texture Features: A Systematic Review. *PLoS One*, 10(5):e0124165, 2015.

[88] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Med Image Anal*, 42:60–88, Dec 2017.

[89] D. Shen, G. Wu, and H. I. Suk. Deep Learning in Medical Image Analysis. *Annu Rev Biomed Eng*, 19:221–248, 06 2017.

[90] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. http://www.deeplearningbook.org.

[91] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[92] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel González Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohé, Xavier Pennec, Maxime Sermesant, Fabian Isensee, Paul Jäger, Klaus H. Maier-Hein, Peter M. Full, Ivo Wolf, Sandy Engelhardt, Christian F. Baumgartner, Lisa M. Koch, Jelmer M. Wolterink, Ivana Išgum, Yeonggul Jang, Yoonmi Hong, Jay Patravali, Shubham Jain, Olivier Humbert, and Pierre-Marc Jodoin. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018.

[93] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[94] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(6), 2020.

[95] Brian McCrindle, Katherine Zukotynski, Thomas E. Doyle, and Michael D. Noseworthy. A radiology-focused review of predictive uncertainty for ai interpretability in computer-assisted segmentation. *Radiology: Artificial Intelligence*, 3(6):e210031, 2021.

[96] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael A. Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced*

Analytics (DSAA), pages 80–89, 2018.

[97] Wojciech Samek and Klaus-Robert Müller. *Towards Explainable Artificial Intelligence*, pages 5–22. Springer International Publishing, Cham, 2019.

[98] Guang Yang, Qinghao Ye, and Jun Xia. Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion*, 77:29–52, 2022.

[99] T. L. Chaunzwa, A. Hosny, Y. Xu, A. Shafer, N. Diao, M. Lanuti, D. C. Christiani, R. H. Mak, and H. J. W. L. Aerts. Deep learning classification of lung cancer histology using CT images. *Sci Rep*, 11(1):5471, 03 2021.

[100] J. Choi, H. H. Cho, J. Kwon, H. Y. Lee, and H. Park. A Cascaded Neural Network for Staging in Non-Small Cell Lung Cancer Using Pre-Treatment CT. *Diagnostics (Basel)*, 11(6), Jun 2021.

[101] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[102] WenQian Liu, Runze Li, Meng Zheng, Srikrishna Karanam, Ziyan Wu, Bir Bhanu, Richard J. Radke, and Octavia I. Camps. Towards visually explaining variational autoencoders. pages 8639–8648, 2020.

[103] Carlo Biffi, Juan J. Cerrolaza, Giacomo Tarroni, Wenjia Bai, Antonio de Marvao, Ozan Oktay, Christian Ledig, Loic Le Folgoc, Konstantinos Kamnitsas, Georgia Doumou, Jinming Duan, Sanjay K. Prasad, Stuart A. Cook, Declan P. O'Regan, and Daniel Rueckert. Explainable anatomical shape analysis through deep hierarchical generative models. *IEEE Transactions on Medical Imaging*, 39:2088–2099, 2020.

[104] Salome Kazeminia, Christoph Baur, Arjan Kuijper, Bram van Ginneken, Nassir Navab, Shadi Albarqouni, and Anirban Mukhopadhyay. Gans for medical image analysis. *Artificial Intelligence in Medicine*, 109:101938, 2020.

[105] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[106] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680. Curran Associates, Inc., 2014.

[107] Gaëtan Hadjeres, Frank Nielsen, and François Pachet. GLSR-VAE: Geodesic latent space regularization for variational autoencoder architectures. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7, 2017.

[108] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. In *Proceedings of the 31st International Conference on Neural Informa-*

tion Processing Systems, NIPS'17, page 5969–5978, Red Hook, NY, USA, 2017. Curran Associates Inc.

[109] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018.

[110] Agisilaos Chartsias, Thomas Joyce, Giorgos Papanastasiou, Scott Semple, Michelle Williams, David E. Newby, Rohan Dharmakumar, and Sotirios A. Tsaftaris. Disentangled representation learning in cardiac image analysis. *Medical Image Analysis*, 58:101535, 2019.

[111] Ashis Pati and Alexander Lerch. Attribute-based Regularization of Latent Spaces for Variational Auto-Encoders. *Neural Computing and Applications*, 2020.

[112] Gaetano Santulli. Epidemiology of cardiovascular disease in the 21st century: Updated updated numbers and updated facts. *Journal of Cardiovascular Disease Research*, 1(1), July 2013.

[113] P. Peng, K. Lekadir, A. Gooya, L. Shao, S. E. Petersen, and A. F. Frangi. A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging. *MAGMA*, 29(2):155–195, Apr 2016.

[114] R. Merai, C. Siegel, M. Rakotz, P. Basch, J. Wright, B. Wong, and P. Thorpe. CDC Grand Rounds: A Public Health Approach to Detect and Control Hypertension. *MMWR Morb. Mortal. Wkly. Rep.*, 65(45):1261–1264, Nov 2016.

[115] A. Wibmer, H. Hricak, T. Gondo, K. Matsumoto, H. Veeraraghavan, D. Fehr, J. Zheng, D. Goldman, C. Moskowitz, S. W. Fine, V. E. Reuter, J. Eastham, E. Sala, and H. A. Vargas. Haralick texture analysis of prostate MRI: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason scores. *Eur Radiol*, 25(10):2840–2850, Oct 2015.

[116] R. Thawani, M. McLane, N. Beig, S. Ghose, P. Prasanna, V. Velcheti, and A. Madabhushi. Radiomics and radiogenomics in lung cancer: A review for the clinician. *Lung Cancer*, 115:34–41, 01 2018.

[117] A. Ahmed, P. Gibbs, M. Pickles, and L. Turnbull. Texture analysis in assessment and prediction of chemotherapy response in breast cancer. *J Magn Reson Imaging*, 38(1):89–101, Jul 2013.

[118] T. P. Coroller, V. Agrawal, E. Huynh, V. Narayan, S. W. Lee, R. H. Mak, and H. J. W. L. Aerts. Radiomic-Based Pathological Response Prediction from Primary Tumors and Lymph Nodes in NSCLC. *J Thorac Oncol*, 12(3):467–476, 03 2017.

[119] H. J. Aerts, E. R. Velazquez, R. T. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, S. Cavalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*, 5:4006, Jun 2014.

[120] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, and H. J. Aerts. Machine Learning methods for Quantitative Radiomic Biomarkers. *Sci Rep*, 5, 2015.

[121] A. Kotrotsou, P. O. Zinn, and R. R. Colen. Radiomics in Brain Tumors: An Emerging Technique for Characterization of Tumor Environment. *Magn Reson Imaging Clin N Am*, 24, 2016.

[122] J. G. Bosch, F. Nijland, S. C. Mitchell, B. P. Lelieveldt, O. Kamp, J. H. Reiber, and M. Sonka. Computer-aided diagnosis via model-based shape analysis: automated classification of wall motion abnormalities in echocardiograms. *Acad Radiol*, 12(3):358–367, Mar 2005.

[123] F. Zhao, H. Zhang, A. Wahle, M. T. Thomas, A. H. Stolpen, T. D. Scholz, and M. Sonka. Congenital aortic disease: 4D magnetic resonance segmentation and quantitative analysis. *Med Image Anal*, 13(3):483–493, Jun 2009.

[124] A. Suinesiaputra, P. Ablin, X. Alba, M. Alessandrini, J. Allen, W. Bai, S. Cimen, P. Claes, B. R. Cowan, J. D'hooge, N. Duchateau, J. Ehrhardt, A. F. Frangi, A. Gooya, V. Grau, K. Lekadir, A. Lu, A. Mukhopadhyay, I. Oksuz, N. Parajali, X. Pennec, M. Pereanez, C. Pinto, P. Piras, M. M. Rohe, D. Rueckert, D. Saring, M. Sermesant, K. Siddiqi, M. Tabassian, L. Teresi, S. A. Tsaftaris, M. Wilms, A. A. Young, X. Zhang, and P. Medrano-Gracia. Statistical shape modeling of the left ventricle: myocardial infarct classification challenge. *IEEE J Biomed Health Inform*, 22(2):503–515, 03 2018.

[125] Karim Lekadir, Xènia Albí, Marco Pereañez, and Alejandro F. Frangi. Statistical shape modeling using partial least squares: Application to the assessment of myocardial infarction. In *Revised Selected Papers of the 6th International Workshop on Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges - Volume 9534*, page 130–139, Berlin, Heidelberg, 2015. Springer-Verlag.

[126] K. Lekadir, C. Hoogendoorn, M. Pereanez, X. Albà, A. Pashaei, and A. F. Frangi. Statistical personalization of ventricular fiber orientation using shape predictors. *IEEE Trans Med Imaging*, 33(4):882–890, Apr 2014.

[127] A. Suinesiaputra, A. F. Frangi, T. A. Kaandorp, H. J. Lamb, J. J. Bax, J. H. Reiber, and B. P. Lelieveldt. Automated detection of regional wall motion abnormalities based on a statistical model applied to multislice short-axis cardiac MR images. *IEEE Trans Med Imaging*, 28(4):595–607, Apr 2009.

[128] Wenjia Bai, Ozan Oktay, and Daniel Rueckert. Classification of myocardial infarcted patients by combining shape and motion features. In Oscar Camara, Tommaso Mansi, Mihaela Pop, Kawal Rhode, Maxime Sermesant, and Alistair Young, editors, *Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges*, pages 140–145, Cham, 2016. Springer International Publishing.

[129] H. Engblom, J. Tufvesson, R. Jablonowski, M. Carlsson, A. H. Aletras, P. Hoffmann, A. Jacquier, F. Kober, B. Metzler, D. Erlinge, D. Atar, H. Arheden, and E. Heiberg. A new automatic algorithm for quantification of myocardial infarction imaged by late gadolinium enhancement cardiovascular magnetic resonance: experimental validation and comparison to expert delineations in multi-center, multivendor patient data. *J Cardiovasc Magn Reson*, 18(1):27, 05 2016.

[130] A. S. Fahmy, H. El-Rewaidy, M. Nezafat, S. Nakamori, and R. Nezafat. mapping images using fully convolutional neural networks. *J Cardiovasc Magn Reson*, 21(1):7, 01 2019.

[131] Hinrich B. Winther, Christian Hundt, Bertil Schmidt, Christoph Czerner, Johann Bauersachs, Frank Wacker, and Jens Vogel-Claussen. *v*-net: Deep learning for generalized biventricular mass and function parameters using multicenter cardiac mri data. *JACC: Cardiovascular Imaging*, 11(7):1036–1038, 2018.

[132] W. Bai, M. Sinclair, G. Tarroni, O. Oktay, M. Rajchl, G. Vaillant, A. M. Lee, N. Aung, E. Lukaschuk, M. M. Sanghvi, F. Zemrak, K. Fung, J. M. Paiva, V. Carapella, Y. J. Kim, H. Suzuki, B. Kainz, P. M. Matthews, S. E. Petersen, S. K. Piechnik, S. Neubauer, B. Glocker, and D. Rueckert. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *J Cardiovasc Magn Reson*, 20(1):65, 09 2018.

[133] Mihaela Pop, Maxime Sermesant, Pierre-Marc Jodoin, Alain Lalande, Xiahai Zhuang, Guang Yang, Alistair A. Young, and Olivier Bernard. Statistical atlases and computational models of the heart. ACDC and MMWHS challenges. In *Lecture Notes in Computer Science*, 2017.

[134] N. Zhang, G. Yang, Z. Gao, C. Xu, Y. Zhang, R. Shi, J. Keegan, L. Xu, H. Zhang, Z. Fan, and D. Firmin. Deep Learning for Diagnosis of Chronic Myocardial Infarction on Nonenhanced Cardiac Cine MRI. *Radiology*, 291(3):606–617, 06 2019.

[135] Alejandra Moreno, Jefferson Rodriguez, and Fabio Martínez. Regional multiscale motion representation for cardiac disease prediction. In *2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA)*, pages 1–5, 2019.

[136] 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, 2019.

[137] R. T. Larue, G. Defraene, D. De Ruysscher, P. Lambin, and W. van Elmpt. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. *Br J Radiol*, 90(1070):20160665, Feb 2017.

[138] M. Vallières, C. R. Freeman, S. R. Skamene, and I. El Naqa. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Phys Med Biol*, 60(14):5471–5496, Jul 2015.

[139] C. Parmar, P. Grossmann, D. Rietveld, M. M. Rietbergen, P. Lambin, and H. J. Aerts. Radiomic Machine-Learning Classifiers for Prognostic Biomarkers of Head and Neck Cancer. *Front Oncol*, 5:272, 2015.

[140] Bettina Baessler, Manoj Mannil, Sabrina Oebel, David Maintz, Hatem Alkadhi, and Robert Manka. Subacute and chronic left ventricular myocardial scar: Accuracy of texture analysis on nonenhanced cine mr images. *Radiology*, 286(1):103–112, 2018.

[141] Ulf Neisius, Hossam El-Rewaidy, Shiro Nakamori, Jennifer Rodriguez, Warren J.

Manning, and Reza Nezafat. Radiomic analysis of myocardial native t1 imaging discriminates between hypertensive heart disease and hypertrophic cardiomyopathy. *JACC: Cardiovascular Imaging*, 12(10):1946–1954, 2019.

[142] C. Sudlow 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 Med.*, 12(3):e1001779, Mar 2015.

[143] S E. Petersen, P M. Matthews, J M. Francis, M D. Robson, F Zemrak, R Boubertakh, A A. Young, S Hudson, P Weale, S Garratt, R Collins, S Piechnik, and S Neubauer. Uk biobank's cardiovascular magnetic resonance protocol. *Journal of Cardiovascular Magnetic Resonance*, 18(1):8, Feb 2016.

[144] Hrvoje Kalinic. Atlas-based image segmentation: A survey. 2009.

[145] W. Bai, W. Shi, A. de Marvao, T. J. Dawes, D. P. O'Regan, S. A. Cook, and D. Rueckert. A bi-ventricular cardiac atlas built from 1000+ high resolution MR images of healthy subjects and an analysis of shape and motion. *Med Image Anal*, 26(1):133–145, Dec 2015.

[146] M. Lorenzo-Valdés, G. I. Sanchez-Ortiz, R. Mohiaddin, and D. Rueckert. Atlas-based segmentation and tracking of 3d cardiac mr images using non-rigid registration. In Takeyoshi Dohi and Ron Kikinis, editors, *Medical Image Computing and Computer-Assisted Intervention — MICCAI 2002*, pages 642–650, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

[147] Isabelle Guyon. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[148] Sebastian Raschka. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *The Journal of Open Source Software*, 3(24), April 2018.

[149] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[150] Padmavathi Janardhanan, Heena L., and Fathima Sabika. Effectiveness of support vector machines in medical data mining. *Journal of Communications Software and Systems*, 11(1):25–30, 3 2015.

[151] Y. J. Son, H. G. Kim, E. H. Kim, S. Choi, and S. K. Lee. Application of support vector machine for prediction of medication adherence in heart failure patients. *Healthc Inform Res*, 16(4):253–259, Dec 2010.

[152] S. E. Petersen, M. M. Sanghvi, N. Aung, J. A. Cooper, J. M. Paiva, F. Zemrak, K. Fung, E. Lukaschuk, A. M. Lee, V. Carapella, Y. J. Kim, S. K. Piechnik, and S. Neubauer. The impact of cardiovascular risk factors on cardiac structure and function: Insights from the UK Biobank imaging enhancement study. *PLoS One*, 12(10):e0185114, 2017.

[153] C. Martin-Isla, V. M. Campello, C. Izquierdo, Z. Raisi-Estabragh, B. Baeßler, S. E. Petersen, and K. Lekadir. Image-Based Cardiac Diagnosis With Machine Learning: A Review. *Front Cardiovasc Med*, 7:1, 2020.

[154] Hugo J. W. L. Aerts. The Potential of Radiomic-Based Phenotyping in Precision

Medicine: A Review. *JAMA Oncology*, 2(12):1636–1642, 12 2016.

[155] T. P. Coroller, P. Grossmann, Y. Hou, E. Rios Velazquez, R. T. Leijenaar, G. Hermann, P. Lambin, B. Haibe-Kains, R. H. Mak, and H. J. Aerts. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother Oncol*, 114(3):345–350, Mar 2015.

[156] S. Napel, W. Mu, B. V. Jardim-Perassi, H. J. W. L. Aerts, and R. J. Gillies. Quantitative imaging of cancer in the postgenomic era: Radio(geno)mics, deep learning, and habitats. *Cancer*, 124(24):4633–4649, 12 2018.

[157] X. Chen, K. Oshima, D. Schott, H. Wu, W. Hall, Y. Song, Y. Tao, D. Li, C. Zheng, P. Knechtges, B. Erickson, and X. A. Li. Assessment of treatment response during chemoradiation therapy for pancreatic cancer based on quantitative radiomic analysis of daily CTs: An exploratory study. *PLoS One*, 12(6):e0178961, 2017.

[158] Z. Raisi-Estabragh and S. E. Petersen. Cardiovascular research highlights from the UK Biobank: opportunities and challenges. *Cardiovasc Res*, 116(1):e12–e15, 01 2020.

[159] S. E. Petersen, N. Aung, M. M. Sanghvi, F. Zemrak, K. Fung, J. M. Paiva, J. M. Francis, M. Y. Khanji, E. Lukaschuk, A. M. Lee, V. Carapella, Y. J. Kim, P. Leeson, S. K. Piechnik, and S. Neubauer. Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort. *J Cardiovasc Magn Reson*, 19(1):18, Feb 2017.

[160] X. N. Shao, Y. J. Sun, K. T. Xiao, Y. Zhang, W. B. Zhang, Z. F. Kou, and J. L. Cheng. Texture analysis of magnetic resonance T1 mapping with dilated cardiomyopathy: A machine learning approach. *Medicine (Baltimore)*, 97(37):e12246, Sep 2018.

[161] Irem Cetin, Steffen E. Petersen, Sandy Napel, Oscar Camara, Miguel Angel González Ballester, and Karim Lekadir. A radiomics approach to analyze cardiac alterations in hypertension. In *16th IEEE International Symposium on Biomedical Imaging, ISBI 2019, Venice, Italy, April 8-11, 2019*, pages 640–643. IEEE, 2019.

[162] Irem Cetin, Gerard Sanroma, Steffen E. Petersen, Sandy Napel, Oscar Camara, Miguel Angel González Ballester, and Karim Lekadir. A radiomics approach to computer-aided diagnosis with cardiac cine-mri. In Mihaela Pop, Maxime Sermesant, Pierre-Marc Jodoin, Alain Lalande, Xiahai Zhuang, Guang Yang, Alistair A. Young, and Olivier Bernard, editors, *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges - 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers*, volume 10663 of *Lecture Notes in Computer Science*, pages 82–90. Springer, 2017.

[163] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(null):2825–2830, nov 2011.

[164] Stephen W. Looney. A statistical technique for comparing the accuracies of several classifiers. *Pattern Recognition Letters*, 8(1):5–9, 1988.

[165] W. G. COCHRAN. The comparison of percentages in matched samples. *Biometrika*, 37(3-4):256–266, Dec 1950.

[166] W. Nadruz, B. Claggett, A. Gonçalves, G. Querejeta-Roca, M. M. Fernandes-Silva, A. M. Shah, S. Cheng, H. Tanaka, G. Heiss, D. W. Kitzman, and S. D. Solomon. Smoking and Cardiac Structure and Function in the Elderly: The ARIC Study (Atherosclerosis Risk in Communities). *Circ Cardiovasc Imaging*, 9(9):e004950, Sep 2016.

[167] A. Larroza, A. Materka, M. P. López-Lereu, J. V. Monmeneu, V. Bodí, and D. Moratal. Differentiation between acute and chronic myocardial infarction by means of texture analysis of late gadolinium enhancement and cine cardiac magnetic resonance imaging. *Eur J Radiol*, 92:78–83, Jul 2017.

[168] P. Medrano-Gracia, B. R. Cowan, B. Ambale-Venkatesh, D. A. Bluemke, J. Eng, J. P. Finn, C. G. Fonseca, J. A. Lima, A. Suinesiaputra, and A. A. Young. Left ventricular shape variation in asymptomatic populations: the Multi-Ethnic Study of Atherosclerosis. *J Cardiovasc Magn Reson*, 16:56, Jul 2014.

[169] A. A. Young and A. F. Frangi. Computational cardiac atlases: from patient to population and back. *Exp Physiol*, 94(5):578–596, May 2009.

[170] K. Gilbert, W. Bai, C. Mauger, P. Medrano-Gracia, A. Suinesiaputra, A. M. Lee, M. M. Sanghvi, N. Aung, S. K. Piechnik, S. Neubauer, S. E. Petersen, D. Rueckert, and A. A. Young. Independent Left Ventricular Morphometric Atlases Show Consistent Relationships with Cardiovascular Risk Factors: A UK Biobank Study. *Sci Rep*, 9(1):1130, 02 2019.

[171] S. H. Lee, H. H. Cho, H. Y. Lee, and H. Park. Clinical impact of variability on CT radiomics and suggestions for suitable feature selection: a focus on lung cancer. *Cancer Imaging*, 19(1):54, Jul 2019.

[172] H. Shakir, Y. Deng, H. Rasheed, and T. M. R. Khan. Radiomics based likelihood functions for cancer diagnosis. *Sci Rep*, 9(1):9501, 07 2019.

[173] A. Lecler, L. Duron, D. Balvay, J. Savatovsky, O. Bergès, M. Zmuda, E. Farah, O. Galatoire, A. Bouchouicha, and L. S. Fournier. Combining Multiple Magnetic Resonance Imaging Sequences Provides Independent Reproducible Radiomics Features. *Sci Rep*, 9(1):2068, 02 2019.

[174] J. Peerlings, H. C. Woodruff, J. M. Winfield, A. Ibrahim, B. E. Van Beers, A. Heerschap, A. Jackson, J. E. Wildberger, F. M. Mottaghy, N. M. DeSouza, and P. Lambin. Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial. *Sci Rep*, 9(1):4800, 03 2019.

[175] J. E. Park, S. Y. Park, H. J. Kim, and H. S. Kim. Reproducibility and Generalizability in Radiomics Modeling: Possible Strategies in Radiologic and Statistical Perspectives. *Korean J Radiol*, 20(7):1124–1137, 07 2019.

[176] J. E. Park, D. Kim, H. S. Kim, S. Y. Park, J. Y. Kim, S. J. Cho, J. H. Shin, and J. H. Kim. Quality of science and reporting of radiomics in oncologic studies: room for

improvement according to radiomics quality score and TRIPOD statement. *Eur Radiol*, 30(1):523–536, Jan 2020.

[177] Rahul Pitale, Harshvardhan Kale, Sakshi Kshirsagar, and Harshal Rajput. A schematic review on applications of deep learning and computer vision. In *2021 Asian Conference on Innovation in Technology (ASIANCON)*, pages 1–6, 2021.

[178] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017.

[179] Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, Bo Zhao, and Hua Xu. Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, 27(3):457–470, 12 2019.

[180] Li Deng and Yang Liu. *Deep Learning in Natural Language Processing*. Springer Publishing Company, Incorporated, 2018.

[181] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[182] Karen López-Linares, Nerea Aranjuelo, Luis Kabongo, Gregory Maclair, Nerea Lete, Mario Ceresa, Ainhoa García-Familiar, Iván Macía, and González Ballester Miguel Angel. Fully automatic detection and segmentation of abdominal aortic thrombus in post-operative cta images using deep convolutional neural networks. *Medical Image Analysis*, 46:202 – 214, 2018.

[183] T. Jo, K. Nho, and A. J. Saykin. Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data. *Frontiers in Aging Neuroscience*, 11:220, 2019.

[184] T. Higaki, Y. Nakamura, J. Zhou, Z. Yu, T. Nemoto, F. Tatsugami, and K. Awai. Deep learning reconstruction at ct : Phantom study of the image characteristics. *Academic Radiology*, 27(1):82–87, 01 2020.

[185] Andreas Kofler, Markus Haltmeier, Tobias Schaeffter, Marc Kachelriess, Marc Dewey, Christian Wald, and Christoph Kolbitsch. Neural networks-based regularization for large-scale medical image reconstruction. *Physics in medicine and biology*, 65(13):135003, jul 2020.

[186] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[187] Christoph Molnar. *Interpretable Machine Learning*. Second edition, 2022.

[188] Serg Masis. *Interpretable Machine Learning with Python*. Packt Publishing, 2021.

[189] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedan-

tam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[190] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014.

[191] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viegas, and Michael Terry. XRAI: Better attributions through regions. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4947–4956, 2019.

[192] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[193] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019.

[194] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: a review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798—1828, August 2013.

[195] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

[196] Hyunjik Kim and Andriy Mnih. Disentangling by Factorising. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2649–2658. PMLR, 10–15 Jul 2018.

[197] Paul K. Rubenstein, Bernhard Schölkopf, and Ilya O. Tolstikhin. Learning disentangled representations with wasserstein auto-encoders. In *ICLR*, 2018.

[198] Ricky T. Q. Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[199] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124. PMLR, 09–15 Jun 2019.

[200] Nathan Painchaud, Nicolas Duchateau, Olivier Bernard, and Pierre-Marc Jodoin. Echocardiography segmentation with enforced temporal consistency. *ArxiV*, 12

2021.

[201] Alain Lalande, Zhihao Chen, Thomas Decourselle, Abdul Qayyum, Thibaut Pommier, Luc Lorgis, Ezequiel de la Rosa, Alexandre Cochet, Yves Cottin, Dominique Ginhac, Michel Salomon, Raphaël Couturier, and Fabrice Meriaudeau. Emidec: A database usable for the automatic evaluation of myocardial infarction from delayed-enhancement cardiac mri. *Data*, 5(4), 2020.

[202] Stan Lipovetsky and Michael Conklin. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001.

[203] Samanta Knapič, Avleen Malhi, Rohit Saluja, and Kary Främling. Explainable artificial intelligence for human decision support system in the medical domain. *Machine Learning and Knowledge Extraction*, 3(3):740–770, 2021.

[204] R. Du, V. H. Lee, H. Yuan, K. O. Lam, H. H. Pang, Y. Chen, E. Y. Lam, P. L. Khong, A. W. Lee, D. L. Kwong, and V. Vardhanabhuti. Radiomics Model to Predict Early Progression of Nonmetastatic Nasopharyngeal Carcinoma after Intensity Modulation Radiation Therapy: A Multicenter Study. *Radiol Artif Intell*, 1(4):e180075, Jul 2019.

[205] I. Palatnik de Sousa, M. Maria Bernardes Rebuzzi Vellasco, and E. Costa da Silva. Local Interpretable Model-Agnostic Explanations for Classification of Lymph Node Metastases. *Sensors (Basel)*, 19(13), Jul 2019.

[206] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

[207] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018.

[208] Titus J. Brinker, Achim Hekler, Alexander H. Enk, Carola Berking, Sebastian Haferkamp, Axel Hauschild, Michael Weichenthal, Joachim Klode, Dirk Schadendorf, Tim Holland-Letz, Christof von Kalle, Stefan Fröhling, Bastian Schilling, and Jochen S. Utikal. Deep neural networks are superior to dermatologists in melanoma image classification. *European Journal of Cancer*, 119:11–17, 2019.

[209] Eali Stephen Neal Joshua, Midhun Chakkravarthy, and Debnath Bhattacharyya. Lung cancer detection using improvised grad-cam++ with 3d cnn class activation. In Sanjoy Kumar Saha, Paul S. Pang, and Debnath Bhattacharyya, editors, *Smart Technologies in Data Science and Communication*, pages 55–69, Singapore, 2021. Springer Singapore.

[210] Carlo Biffi, Ozan Oktay, Giacomo Tarroni, Wenjia Bai, Antonio de Marvao, Georgia Doumou, Martin Rajchl, Reem Bedair, Sanjay K. Prasad, Stuart A. Cook, Declan P. O'Regan, and Daniel Rueckert. Learning interpretable anatomical features through deep generative models: Application to cardiac remodeling. *ArXiv*, abs/1807.06843, 2018.

[211] James R. Clough, Ilkay Oksuz, Esther Puyol-Antón, Bram Ruijsink, Andrew P. King, and Julia A. Schnabel. Global and local interpretability for cardiac mri clas-

sification. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 656–664, Cham, 2019. Springer International Publishing.

[212] Mahsa Shakeri, Herve Lombaert, Shashank Tripathi, and Samuel Kadoury. Deep spectral-based shape features for alzheimer's disease classification. In Martin Reuter, Christian Wachinger, and Hervé Lombaert, editors, *Spectral and Shape Analysis in Medical Imaging*, pages 15–24, Cham, 2016. Springer International Publishing.

[213] Esther Puyol-Antón, Chen Chen, James R. Clough, Bram Ruijsink, Baldeep S. Sidhu, Justin Gould, Bradley Porter, Marc Elliott, Vishal Mehta, Daniel Rueckert, Christopher A. Rinaldi, and Andrew P. King. Interpretable deep models for cardiac resynchronisation therapy response prediction. pages 284–293, 2020.

[214] Chris Donahue, Zachary Chase Lipton, Akshay Balsubramani, and Julian McAuley. Semantically decomposing the latent spaces of generative adversarial networks. *ArXiv*, abs/1705.07904, 2018.

[215] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *NIPS 2016*, pages 2172–2180, 2016.

[216] Jesse Engel, Matthew D. Hoffman, and Adam Roberts. Latent constraints: Learning to generate conditionally from unconditional generative models. *ArXiv*, abs/1711.05772, 2018.

[217] Tejas D Kulkarni, William F. Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[218] Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem. Disentangling factors of variations using few labels. In *International Conference on Learning Representations*, 4 2020.

[219] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[220] Xiao Liu, Pedro Sanchez, Spyridon Thermos, Alison Q. O'Neil, and Sotirios A. Tsaftaris. Learning disentangled representations in the imaging domain. *CoRR*, abs/2108.12043, 2021.

[221] Christopher P. Burgess, Irina Higgins, Arka Pal, Loïc Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *CoRR*, abs/1804.03599, 2018.

[222] Irina Higgins, Nicolas Sonnerat, Loïc Matthey, Arka Pal, Christopher P. Burgess, Matthew M. Botvinick, Demis Hassabis, and Alexander Lerchner. SCAN: Learning

abstract hierarchical compositional visual concepts. *ArXiv*, 2017.

[223] M. L. Huang, Y. H. Hung, W. M. Lee, R. K. Li, and B. R. Jiang. SVM-RFE based feature selection and Taguchi parameters optimization for multiclass SVM classifier. *ScientificWorldJournal*, 2014:795624, 2014.

[224] Mouhamadou Lamine Samb, Fodé Camara, Samba Ndiaye, Yahya Slimani, Mohamed Amir Esseghir, and Cheikh Anta. A novel RFE-SVM-based feature selection approach for classification. *International Journal of Advanced Science and Technology*, 43, Jun 2012.

[225] X. Yang. Identification of risk genes associated with myocardial infarction based on the recursive feature elimination algorithm and support vector machine classifier. *Mol Med Rep*, 17(1):1555–1560, Jan 2018.

[226] Irem Cetin, Zahra Raisi-Estabragh, Steffen E. Petersen, Sandy Napel, Stefan K. Piechnik Piechnik, Stefan Neubauer, Miguel Angel Gonzalez Ballester, Oscar Camara, and Karim Lekadir. Radiomics signatures of cardiovascular risk factors in cardiac mri: Results from the uk biobank. *Frontiers in Cardiovascular Medicine*, 7, 2020.

[227] G. Xiao, W. C. Rong, Y. C. Hu, Z. Q. Shi, Y. Yang, J. L. Ren, and G. B. Cui. MRI Radiomics Analysis for Predicting the Pathologic Classification and TNM Staging of Thymic Epithelial Tumors: A Pilot Study. *AJR Am J Roentgenol*, 214(2):328–340, 02 2020.

[228] Wei Chen, Boqiang Liu, Suting Peng, Jiawei Sun, and Xu Qiao. Computer-aided grading of gliomas combining automatic segmentation and radiomics. *International Journal of Biomedical Imaging*, 2018:2512037, 2018.

[229] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[230] Karl Ridgeway and Michael C Mozer. Learning deep disentangled embeddings with the f-statistic loss. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[231] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *CoRR*, abs/1711.00848, 2017.

[232] Tameem Adel, Zoubin Ghahramani, and Adrian Weller. Discovering interpretable representations for both deep generative and discriminative models. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 50–59. PMLR, 10–15 Jul 2018.

[233] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, vol-

ume 19. MIT Press, 2007.

[234] M. Reyes, R. Meier, S. Pereira, C. A. Silva, F. M. Dahlweid, H. von Tengg-Kobligk, R. M. Summers, and R. Wiest. On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. *Radiol Artif Intell*, 2(3):e190043, May 2020.

[235] Ana Lourenço, Eric Kerfoot, Irina Grigorescu, Cian M. Scannell, Marta Varela, and Teresa M. Correia. Automatic myocardial disease prediction from delayed-enhancement cardiac mri and clinical information. In Esther Puyol Anton, Mihaela Pop, Maxime Sermesant, Victor Campello, Alain Lalande, Karim Lekadir, Avan Suinesiaputra, Oscar Camara, and Alistair Young, editors, *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, pages 334–341, Cham, 2021. Springer International Publishing.

[236] Jixi Shi, Zhihao Chen, and Raphaël Couturier. Classification of pathological cases of myocardial infarction using convolutional neural network and random forest. In Esther Puyol Anton, Mihaela Pop, Maxime Sermesant, Victor Campello, Alain Lalande, Karim Lekadir, Avan Suinesiaputra, Oscar Camara, and Alistair Young, editors, *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, pages 406–413, Cham, 2021. Springer International Publishing.

[237] Matthias Ivantsits, Markus Huellebrand, Sebastian Kelle, Stefan O. Schönberg, Titus Kuehne, and Anja Hennemuth. Deep-learning-based myocardial pathology detection. In Esther Puyol Anton, Mihaela Pop, Maxime Sermesant, Victor Campello, Alain Lalande, Karim Lekadir, Avan Suinesiaputra, Oscar Camara, and Alistair Young, editors, *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, pages 369–377, Cham, 2021. Springer International Publishing.

[238] Rishabh Sharma, Christoph F. Eick, and Nikolaos V. Tsekos. SM2N2: A stacked architecture for multimodal data and its application to myocardial infarction detection. In Esther Puyol Anton, Mihaela Pop, Maxime Sermesant, Victor Campello, Alain Lalande, Karim Lekadir, Avan Suinesiaputra, Oscar Camara, and Alistair Young, editors, *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, pages 342–350, Cham, 2021. Springer International Publishing.

[239] Alain Lalande, Zhihao Chen, Thibaut Pommier, Thomas Decourselle, Abdul Qayyum, Michel Salomon, D. Ginhac, Youssef Skandarani, Arnaud Boucher, Khawla Brahim, Marleen de Bruijne, Robin Camarasa, Teresa Correia, Xue Feng, Kibrom Berihu Girum, Anja Hennemuth, Markus Huellebrand, Raabid Hussain, Matthias Ivantsits, Jun Ma, Craig H. Meyer, Rishabh Sharma, Jixi Shi, Nikolaos V. Tsekos, Marta Varela, Xiyue Wang, Sen Yang, Hannu Zhang, Yichi Zhang, Yuncheng Zhou, Xiahai Zhuang, Raphaël Couturier, and Fabrice Mériaudeau. Deep learning methods for automatic evaluation of delayed enhancement-mri. the results of the emidec challenge. *CoRR*, abs/2108.04016, 2021.

[240] Kibrom Berihu Girum, Youssef Skandarani, Raabid Hussain, Alexis Bozorg

Grayeli, Gilles Créhange, and Alain Lalande. Automatic myocardial infarction evaluation from delayed-enhancement cardiac mri using deep convolutional networks. In Esther Puyol Anton, Mihaela Pop, Maxime Sermesant, Victor Campello, Alain Lalande, Karim Lekadir, Avan Suinesiaputra, Oscar Camara, and Alistair Young, editors, *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, pages 378–384, Cham, 2021. Springer International Publishing.

[241] Víctor M. Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martín-Isla, Alireza Sojoudi, Peter M. Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, Mario Parreño, Alberto Albiol, Fanwei Kong, Shawn C. Shadden, Jorge Corral Acero, Vaanathi Sundaresan, Mina Saber, Mustafa Elattar, Hongwei Li, Bjoern Menze, Firas Khader, Christoph Haarburger, Cian M. Scannell, Mitko Veta, Adam Carscadden, Kumaradevan Punithakumar, Xiao Liu, Sotirios A. Tsaftaris, Xiaoqiong Huang, Xin Yang, Lei Li, Xiahai Zhuang, David Viladés, Martín L. Descalzo, Andrea Guala, Lucia La Mura, Matthias G. Friedrich, Ria Garg, Julie Lebel, Filipe Henriques, Mahir Karakas, Ersin Çavuş, Steffen E. Petersen, Sergio Escalera, Santi Seguí, José F. Rodríguez-Palomares, and Karim Lekadir. Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge. *IEEE Transactions on Medical Imaging*, 40(12):3543–3554, 2021.

[242] Zohaib Salahuddin, Henry C. Woodruff, Avishek Chatterjee, and Philippe Lambin. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in Biology and Medicine*, 140:105111, 2022.

[243] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9(61):1871–1874, 2008.

[244] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, 16:1190–1208, 1995.

[245] X. Li, G. Zhao, J. Zhang, Z. Duan, and S. Xin. Prevalence and trends of the abdominal aortic aneurysms epidemic in general population - a meta-analysis. *PLoS ONE*, 8(12):e81260, 2013.

[246] SB White and SW Stavropoulos. Management of endoleaks following endovascular aneurysm repair. *Semin Intervent Radiol*, 26(1):33–38, 2009.

[247] Chintan A. Parmar, Ralph T. H. Leijenaar, Patrick Grossmann, Emmanuel Rios Velazquez, Johan Bussink, Derek H. F. Rietveld, M M Rietbergen, Benjamin Haibe-Kains, Philippe Lambin, and Hugo J. W. L. Aerts. Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. In *Scientific reports*, volume 5, 2015.

[248] Ming Zhou, Jacob G Scott, Baishali Chaudhury, Lawrence O Hall, Dmitry B Goldgof, Kristen W. Yeom, M Iv, Yangming Ou, Jayashree Kalpathy-Cramer, Sandy Napel, Robert James Gillies, Olivier Gevaert, and Robert A. Gatenby. Radiomics in brain tumor: Image assessment, quantitative feature descriptors, and machine-learning approaches. *AJNR*, 39 2:208–216, 2018.

[249] Damini Dey and Frederic Commandeur. Radiomics to identify high-risk atherosclerotic plaque from computed tomography. *Circulation: Cardiovasc Imaging*, 10(12), 2017.

[250] Giampaolo Martufi, Moritz Lindquist Liljeqvist, Natzi Sakalihasan, Giuseppe Panuccio, Rebecka Hultgren, Joy Roy, and T. Christian Gasser. Local diameter, wall stress, and thrombus thickness influence the local growth of abdominal aortic aneurysms. *J Endovasc Ther*, 23(6):957–966, 2016.

[251] G. García, J. Maiora, A. Tapia, M. Graña, and M. De Blas. Computer-aided diagnosis of abdominal aortic aneurysm after endovascular repair using active learning segmentation and texture analysis. In *XIII Mediterranean Conference on Medical and Biological Engineering and Computing 2013*, pages 186–189, 2014.

[252] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *J Mach Learn Res*, 3:1157–1182, 2003.

[253] T. Fukuda, H. Matsuda, Y. Sanda, Y. Morita, K. Minatoya, J. Kobayashi, and H. Naito. CT Findings of Risk Factors for Persistent Type II Endoleak from Inferior Mesenteric Artery to Determine Indicators of Preoperative IMA Embolization. *Ann Vasc Dis*, 7(3):274–279, 2014.

Curriculum Vitae

Irem Cetin received her B.Sc. degree in Computer Engineering from Izmir University of Economics (Izmir, Turkey) in 2011, and her M.Sc. degree in Computer science from the University of Bonn (Bonn, Germany) in 2016. She then joined the PhD program at the Universitat Pompeu Fabra (Barcelona, Spain) in 2017, under the supervisions of Prof. Oscar Camara and Prof. Miguel Angel Gonzalez Ballester. Her work is focused on the analysis of cardiovascular imaging data with radiomics and artificial intelligence algorithms, with a focus on development of explainable artificial intelligence (XAI) to identify the most relevant and interpretable biomarkers from large multi-scale datasets.

Publications

Journal papers

1. **Cetin, I.**, Camara, O., Gonzalez Ballester, M. A., Attri-VAE: attribute-based, disentangled and interpretable representations of medical images with variational autoencoders. Medical Image Analysis. (2022) [Submitted]

2. **Cetin. I.**, Raisi-Estabragh, Z., Petersen, S.E., Napel, S., Piechnik, S.K., Neubauer S., Gonzalez Ballester, M. A. Camara, O., Lekadir, K. Radiomics signatures of cardiovascular risk factors in cardiac MRI: Results from the UK Biobank. Frontiers in Cardiovascular Medicine, Volume 7. (2020)

3. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P., **Cetin, I.**, Lekadir, K., Camara, O., González Ballester, M.A., Sanromá, G., Napel, S., Petersen, S.E., Tziritas, G., Grinias, E., Khened, M., Kollerathu, V.A., Krishnamurthi, G., Rohé, M., Pennec, X., Sermesant, M., Isensee, F., Jäger, P.F., Maier-Hein, K., Full, P.M., Wolf, I., Engelhardt, S., Baumgartner, C.F., Koch, L.M., Wolterink, J.M., Išgum, I., Jang, Y., Hong, Y., Patravali, J., Jain, S., Humbert, O., Jodoin, P. Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? IEEE Transactions on Medical Imaging, 37, 2514-2525. (2018)

Conference papers and abstracts

1. **Cetin, I.**, Petersen, S.E., Napel, S., Camara, O., Gonzalez Ballester, M. A., Lekadir, K. A Radiomics Approach to Analyze Cardiac Alterations in Hypertension. International Symposium on Biomedical Imaging (ISBI). (2019)

2. **Cetin, I.**, Petersen S.E., Camara O., González Ballester M.A., Lekadir K. Identifying alterations in the cardiac ventricles in atrial fibrillation: a radiomics approach. International Journal of Computer Assisted Radiology and Surgery. (2019)

3. Masias M., **Cetin, I.**, Petersen S.E., González Ballester M.A., Piella G., Lekadir K. Can one predict brain disease based on cardiac imaging data? A proof-of-concept study Cardiovascular Imaging. International Journal of Computer Assisted Radiology and Surgery. (2019)

4. Rodriguez Martin R., del Rio Barquero L., **Cetin, I.**, Ruiz Wills C., González Ballester M.A., Noailly J., Lekadir K. Modelado predictivo de la fractura de fémur a partir de imágenes DXA usando radiomica y técnicas de aprendizaje supervisadas. Revista de Osteoporosis y Metabolismo Mineral. (2018)

5. **Cetin, I.**, Sanroma, G., Petersen, S.E., Napel, S., Camara, O., Gonzalez Ballester, M. A., Lekadir, K. A Radiomics Approach to Computer-Aided Diagnosis with Cardiac Cine-MRI. International Workshop on Statistical Atlases and Computational Models of the Heart. (2017)