Differentiation of Hypertrophic Cardiomyopathy Mutation Carriers without Left Ventricular Hypertrophy and Healthy Controls on CMR using Radiomics

By

Amber Heijdra

to obtain the degree of
Master of Science
in Biomedical Engineering
at the Delft University of Technology,
to be defended publicly on Friday January 21, 2022 at 11:00am.

Student number:	4607163	
Project duration:	February 2021 – January 2022	
Thesis committee:	Dr. F. M. Vos,	TU Delft, chair of thesis committee
	Ir. M. P. A. Starmans,	Erasmus MC Rotterdam, daily supervisor
	Dr. Ir. Rob J. van der Geest,	Leids Universitair Medisch Centrum
	Dr. Ir. S. Klein	Erasmus MC Rotterdam, supervisor
	Dr. A. Hirsch	Erasmus MC Rotterdam, supervisor

TU Delft

Erasmus MC
University Medical Center Rotterdam

Abstract

Hypertrophic cardiomyopathy (HCM) is known as a frequent, genetic cardiovascular disease, often caused by mutations of sarcomere protein genes. HCM is primarily characterized by the presence of an increased left ventricular wall thickness, i.e. left ventricular hypertrophy (LVH). However, the disease appears to be asymptomatic in some patients, which makes it a diagnostic challenge. Mutation carriers of HCM who have not yet developed LVH are called genotype-positive left ventricular hypertrophy-negative (G+/LVH-) patients. The primary aim of this study was to investigate whether a radiomics model is able to distinguish between G+/LVH- patients and healthy controls, based on cardiac magnetic resonance (CMR) images.

In total three datasets are analysed. A development dataset was used to develop different radiomics models and to evaluate the performance of the models. The models were validated on both the prospective validation dataset and external validation dataset. G+/LVH- patients had to be known to carry a class 4 (likely pathogenic) or class 5 (pathogenic) gene mutation for HCM and a maximum left ventricular wall thickness of <13mm. Endocardial and epicardial borders were manually and automatically segmented on long-axis view (2-chamber (2CH), 3-chamber (3CH), and 4-chamber (4CH)) and short-axis (SA) view in both end-diastolic (ED) and end-systolic (ES) phase. From these segmentation 555 features including shape, intensity and texture were extracted. Evaluation of radiomics models was performed through a 100x stratified random-split cross-validation in development dataset. Next, the models were validated on prospective validation dataset and external validation dataset.

The radiomics model with best performance developed on development dataset had a mean area under the receiver operating characteristic curve (AUC) of 0.89. A similar performance in prospective validation was found (mean AUC of 0.89), while a lower performance was found in external validation dataset (mean AUC of 0.63). In addition, the radiomics models performed with automatic segmentation showed all a decrease in performance; mean AUC of 0.75, 0.77 and 0.50 in development dataset, prospective validation dataset and external validation dataset, respectively.

Our radiomics models using CMR images can non-invasively distinguish between +/LVH- patients and healthy controls on both development dataset and prospective validation dataset. However, it was not able to distinguish on external validation dataset.

Keywords: Hypertrophic cardiomyopathy, Cardiovascular magnetic resonance imaging, Radiomics, Machine learning

Contents

1. Introduction

Hypertrophic cardiomyopathy (HCM) is described as a frequent, autosomal dominant inherited cardiovascular disease affecting the heart muscle, i.e., myocardium [1, 2]. Among the different cardiomyopathies, HCM is the most common with a prevalence of 1 in 500 people worldwide [3, 4]. In most cases, HCM is caused by mutations of sarcomere protein genes [2, 5]. These are known as the functional units responsible for the contraction of muscle tissue and can be found with genetic testing [6, 7]. However, genetic testing is not widely available [8]. Moreover, patients may not prefer a genetic test because the result may affect issues such as life insurance and professional opportunities [9].

Another way to determine HCM is by visual assessment of medical images such as cardiac magnetic resonance (CMR). HCM is characterized by various phenotypic manifestations, but mainly by the occurrence of an increased left ventricular wall thickness, i.e. left ventricular hypertrophy (LVH) [3, 10]. Most thickenings are asymmetrical and involve the basal interventricular septum [7, 11]. To establish a clinical diagnosis of HCM, the maximum left ventricle wall thickness (MLVWT) is used as key factor, which is often measured with CMR [12, 13]. The clinical measurement of the MLVWT is preferably performed on the short axis and is highly dependent on the accuracy of the assessment, as the distance between the endocardial and epicardial borders is measured manually [3, 14]. In general, HCM is defined by the presence of a MLVWT of ≥15 mm or ≥13 mm in relatives of HCM patients, measured in one or more myocardial segments [5, 7]. These diagnostic conditions are recommended by The European Society of Cardiology guidelines [9]. When both a sarcomere mutation and LVH are present, we refer to the patient as genotype-positive left ventricular hypertrophy-positive (G+/LVH+) patient. Mutation carriers who have not yet developed LVH are called genotype-positive left ventricular hypertrophy-negative (G+/LVH-) patients [15]. With the late gadolinium enhancement (LGE) CMR sequence, tissue abnormalities, such as myocardial fibrosis, can be assessed using administration of a gadolinium-based contrast agent (GBCA) [16, 17]. Myocardial fibrosis is another common phenotypic manifestation and is used as a prognostic marker for HCM patients, as the presence of myocardial fibrosis indicates an increased risk of sudden cardiac death (SCD) [18,19]. However, since myocardial fibrosis can develop over time, repeated LGE with GBCA administration is often required to monitor this, which involves additional health risks, such as nephrogenic systemic fibrosis [20].

To be more specific, HCM appears to be a disease with a heterogeneous pathological and clinical profile, with a highly variable degree of manifestation [21, 22]. For example, the same mutation may manifest differently between family members and lead to a different clinical outcome [6, 23]. In addition, since HCM remains asymptomatic throughout life in many patients, the disease is often discovered by coincidence in families with HCM [6, 9]. Despite the absence of manifestation in some HCM patients and the fact that many HCM patients have a normal life expectancy, HCM is considered a leading cause of SCD observed in young people and athletes [3, 24]. Moreover, HCM is sometimes difficult to distinguish from an "athlete's heart" or diseases such as hypertension heart disease (HHD) which also involve LVH.

Since the diagnosis of HCM is solely based on observational data from the medical images, diagnosing HCM is highly observer dependent. Moreover, subtle pathologic features are often difficult or impossible to detect with the naked eye, which makes it a diagnostic challenge [25, 26]. Because of the limited phenotypic manifestations and the limitations of qualitative assessment, there is a great need for tools to improve the diagnosis and clinical understanding of HCM [24]. To our knowledge, no studies have yet been done on quantitative assessment in HCM patients in whom the disease has not yet manifested with LVH.

However, previous studies have already shown that quantitative analysis of medical images using radiomics can be used to diagnose HCM in patients with presence of LVH [3, 12, 18, 27, 28, 29]. Radiomic aims to predict clinical labels or outcomes using quantitative medical imaging features, which are obtained from medical examinations such as CMR using machine-learning methods [30, 31]. Moreover, radiomics is based on the assumption that quantitative medical image features are linked to disease-specific processes, such as genetic aberrations, that are difficult to quantify or recognize by the human eye [26, 32]. Baeßler et al. (2018) [3] and Amano et al. (2021) [28] showed that radiomics can differentiate HCM patients with and without presence of myocardial fibrosis from healthy controls on native CMR images. In addition, Neisius et al. (2019) [12] and Schofield et al. (2019) [27] have demonstrated on native CMR images that in addition to distinguishing HCM from healthy controls, HCM patients can also be distinguished from diseases, such as HHD, which manifest similarly. These results showed that, despite similar phenotypic manifestations and the omission of the administration of GBCA, it was possible to distinguish HCM patients from healthy controls and other diseases.

Based on these results, it was hypothesized that radiomics could also be used to distinguish HCM mutation carriers without manifestation of LVH from healthy controls based on native CMR images. The primary aim of this study was to investigate whether a radiomics model based on CMR images is able to distinguish between G+/LVH- patients and healthy controls. Additionally, the generalizability of the results was evaluated by performing validation on independent, unseen data. Finally, with the same radiomics models we compared the performance of manual segmentation to the performance of automatic segmentation, which are in general more consistent and reproducible.

2. Method

2.1 Study population

In this study we use three independent datasets: 1) development dataset; 2) prospective validation dataset; and 3) external validation dataset. The development dataset obtained from Erasmus Medical Center (EMC) (Rotterdam, the Netherlands) was used to develop the radiomics model and to evaluate the performance of the model. To ensure the performance of the applied approach, the model was validated on both the prospective validation dataset and external validation dataset. The prospective validation dataset also consisted of data collected from the EMC, while the external validation dataset included a two-center dataset obtained from the VU University Medical Center (VUmc) (Amsterdam, the Netherlands) and the Radboud university Medical Center (RUMC) (Nijmegen, the Netherlands). An overview of the datasets is shown in Table 1.

Table 1: *Overview of the three different datasets used in this study; development dataset used to develop the model, prospective validation dataset and external validation dataset used to validate the model.*

Dataset	Medical center	Number of G+/LVH- patients	Scanning period	Number of healthy controls	Scanning period
Development	Erasmus Medical Center, Rotterdam, The Netherlands	57	12/2008 - 09/2020	40	06/2018 - 11/2019
Prospective validation	Erasmus Medical Center, Rotterdam, The Netherlands	11	01/2021 - 08/2021	25	06/2018 - 06/2019
External validation	VU University Medical Center, Amsterdam, The Netherlands	18	05/2005 - 10/2011	14	10/2006 - 04/2007
	Radboud University Medical Center, Nijmegen, The Netherlands	9	10/2012 - 07/2013	10	07/2020 - 08/2021

Abbreviations: G+/LVH-, genotype-positive left ventricular hypertrophy-negative.

This study was approved by the institutional review boards of the EMC (MEC-2014-096). For all datasets, informed consent was obtained from both the G+/LVH- patients and healthy controls.

Inclusion criteria for G+/LVH- patients and healthy controls were: presence of a scanned balanced steady-state free precession (bSSFP) CMR in long-axis view (2-chamber (2CH), 3-chamber (3CH), and 4-chamber (4CH)) and short-axis (SA) view to perform the analysis on; presence of complete cardiac cycle to determine end-diastolic (ED) and end-systolic (ES) phase; and a measured MLVWT of <13mm, such that HCM patient has not yet been identified with LVH according to the diagnostic conditions recommended by The European Society of Cardiology guidelines [9]. In addition, G+/LVH- patients had to be known to carry a class 4 (likely pathogenic) or class 5 (pathogenic) gene mutation for HCM, according to the recommendations of the American College of Medical Genetics and Genomics [33]. Control groups had to be unrelated healthy controls without cardiovascular disease. However, healthy controls were not required to be genetically tested for pathogenic DNA variants that cause HCM.

The ED and ES phases were determined on the SA series, ensuring that the same phase was used within each patient. The ED and ES phases were determined by selecting the phase with the visually highest and lowest estimated left ventricular volume, respectively, obtained on the SA view. In other words, the phases in which the bloodpool was depicted largest and smallest were chosen as ED and ES phase, respectively. The phases obtained on the SA view were then selected and chosen as ED and ES phase for the long-axis views as well.

2.1.1 Development dataset

The dataset as described in N. van der Velde et al. (2021) [34], was used as model development dataset. This dataset consists of 57 G+/LVH- patients, with a CMR performed between October 2008 and

September 2020. The control group consisted of a total of 40 unrelated healthy controls, scanned between June 2018 and November 2019. The control group was matched for age and sex on group level to the patient group.

In addition to the CMR images, manually scored morphological features of the development dataset obtained from N. van de Velde et al. (2021) [34] were collected. These included: 1) right-to-left ventricular ratio; 2) presence of ≥2 myocardial crypts; 3) presence of hooked thickening basal anterior wall; 4) maximal wall thickness/posterior wall thickness ratio; and 5) maximal wall thickness/left ventricular mass indexed by body surface area and normalized by sex ratio. An example of an anterobasal hook and a myocardial crypt is depicted in Appendix A (Figure A1).

2.1.2 Prospective validation dataset

The prospective validation dataset consisted of eleven G+/LVH- patients and 25 healthy controls, scanned in the EMC, between January 2021 and August 2021 for G+/LVH- patients and between June 2018 and June 2019 for healthy controls.

2.1.3 External validation dataset

The external validation dataset consisted of a multi-center imaging dataset obtained from two different medical centers with a total of 27 G+/LVH- patients and 24 healthy controls. Of these, eighteen G+/LVH- patients and fourteen healthy controls were collected from VUmc, performed between May 2005 and October 2011, and between October 2006 and April 2007, respectively. The remaining nine G+/LVH- patients and ten healthy controls were collected from RUMC, scanned between October 2012 to July 2013, and between July 2020 and August 2021, respectively.

2.2 Segmentation

2.2.1 Manual segmentation

After collecting the CMR images and defining ED and ES phase, segmentation was performed. Segmentation of endocardial and epicardial borders of the datasets was done manually on both the ED and ES phase slice-by-slice, resulting in a 3D volume for SA view and a 2D volume for 2CH, 3CH and 4CH view. A medical PhD student performed segmentation on the SA view for the developmental dataset; the long-axis view for the development dataset and all views for the two validation datasets were segmented by a biomedical engineering student.

Based on the endocardial and epicardial borders, two segmentations were composed: 1) $LV_{Bloodpool}$: all voxels within the endocardial border; and 2) $LV_{Myocard}$: all voxels between the epicardial and endocardial borders. In addition, $LV_{Bloodpool+Myocard}$ denotes a combination of the two segmentations, on which both analysis was performed. Examples from the obtained segmentations are shown in Figure 1. The segmentations were performed in Medis Suite MR software (Qmass software version 8.1, Medis, Leiden, the Netherlands).

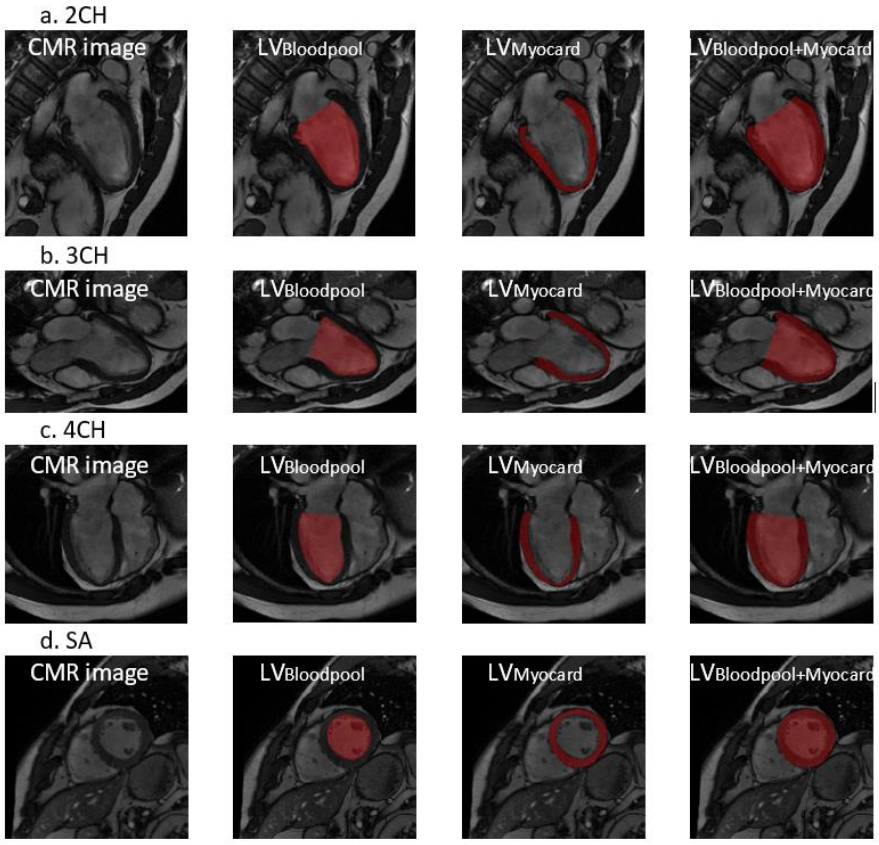

Figure 1: *Examples of the 2D slices of retrieved segmentations with $LV_{Bloodpool}$ and $LV_{Myocard}$ in (a) 2CH, (b) 3CH, (c) 4CH and (d) SA view depicted in red. Abbreviations: CMR, Cardiac Magnetic Resonance; 2CH, 2-chamber; 3CH, 3-chamber; 4CH, 4-chamber; SA, short axis*

2.2.2 Automatic segmentation

In addition to manual segmentation of the datasets, automatic segmentation was performed on 2CH, 3CH, 4CH and SA views, in the Medis Suite MR software. To assess the agreement between manual and automatic segmentation, the pairwise Dice Similarity Coefficient (DSC) was calculated in Python 3.8. The DSC results in a value between 0 and 1, with a value close to 1 indicating high agreement [35, 36]. In addition, manual and automatic segmentations were visually assessed to determine common areas of disagreement.

2.3 Feature extraction

Before feature extraction, pre-processing was performed. Since CMR images does not have a fixed unit and scale, images are normalized using z-scoring, to obtain image intensities with similar scale.

2.3.1 Image features

After delineation of the borders and image pre-processing, imaging features were extracted. By default, for each segmentation, 564 features are extracted using the Workflow for Optimal Radiomics Classification (WORC) toolbox [30]. These features quantify shape, intensity, texture and orientation texture [37, 38]. However, since orientation features depend on the location and orientation of the segmentation, they were omitted to avoid bias. As result, a total of 555 imaging features were extracted in this study, details can be found in Appendix B (Table B1). Feature extraction was performed using the open-source Python packages PyRadiomics [37] and PREDICT [39].

A total of 35 shape features were extracted. The shape features describe morphological characteristics of the segmentation, which include features such as volume, the surface-to-volume ratio, sphericity and compactness. These features are extracted based solely on the segmentations and do not require the underlying image intensity data.

For the intensity features, thirteen features are included. Intensity features are extracted directly from the image, or derived from the intensity histogram. These features consist of various first-order statistics to quantify the raw intensity distribution within the segmentation. The most basic intensity features include features such as the mean, maximum, minimum and standard deviation (std). In addition, features that measure the distribution of the histogram are extracted, such as skewness and kurtosis.

Texture features, also called second-order features, consist of a total of 507 features. These features quantify intensity properties between surrounding pixels. The algorithms that are included are; the Gray Level Co-occurrence Matrix [40], the Grey-Level Run-Length Matrix [40], the Gray Level Size Zone Matrix [40], the Neighbouring Grey Tone Difference Matrix [40], the Gray Level Dependence Matrix [40], the Gabor filter [40], the Laplacian of Gaussian filter [41], Local Binary Patterns [42], local phase features [43, 44] and vessel filter features [41]. These features include more complex patterns such as heterogeneity and are extracted based on the spatial relationship between neighbouring pixels within the segmentation.

2.4 Decision model creation

In this study, a decision model is created within the Workflow for Optimal Radiomics Classification (WORC), i.e., a computational radiomics workflow. An schematic overview of the methodology is depicted in Figure 2.

The decision model creation consists of several steps, including feature selection and machine learning. For example, within these steps, the features with the highest predictive value are selected and with machine learning, patterns in these features are identified that differentiate between G+/LVH patients and healthy controls. Numerous algorithms are involved in these steps. With the WORC toolbox, different algorithms are automatically analysed and automated machine learning is used to determine which combination of algorithms yields the best prediction performance on the development dataset. Since a single best solution may be a coincidence, the resulting radiomics model is an ensemble of the 50 best performing solutions. For more details, refer to M.P.A. Starmans (in press) [30].

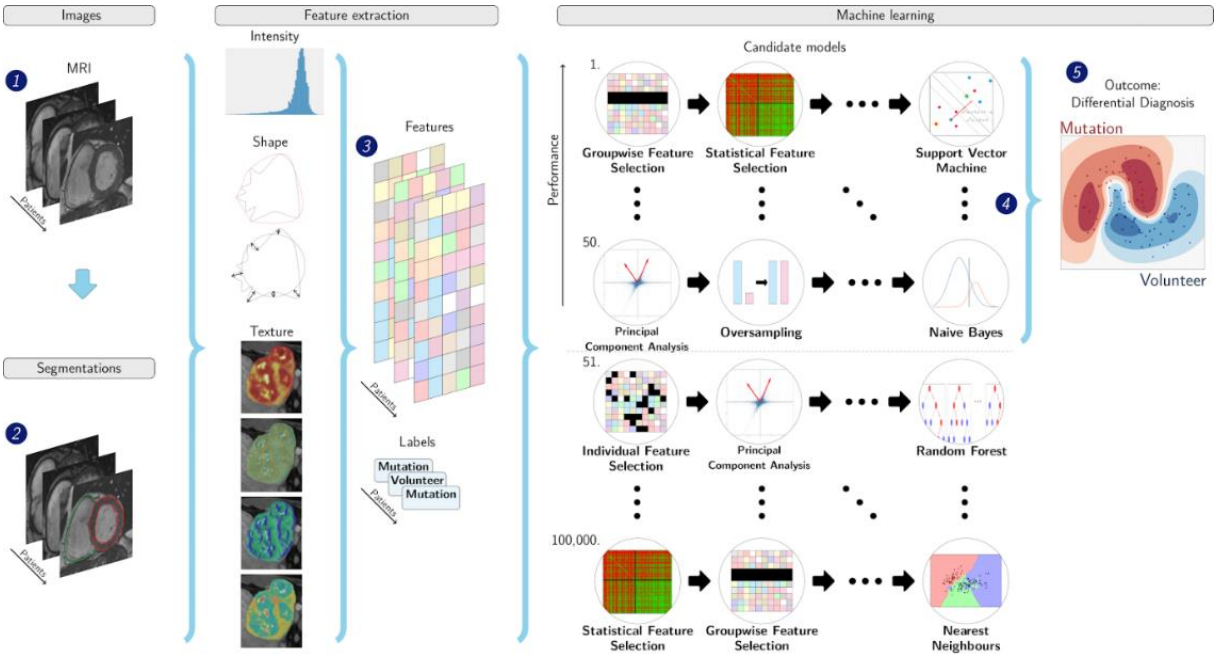

Figure 2: *Schematic overview of the radiomics approach: adapted from M. Vos (2019) [25]. Input to the algorithm are (1) cardiac magnetic resonance images and (2) segmentations by delineation of the endocardial and epicardial borders. The processing steps include (3) feature extraction and (5) the creation of a machine learning decision model, using (4) an ensemble of the best 100 workflows from 1.000 candidate workflows, in which the workflows are different combinations of the different analysis steps (e.g. the classifier used).*

2.5 Experimental setup

First, to assess which segmentations, phases and sequences had the most predictive value, radiomics models were created based on all different combinations of segmentation ($LV_{Bloodpool}$, $LV_{Myocard}$ or $LV_{Bloodpool+Myocard}$), phase (ED or ES), and sequence (SA, 2CH, 3CH or 4CH), totalling 24 models, called baseline models. The baseline models with the highest performance were selected for further evaluation. Second, based on the results of these models, three radiomics models were externally validated: 1) based on the single sequence, view, and phase with the highest performance ; 2) for each view, the segmentation with the highest performance within the phase with highest mean performance; and 3) for each view, in both ED and ES phase, the segmentation with the highest performance. Because the different views are highly dependent on how the CMR is planned and not all morphological features are visible on a single view, model 2 and 3 are expected to extract more information from the imaging features and thus obtain higher performance. Model 2 is included to obtain the added value of the different views within a single phase. Model 3 is included to obtain the added value of the CMR as a totality.

Next, manually scored morphological features obtained from N. van der Velde et al. (2021) [34] were used to create two additional models: 4) based on manually scored features solely, and 5) a combination of manually scored features and the setup used in model 3. To compare the performance of the WORC methodology with the performance of the methodology applied by N. van der Velde et al. (2021) [34], model 4 was performed. Finally, model 5 is evaluated and seen as the most complete model comparable to the clinic, with both CMR image features and manually scored features determining performance.

The predictive value of models 1 through 5 were assessed by training and evaluating the development dataset. Moreover, the models 1 to 3 were evaluated with the prospective and external validation dataset. In addition, these models were also applied with automatic segmentation.

2.5.1 Elimination of biases

Due to the use of data from other medical centers in external validation dataset, the use of different scanning parameters may cause the CMR images to be displayed differently compared to the developmental dataset. As a result, this may affect the predictive value of intensity and texture features. In addition, some features are related to slice thickness, and this may be different in the different datasets. Therefore, three additional models were evaluated with external validation dataset, with broadly adopting the setup of model 1 for simplicity. These additional models were performed to eliminate biases due to different signal intensities resulting from different scan parameters and biases due to different slice thicknesses. The additional models were based on: 1A) shape features extraction only, 1B) manual adjustment of slice thickness in both development and external validation dataset to create same slice thickness, and 1C) shape features extraction performed with the adjusted slice thickness.

In addition, the difference in signal intensity is examined more in detail by focussing on two intensity features. Within the intensity features, hf_mean and hf_std were considered, to obtain a better insight into the histogram distribution within the different datasets. Furthermore, two shape features, sf_Max2DDiameterSlice and sf_area_avg_2D, were also examined to eliminate difference in cardiac size within the two datasets. The median and interquartile range (IQR) of the features for both $LV_{Myocard}$ and $LV_{Bloodpool}$ segmentation were examined.

Finally, given that G+/LVH- patients during standard examinations received administration of GBCA before scanning the SA view, for detection of myocardial fibrosis on the LGE sequence, it is considered that GBCA is present on SA views in G+/LVH- patients. It is assumed that this will affect performance, as healthy controls will not receive unnecessary administration of GBCA. If confirmed in baseline models, SA views are omitted from radiomics models.

2.6 Evaluation

2.6.1 Cross-validation set-up

By default in WORC, the performance of the models on the development dataset were evaluated through a 100x stratified random-split cross-validation. In each iteration, the dataset was randomly split in 80% for training and 20% for testing, as shown in Figure 3A. To eliminate the risk of overfitting on the test set, model optimization was performed within the training set, in which an internal 5x random split cross-validation was performed. The training set was split in 15% for validation to optimize the model hyperparameters and 85% was used for actual training. In the evaluation of the prospective validation and external validation dataset, which both used a fixed, independent training and testing set, only an internal 5x random split cross-validation was performed, see Figure 3B.

The performance of all models on development dataset, prospective validation dataset and external validation dataset were assessed by using the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve, sensitivity, specificity, and accuracy. ROC confidence bands were constructed using fixed-width bands [30, 35]. The G+/LVH- patients were defined as positive class.

For development dataset, the corrected resampled t-test, which takes into account that the cross-validation samples are not statistically independent, were used to construct the 95% confidence intervals (CIs) of the performance metrics [35, 36]. To evaluate whether the developed models on the development dataset generalize well to independent, unseen data, the models were validated on both a prospective validation and external validation dataset. For these datasets, 95% Cis were constructed using 1000x bootstrap resampling [30, 45]. Only models 1, 2, and 3 could be evaluated within the validation datasets, due to the absence of manually scored morphological features in these datasets. Performance on the three different datasets was assessed for both manually and automatically generated segmentations.

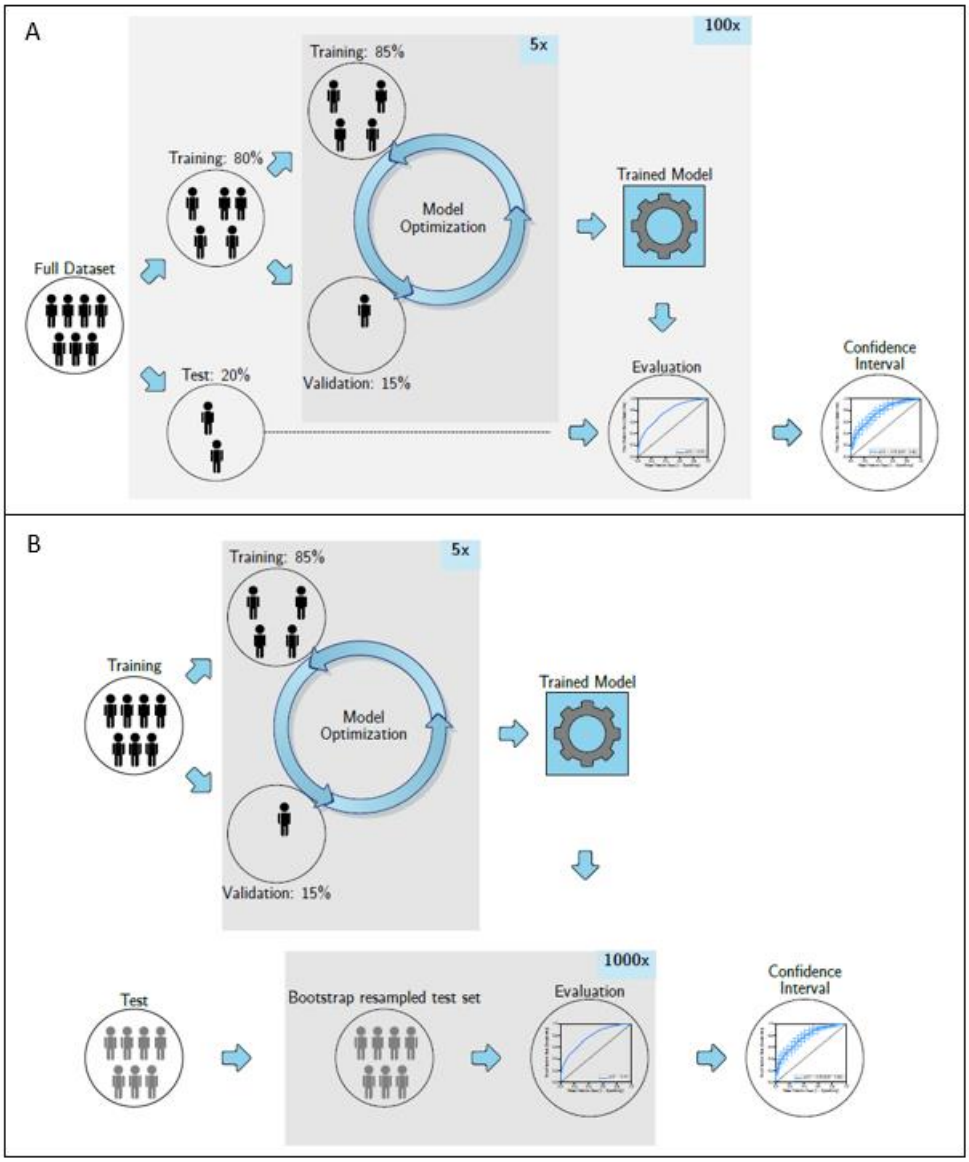

Figure 3: *Schematic overview of the cross-validation setup used by the WORC framework for optimization and evaluation. (A) The 100x random-split cross-validation used in development dataset; (B) and the 1000x bootstrap resampling in prospective validation and external validation dataset. Adapted from M.P.A. Starmans et al. (in press) [30].*

2.6.2 Model insights

The Mann–Whitney U test was used in order to assess the predictive value of the individual features for the different datasets. To correct the p-values for multiple testing, Bonferroni correction was used, i.e., multiplying the p-values by the number of tests. After correction, p-values were considered statistically significant at p-value ≤ 0.05.

Furthermore, to gain insight into the models performed on development dataset, patients were ranked from typical to atypical, based on the consistency of the model predictions. The consistency was determined by the percentage that a patient was classified correctly when occurring in the test set. Typical examples were patients who were always correctly classified, patients who were always classified incorrectly were considered atypical..

3. Results

3.1 Study population

An overview of clinical and imaging characteristics of the three different datasets is summarized in Table 2.

Within these three datasets, a total of 184 CMR scans were performed. Within the development dataset and prospective validation dataset, CMR was performed with General Electric (GE), while the external validation dataset was performed with Siemens. The median age for the development dataset was 46 years (IQR 32–54 years) and 34% of the patients were male. Development dataset and prospective validation dataset, both scanned at EMC, were comparable in image acquisition protocols. For the external validation dataset, consisting of a dataset from two centers, other image acquisition protocols were used but were homogeneous within the dataset.

Table 2: *Clinical characteristics and CMR scan parameters of the three datasets included in this study.*

Dataset	Development dataset		Prospective validation dataset		External validation dataset			
	G+/LVH- patients	Healthy controls	G+/LVH- patients	Healthy controls	G+/LVH- patients (n = 27)		Healthy controls (n = 24)	
					VU	RUMC	VU	RUMC
Number of subjects	57	40	11	25	18	9	14	10
Age (years) *	46 [32-53]	45 [33-55]	N/A	26 [28 - 30]	N/A	N/A	N/A	N/A
Male sex (%)	18 (32%)	15 (38%)	N/A	19 (76%)	5 (28%)	4 (44%)	9 (64%)	3 (30%)
Length (cm) *	171 [165-182]	174 [169-181]	N/A	185 [180 - 189]	N/A	N/A	N/A	N/A
Weight (kg) *	71 [62-85]	70 [65-79]	N/A	84 [74 - 85]	N/A	N/A	N/A	N/A
Magnetic field strength								
1.5 Tesla	51	40	11	25	18	9	14	10
3.0 Tesla	6							
Manufacturer								
Siemens					18	9	14	10
GE	57	40	11	25				
Slice thickness(mm)	8	8	8	8	5	6	5	5
Pixel spacing (mm) *	0.70 [0.70-0.70]	0.70 [0.70-0.70]	0.70 [0.70-0.70]	0.70 [0.70-0.70]	1.33 [1.33-1.33]	1.45 [1.45-1.45]	1.33 [1.33-1.39]	1.37 [1.37-1.37]
Repetition time(ms) *	3.78 [3.65-3.78]	3.78 [3.77-3.78]	3.78 [3.75-3.78]	3.78 [3.75-3.78]	47.10 [47.10-47.10]	42.00 [42.00-42.00]	47.10 [46.20-47.10]	44.55 [44.55-44.55]
Echo time (ms) *	1.69 [1.64-1.70]	1.69 [1.68-1.70]	1.69 [1.68-1.70]	1.69 [1.68-1.70]	1.57 [1.57-1.57]	1.18 [1.18-1.18]	1.57 [1.55-1.57]	1.25 [1.25-1.25]
Flip angle *	65 [65-65]	65 [65-65]	65 [65-65]	65 [65-65]	60 [60-60]	80 [79-80]	60 [60-60]	78 [78-78]

Abbreviations: G+/LVH-, genotype-positive left ventricular hypertrophy-negative.

** Values are given in median (interquartile range). Other values than those given in the median and interquartile range do occur.*

3.2 Development of radiomics models

Based on the AUC values of the 24 baseline models, the models for further evaluation were established. The performance of these 24 baseline models can be found in Appendix C (Table C1). As predicted, the SA image showed extremely high predictive values (mean AUC between 0.99 and 1.00), which can be attributed to the presence of GBCA on the CMR images. As a result, SA views were omitted from the radiomics models for validation in this study. Overall, the 2CH, 3CH and 4CH views all showed high performance with mean AUCs between 0.77 and 0.86 for ES phase, while ED phase showed lower performance with mean AUCs between 0.65 and 0.81. The same was apparent for other performance metrics in ED and ES phase; a mean accuracy between 0.59 and 0.75 vs. 0.69 and 0.79, a mean sensitivity between 0.57 and 0.81 vs. 0.68 and 0.84, and a mean specificity between 0.59 and 0.77 vs. 0.65 and 0.78, respectively. This is noteworthy because it shows that a better predictive value is found in the ES phase, whereas for the determination of certain phenotypic manifestations the ED phase is preferentially used in clinical practice [46, 47]. In addition, within the three views (2CH, 3CH and 4CH) and the two phases the highest performance was obtained in $LV_{BLoodpool+Myocard}$ segmentation in four out of six setups. This shows that features extracted from both segmentations have predictive value and that the predictive value generally increases when they are combined.

After selecting the baseline models with the highest performance, the following models were composed: model 1) $LV_{Bloodpool+Myocard}$ in 2CH view in ES phase; model 2) $LV_{Bloodpool+Myocard}$ in 2CH, 3CH and 4CH view in ES phase; model 3) $LV_{Bloodpool+Myocard}$ in 2CH, 3CH and 4CH view in ES phase, and $LV_{Myocard}$ in 2CH view, $LV_{Bloodpool}$ in 3CH view and $LV_{Bloodpool+Myocard}$ in 4CH view all three in ED phase; model 4) manually scored features from N. van der Velde et al. (2021) [34], and model 5) a combination of manually scored features and setup used for model 3.

3.3 Evaluation of radiomics models

The results of the radiomics models performed with manual segmentation for development dataset, prospective validation dataset and external validation dataset are depicted in Table 3. The obtained ROC curves are shown in Figure 4. In addition, the performance values and ROC curves obtained from radiomics models performed with automatic segmentation are given in Appendix D (Table D1 and Figure D1). Lastly, the performance values and ROC curves obtained from the additional models to eliminate biases in external validation dataset are found in Appendix E (Table E1 and Figure E1)

Table 3: *Performance values of the three datasets performed with manual segmentation based on: model 1: 2CH view $LV_{Bloodpool+Myocad}$ in ES phase; model 2: 2CH, 3CH and 4CH view in $LV_{Bloodpool+Myocad}$ ES phase; model 3: 2CH, 3CH and 4CH view in $LV_{Bloodpool+Myocad}$ ES phase and 2CH view $LV_{Myocard}$, 3CH view $LV_{Bloodpool}$ and 4CH view $LV_{Bloodpool+Myocard}$ in ED phase; model 4: manually scored features from N. van der Velde et al. [34]; and model 5 a combination of manually scored features and setup used for model 3.*

	Development dataset				
	Model 1	Model 2	Model 3	Model 4	Model 5
AUC	0.86 [0.78, 0.94]	0.89 [0.81, 0.97]	0.88 [0.81, 0.96]	0.86 [0.77, 0.94]	0.89 [0.81, 0.96]
Accuracy	0.79 [0.70, 0.88]	0.79 [0.69, 0.89]	0.77 [0.66, 0.87]	0.75 [0.65, 0.84]	0.78 [0.69, 0.87]
Sensiticity	0.79 [0.66, 0.93]	0.79 [0.65, 0.93]	0.74 [0.58, 0.90]	0.72 [0.59, 0.85]	0.78 [0.64, 0.91]
Specificity	0.78 [0.64, 0.93]	0.79 [0.65, 0.93]	0.80 [0.67, 0.93]	0.79 [0.64, 0.93]	0.79 [0.66, 0.92]
	Prospective validation dataset				
	Model 1	Model 2	Model 3		
AUC	0.83 [0.69, 0.98]	0.89 [0.74, 1.03]	0.85 [0.66, 1.03]		
Accuracy	0.78 [0.64, 0.92]	0.83 [0.71, 0.96]	0.86 [0.75, 0.98]		
Sensiticity	0.64 [0.34, 0.93]	0.55 [0.24, 0.85]	0.55 [0.23, 0.86]		
Specificity	0.84 [0.70, 0.98]	0.96 [0.88, 1.04]	1.00 [-, -]		
	External validation dataset				
	Model 1	Model 2	Model 3		
AUC	0.58 [0.42, 0.75]	0.63 [0.47, 0.79]	0.65 [0.48, 0.81]		
Accuracy	0.53 [0.39, 0.67]	0.53 [0.39, 0.67]	0.53 [0.39, 0.67]		
Sensiticity	1.00 [-, -]	1.00 [-, -]	1.00 [-, -]		
Specificity	0.00 [-, -]	0.00 [-, -]	0.00 [-, -]		

Abbreviations: AUC, Area Under the Curve

** Outcomes are presented with the 95% confidence interval.*

3.3.1 Evaluation of radiomics models on development dataset

Model 1, based on one long-axis view in one phase, resulted in a mean AUC of 0.86 (95% CI: 0.78-0.94), mean accuracy of 0.79 (95% CI: 0.70-0.88), mean sensitivity of 0.79 (95% CI: 0.66-0.93) and mean specificity of 0.78 (95% CI: 0.64-0.93). Model 2, based on the three long-axis views in one phase, had a slightly higher mean AUC of 0.89 (95% CI: 0.81-0.97). Moreover, no improvement was detected in accuracy, sensitivity and specificity. Model 3, based on three long-axis views in both ED and ES phase, had similar performance with an AUC of 0.88 (95% CI: 0.81, 0.96). The manually scored features in

model 4 resulted in an AUC of 0.86 (95% CI: 0.77-0.94), where the obtained AUC with the radiomics method was lower compared to the methodology described in N. van der Velde et al. (2021) [34]-(mean AUC of 0.92 (95% CI: 0.87-0.97)). Finally, model 5, combining the manually scored features and the imaging features extracted from both phases in all three long-axis views, did not further improve the performance (AUC of 0.89 (95% CI: 0.81, 0.96)).

3.3.2 Evaluation of radiomics models on prospective and external validation dataset

Model 1, based on one long-axis view in one phase, resulted in a mean AUC of 0.83 (95% CI: 0.69-0.98) for prospective validation dataset and an AUC of 0.58 (95% CI: 0.42-0.75) for external validation dataset was obtained. Model 2, based on the three long-axis views in one phase, had a slightly higher mean AUC as well for both validation datasets with an AUC of 0.89 (95% CI: 0.74-1.03) and 0.63 (95% CI: 0.47-0.79), respectively. Finally, model 3, based on three long-axis views in both ED and ES phase, had again a similar performance to model 2 with an AUC of 0.85 (95% CI: 0.66-1.03) for the prospective validation dataset and an AUC of 0.65 (95% CI: 0.48-0.81) for the external validation dataset.

The three models evaluated with the prospective validation and external validation dataset were found to have varying performances. Overall, the performance of the prospective validation dataset was similar to the development dataset, while the external validation dataset was slightly above random guessing.

The ROC curves for the models showed some sharp bends, while the curves for the development dataset were smoother. All models on the external validation dataset had a sensitivity of 1.00 and a specificity of 0.00. This means that each model within this dataset correctly classifies all G+/LVH- patients, while all healthy controls are incorrectly classified as G+/LVH- patients. In addition, prospective validation dataset showed that with multiple views, sensitivity decreased from 0.64 to 0.55, while specificity increased from 0.84 to 0.96 obtained from model 1 and model 2, respectively.

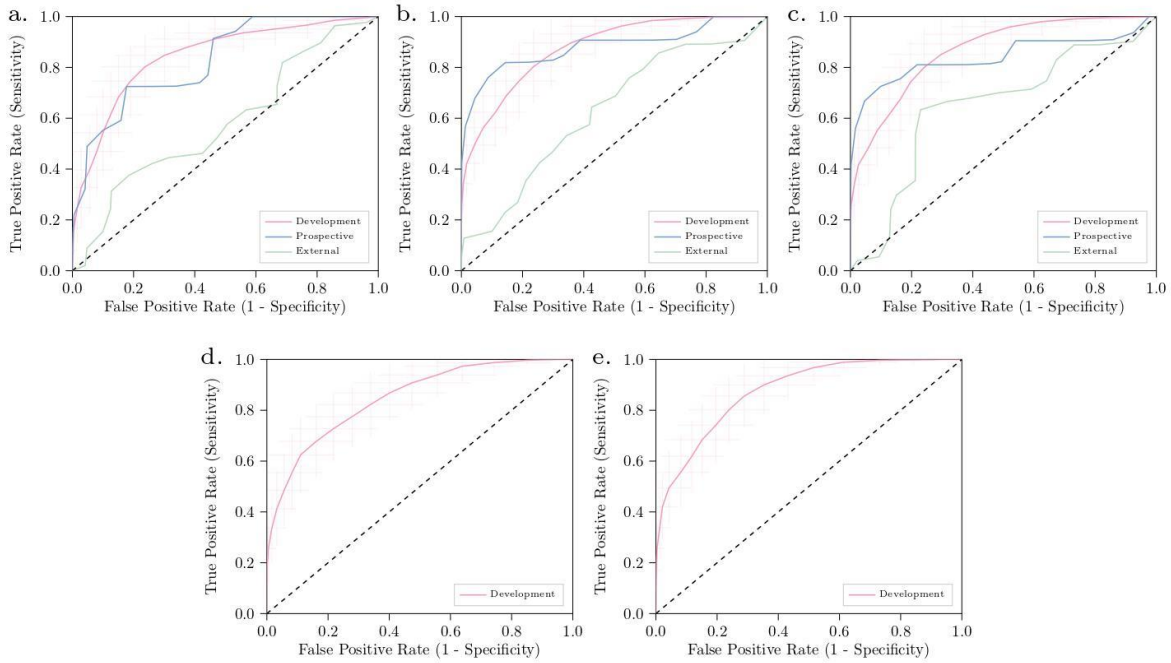

Figure 4: *ROC curves of manual segmentation present the model on development dataset (red), prospective validation dataset (blue) and external validation dataset (green). The ROC curves show: a) model 1: 2CH view $LV_{Bloodpool+Myocad}$ in ES phase; b) model 2: 2CH, 3CH and 4CH view in $LV_{Bloodpool+Myocad}$ ES phase; c) model 3: 2CH, 3CH and 4CH view in $LV_{Bloodpool+Myocad}$ ES phase and 2CH view $LV_{Myocard}$, 3CH view $LV_{Bloodpool}$ and 4CH view $LV_{Bloodpool+Myocard}$ in ED phase; d) model 4: manually scored features from N. van der Velde et al. [34]; and e) model 5 a combination of manually scored features and setup used for model 3.*

3.3.3 Evaluation of radiomics models with automatic segmentation

The obtained mean ± standard deviation of the DSC, given in Appendix D (Table D2), indicated high agreement with an overall DSC of 0.93±0.05, 0.93±0.05 and 0.92±0.04 for the development, prospective validation and external validation dataset, respectively. Visual inspection showed that differences between manual and automatic segmentations were mainly located at the ends of basal wall, shown in Appendix D (Figure D2). In addition, the delineation of the automatic segmentation was smoother, which resulted in less prominent myocardial recesses in the segmentation.

Despite high agreement, performance was generally lower compared to manual segmentation. For example, an AUC of 0.73 (95% CI: 0.62-0.85) was found with automatic segmentation in model 1 while an AUC of 0.86 (95% CI: 0.78-0.94) was obtained with manual segmentation.

3.3.4 Evaluation of biases on external validation dataset

Model 1A, based on shape feature extraction solely, was evaluated by the external validation dataset and showed an even slightly lower performance with an AUC of 0.46 (95% CI: 0.29-0.62) than the original model 1. Model 1B, based on a manually adjusted slice thickness, in both training and evaluating, resulted in a slightly higher performance with an AUC of 0.63 (95% CI: 0.47-0.78). Finally, model 1C, based on a manually adjusted slice thickness and shape feature extraction solely, resulted in a similar performance as model 1A, with a AUC of 0.47 (95% CI: 0.31-0.63).

In addition, similarly, the sensitivity and specificity were found to be 1.00 and 0.00 for both model 1A and 1B, respectively. However, in model 1C, a slight decrease in sensitivity and a slight increase in specificity was found resulting in mean sensitivity of 0.93 and mean specificity of 0.08.

Finally, the obtained values of the two shape features (sf_Max2DDiameterSlice and sf_area_avg_2D) and two intensity features (hf_mean and hf_std) are shown in Table 4. Within the shape features, the IQR of the healthy controls from the external validation dataset were more similar to the IQR of the G+/LVH- patients from the development dataset than those of the healthy controls. Overall, the IQR of the two shape features in healthy controls within the development dataset is higher than in the other subjects. Within the different datasets, the two intensity features were found within the same range, with the largest different visible in hf_mean value of $LV_{Bloodpool}$ in G+/LVH- patients (IQR of 0.82-1.25 in development dataset and IQR of 0.99-1.70 in external validation dataset).

Dataset	Subjects	$LV_{Myocard}$		$LV_{Bloodpool}$	
		sf_Maximum2DDiameterSlice	sf_area_avg_2D	sf_Maximum2DDiameterSlice	sf_area_avg_2D
Development	G+/LVH-	81.23 [77.46, 88.87]	3229.22 [2948.18, 3564.82]	71.58 [68.08, 78.91]	1913.37 [1677.02, 2236.66]
	Healthy control	87.81 [83.46, 92.32]	3526.62 [3206.35, 4107.00]	81.93 [76.38, 86.63]	2205.82 [1969.86, 2618.63]
External Validation	G+/LVH-	83.25 [77.70, 86.61]	3205.04 [2994.69, 3508.10]	73.64 [65.96, 76.71]	2032.91 [1672.49, 2147.89]
	Healthy control	86.51 [78.23, 91.95]	3356.01 [2905.21, 3865.83]	75.46 [66.03, 79.77]	1952.12 [1651.35, 2373.26]
		hf_mean	hf_std	hf_mean	hf_std
Development	G+/LVH-	-0.11 [-0.26, 0.01]	0.22 [0.19, 0.26]	1.07 [0.82, 1.25]	0.50 [0.46, 0.58]
	Healthy control	-0.10 [-0.20, -0.05]	0.23 [0.22, 0.25]	1.10 [1.03, 1.20]	0.53 [0.48, 0.58]
External Validation	G+/LVH-	-0.13 [-0.21, 0.01]	0.22 [0.19, 0.25]	1.23 [0.99, 1.70]	0.68 [0.63, 0.80]
	Healthy control	-0.25 [-0.33, -0.19]	0.21 [0.19, 0.26]	0.98 [0.72, 1.30]	0.66 [0.53, 0.74]

Table 4: *Values of two shape (sf_Max2DDiameterSlice and sf_area_avg_2D) and two intensity (hf_mean and hf_std) features in both $LV_{Myocard}$ and $LV_{Bloodpool}$ segmentation obtained from development dataset and external validation dataset.*

Abbreviations: G+/LVH-, genotype-positive left ventricular hypertrophy-negative.

** Values are median [interquartile range].*

3.4 Model analysis

As the models within datasets showed a similar performance, the model insight analysis was only conducted for model 1. After Bonferroni correction, a total of 39 features showed statistically significant differences in model 1 performed with manual segmentation, with p-values from 7.00×10^{-5} to 4.11×10^{-2}. These included eight shape feature, one intensity feature and 30 texture features. In addition, a total of 3 features showed statistically significant differences in model 1 performed with automatic segmentation, with p-values from 8.56×10^{-3} to 2.93×10^{-2}. These included two shape feature and one texture feature. The features with p-values ≤0.05 of the Mann-Whitney U test obtained from manual and automatic segmentation are shown in Appendix F (Table F1 and Table F2). The majority of the features that were considered statistically significant consisted of features extracted from the $LV_{Bloodpool}$ segmentation, with 34 out of 39 features from manual segmentation and two out of three from automatic segmentation.

Of the 97 subjects in the development dataset with manual segmentation, 47 were always classified correctly, i.e. in all 100 cross-validation iterations. In contrast, eight subjects were always classified incorrectly, including four G+/LVH- patients and four healthy controls. Examples of the CMR images from G+/LVH- patients and healthy controls that were labelled as typical or atypical are shown in Figure 5. In addition, with automatic segmentation 33 out of 97 subjects were always classified correctly and twelve were always classified incorrectly. The subjects who were always incorrectly classified were visually checked for possible segmentation errors or unusual features. Based on visual inspection of the CMR images, there was no clear relationship visible. In addition, no clear relationship was found between atypical and typical subjects in terms of scan parameters or clinical characteristics.

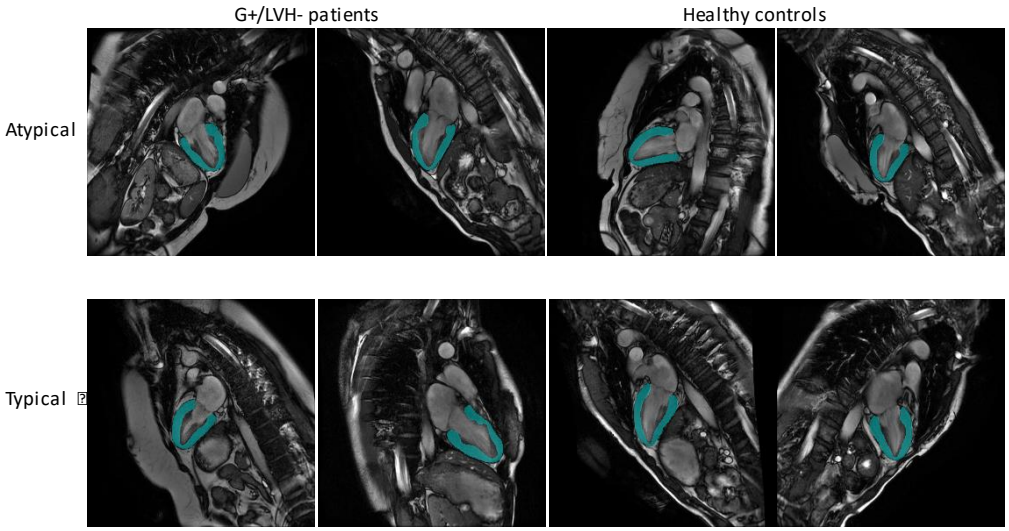

Figure 5: *Examples of typical and atypical subjects within development dataset. Abbreviations: G+/LVH-, genotype-positive left ventricular hypertrophy-negative.*

4. Discussion

The primary aim of this study was to investigate whether radiomics is able to distinguish between G+/LVH- patients and healthy controls based on native CMR images. In this study, we showed that our radiomics models can distinguish G+/LVH- patients from healthy controls, in an internal cross-validation and in prospective validation dataset.

Previous studies have already evaluated radiomics for the differentiation HCM patients and healthy controls. These studies showed promising results [3, 12, 27, 28]. However, these studies did not involve external validation, leaving a gap in the reproducibility of the results. In addition, these studies included HCM patients with presence of LVH, while our study was focusing on HCM patients in whom the disease did not yet manifested with a thickened left ventricle.

The models trained and evaluated on development dataset and models validated on prospective validation dataset, performed all above random guessing (0.50). The best performance was found in model 2, using 2CH, 3CH and 4CH view in ES phase. However, model 1, using a single view, showed similar performance. As demonstrated, radiomics managed to achieve good performance in both the development dataset and the prospective validation dataset with a single view, while adding additional views did not improve performance. Adding additional views might even lead to too much noise, resulting in overfitting. With model insight analysis, it was excluded that the subjects classified as atypical/typical were related to scan parameters or patient characteristics. Moreover, no deviations were apparent using visual inspection. In addition, model insight analysis also revealed that in development dataset the majority of features that showed a statistically significant differences were derived from the $LV_{Bloodpool}$ segmentation. These results suggest that a larger difference in structure and thus more tissue abnormalities are found in $LV_{Bloodpool}$ compared to $LV_{Myocard}$. Further, for prospective validation dataset none of the features showed statistically significant differences. This may be due to the lack of power for these radiomics models.

In addition to the high performance obtained for development dataset and prospective validation dataset, the external validation dataset in contrast showed a low performance. The external validation dataset was not able to distinguish between G+/LVH- patients and healthy controls, with performance slightly above random guessing. It was assumed that this is the result of training models on a homogeneous dataset, while external validation dataset was performed with different scan parameters. Since different scan parameters were used in development dataset and external validation dataset, it may be assumed that this will have an effect on the histogram distribution and thus extraction of intensity and texture features. To eliminate this bias, a model was performed from which only shape features were extracted. This model showed that the performance in external validation dataset remained low. As a result, it was suggested that the cardiac shape in general was different in external validation dataset. Moreover, examination of two shape features showed that there was a slight difference in maximum diameter and area. These values demonstrated that the median and IQR of the shape features obtained from healthy controls in developments dataset differed somewhat from the other subjects. This may be the result of a difference in CMR planning or a difference in study population. Further research should reveal whether these differences are the reason for low performance. The two intensity features, mean and std, were used to assess information about signal intensities. The values of the two intensity features showed a minimal difference. Nevertheless, the histogram distributions can still be very different within the different datasets. Therefore, this is not sufficient to completely eliminate the effect of variation in scan acquisition protocols on performance. Finally, it was demonstrated with the additional model in which slice thickness was manually adjusted, that the different slice thickness between the different datasets

had a minimal negative impact on performance. As a result, the slice thickness can be eliminated as a bias.

To compare the radiomics models with more consistent and reproducible segmentations, the performance is compared to radiomics models performed with automatic segmentation. The different segmentation methods showed high agreement using DSC. Despite the high agreement, a generally lower performance was found. Visual inspection revealed a minimal difference in segmentation, with difference mainly visible at the ends of the basal wall. In addition, the delineation was generally tighter around endocardial and epicardial borders compared to manual segmentation. Despite this minimal difference, due to the tighter delineation, information from pixels/shape of crypts and anterobasal hooks may be missed and therefore result in lower performance. It would be of great benefit to clinical practice if equally high performance could be achieved with automatic segmentation, as there is no dependence on inter- and intra-observer variation and it is less time consuming. It is hypothesized that the application of semi-automatic segmentation, another less time consuming application, would be a good option, where delineation around crypts/hooks would have to adjusted manually.

Our study has several limitations. First, our radiomics models were developed and evaluated on a homogeneous dataset, which reduces the chance of reproducing the obtained performance in a clinical setting. This could therefore be reflected in the drop in performance on the external validation dataset where other MRI scanners were used, resulting in a lack of generalizability. Second, due to bias in SA views by contrast agents, these views were not included in the radiomics models. Whereas, this was often the view in which other studies preferably measured the MLVWT or examined the predictive value of radiomics for HCM patients [46, 47, 48]. Third, 3D features could not be extracted, since only 2D views were used. However, these 3D features may provide additional predictive value. In addition, orientation-dependent features were included. These are dependent on orientation of segmentation and may have a negative effect on the predictive value. To exclude these biases it is recommended not to include features that are orientation dependent in the future. Another limitation is that healthy controls were not required to be tested for presence of pathogenic DNA variants. However, due to the low prevalence its impact may be neglected. Finally, this study demonstrates how strongly the performance of radiomics models is affected by segmentation. Despite the high DSC, a decrease in performance was observed in models performed with automatic segmentation. This suggests that radiomics actually demonstrates that there are important differences between the two segmentations that affect performance.

Future research, in addition to the previously mentioned points, should include training the radiomics model on multi-center data, including different scan parameters, field strengths and manufacturers, to obtain better generalizability of the results. In addition, it would also be interesting to see if radiomics can distinguish between the different mutation types in G+/LVH- patients. Considering that Wang et al. (2020) [5] demonstrated that this is possible in patients with G+/LVH+, it is assumed that it should also be possible with G+/LVH- patients. Finally, the ultimate goal for future research would be the prediction of manifestation using radiomics, including predictive value for SCD. With the aim of gaining a better clinical understanding of HCM and providing better treatment.

5. Conclusion

In conclusion, our radiomics model based on native CMR images was able to distinguish between G+/LVH- patients and healthy controls, in an internal cross-validation and in prospective validation dataset. The model was not able to differentiate G+/LVH- patients in external validation dataset. Further research is needed to improve the generalizability of results.

Bibliography

1 Amano, Y., Suzuki, Y., Yanagisawa, F., Omori, Y., & Matsumoto, N. (2018). Relationship between Extension or Texture Features of Late Gadolinium Enhancement and Ventricular Tachyarrhythmias in Hypertrophic Cardiomyopathy. BioMed Research International, 2018, 1–6. https://doi.org/10.1155/2018/4092469

2 Zhou, H., Li, L., Liu, Z., Zhao, K., Chen, X., Lu, M., Yin, G., Song, L., Zhao, S., Zheng, H., & Tian, J. (2020). Deep learning algorithm to improve hypertrophic cardiomyopathy mutation prediction using cardiac cine images. European Radiology, 31(6), 3931–3940. https://doi.org/10.1007/s00330-020-07454-9

3 Baeßler, B., Mannil, M., Maintz, D., Alkadhi, H., & Manka, R. (2018). Texture analysis and machine learning of non-contrast T1-weighted MR images in patients with hypertrophic cardiomyopathy—Preliminary results. European Journal of Radiology, 102, 61–67. https://doi.org/10.1016/j.ejrad.2018.03.013

4 Kochav, S. M., Raita, Y., Fifer, M. A., Takayama, H., Ginns, J., Maurer, M. S., Reilly, M. P., Hasegawa, K., & Shimada, Y. J. (2021). Predicting the development of adverse cardiac events in patients with hypertrophic cardiomyopathy using machine learning. International Journal of Cardiology, 327, 117–124. https://doi.org/10.1016/j.ijcard.2020.11.003

5 Wang, J., Yang, F., Liu, W., Sun, J., Han, Y., Li, D., Gkoutos, G. V., Zhu, Y., & Chen, Y. (2020). Radiomic Analysis of Native T 1 Mapping Images Discriminates Between MYH7 and MYBPC3 -Related Hypertrophic Cardiomyopathy. Journal of Magnetic Resonance Imaging, 52(6), 1714–1721. https://doi.org/10.1002/jmri.27209

6 Sabater-Molina, M., Pérez-Sánchez, I., Hernández Del Rincón, J., & Gimeno, J. (2017). Genetics of hypertrophic cardiomyopathy: A review of current state. Clinical Genetics, 93(1), 3–14. https://doi.org/10.1111/cge.13027

7 Marian, A. J., & Braunwald, E. (2017). Hypertrophic Cardiomyopathy. Circulation Research, 121(7), 749–770. https://doi.org/10.1161/circresaha.117.311059

8 Yu, F., Huang, H., Yu, Q., Ma, Y., Zhang, Q., & Zhang, B. (2021). Artificial intelligence-based myocardial texture analysis in etiological differentiation of left ventricular hypertrophy. Annals of Translational Medicine, 9(2), 108. https://doi.org/10.21037/atm-20-4891

9 2014 ESC Guidelines on diagnosis and management of hypertrophic cardiomyopathy. (2014). European Heart Journal, 35(39), 2733–2779. https://doi.org/10.1093/eurheartj/ehu284

10 Shi, R. Y., Wu, R., An, D. A., Chen, B. H., Wu, C. W., Du, L., Jiang, M., Xu, J. R., & Wu, L. M. (2021). Texture analysis applied in T1 maps and extracellular volume obtained using cardiac MRI in the diagnosis of hypertrophic cardiomyopathy and hypertensive heart disease compared with normal controls. Clinical Radiology, 76(3), 236.e9-236.e19. https://doi.org/10.1016/j.crad.2020.11.001

11 Hensley, N., Dietrich, J., Nyhan, D., Mitter, N., Yee, M. S., & Brady, M. (2015). Hypertrophic Cardiomyopathy. Anesthesia & Analgesia, 120(3), 554–569. https://doi.org/10.1213/ane.0000000000000538

12 Neisius, U., El-Rewaidy, H., Nakamori, S., Rodriguez, J., Manning, W. J., & Nezafat, R. (2019). Radiomic Analysis of Myocardial Native T1 Imaging Discriminates Between Hypertensive Heart Disease and Hypertrophic Cardiomyopathy. JACC: Cardiovascular Imaging, 12(10), 1946–1954. https://doi.org/10.1016/j.jcmg.2018.11.024

13 Augusto, J. B., Davies, R. H., Bhuva, A. N., Knott, K. D., Seraphim, A., Alfarih, M., Lau, C., Hughes, R. K., Lopes, L. R., Shiwani, H., Treibel, T. A., Gerber, B. L., Hamilton-Craig, C., Ntusi, N. A. B., Pontone, G., Desai, M. Y., Greenwood, J. P., Swoboda, P. P., Captur, G., . . . Moon, J. C. (2021). Diagnosis and risk stratification in hypertrophic cardiomyopathy using machine learning wall thickness measurement: a comparison with human test-retest performance. The Lancet Digital Health, 3(1), e20–e28. https://doi.org/10.1016/s2589-7500(20)30267-3

14 You, Y., Viktorovich, L. A., Qiu, J., Nikolaevich, K. A., & Vladimirovich, B. Y. (2021). Cardiac magnetic resonance image diagnosis of hypertrophic obstructive cardiomyopathy based on a double-branch neural network. Computer Methods and Programs in Biomedicine, 200, 105889. https://doi.org/10.1016/j.cmpb.2020.105889

15 Soler, R., Méndez, C., Rodríguez, E., Barriales, R., Ochoa, J. P., & Monserrat, L. (2018). Phenotypes of hypertrophic cardiomyopathy. An illustrative review of MRI findings. Insights into Imaging, 9(6), 1007–1020. https://doi.org/10.1007/s13244-018-0656-8

16 Neisius, U., El-Rewaidy, H., Kucukseymen, S., Tsao, C. W., Mancio, J., Nakamori, S., Manning, W. J., & Nezafat, R. (2020). Texture signatures of native myocardial T 1 as novel imaging markers for identification of hypertrophic cardiomyopathy patients without scar. Journal of Magnetic Resonance Imaging, 52(3), 906–919. https://doi.org/10.1002/jmri.27048

17 Jiang, B., Guo, N., Ge, Y., Zhang, L., Oudkerk, M., & Xie, X. (2020). Development and application of artificial intelligence in cardiac imaging. The British Journal of Radiology, 93(1113), 20190812. https://doi.org/10.1259/bjr.20190812

18 Mancio, J., Pashakhanloo, F., El-Rewaidy, H., Jang, J., Joshi, G., Csecs, I., Ngo, L., Rowin, E., Manning, W., Maron, M., & Nezafat, R. (2021). Machine learning phenotyping of scarred myocardium from cine in hypertrophic cardiomyopathy. European Heart Journal - Cardiovascular Imaging. https://doi.org/10.1093/ehjci/jeab056

19 Maron, M. S. (2012). Clinical Utility of Cardiovascular Magnetic Resonance in Hypertrophic Cardiomyopathy. Journal of Cardiovascular Magnetic Resonance, 14(1). https://doi.org/10.1186/1532-429x-14-13

20 Zhou, H., An, D., Ni, Z., Xu, J., Fang, W., Lu, R., Ying, L., Huang, J., Yao, Q., Li, D., Chen, B., Shen, J., Jin, H., Wei, Y., Hu, J., Fahmy, L. M., Wesemann, L., Qi, S., Wu, L., & Mou, S. (2021). Texture Analysis of Native T1 Images as a Novel Method for Noninvasive Assessment of Uremic Cardiomyopathy. Journal of Magnetic Resonance Imaging, 54(1), 290–300. https://doi.org/10.1002/jmri.27529

21 Limongelli, G., Pacileo, G., Cerrato, F., Verrengia, M., Di Simone, A., Severino, S., Sarubbi, B., & Calabrò, R. (2003). Myocardial ultrasound tissue characterization in patients with hypertrophic cardiomyopathy: noninvasive evidence of electrical and textural substrate for ventricular arrhythmias. Journal of the American Society of Echocardiography, 16(8), 803–807. https://doi.org/10.1067/s0894-7317(03)00213-x

22 Cheng, S., Fang, M., Cui, C., Chen, X., Yin, G., Prasad, S. K., Dong, D., Tian, J., & Zhao, S. (2018). LGE-CMR-derived texture features reflect poor prognosis in hypertrophic cardiomyopathy patients with systolic dysfunction: preliminary results. European Radiology, 28(11), 4615–4624. https://doi.org/10.1007/s00330-018-5391-5

23 Gao, J., Collyer, J., Wang, M., & Xu, F. (2020). Genetic Dissection of Hypertrophic Cardiomyopathy with Myocardial RNA-Seq. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3514605

24 Lyon, A., Mincholé, A., Bueno-Orovio, A., & Rodriguez, B. (2019). Improving the clinical understanding of hypertrophic cardiomyopathy by combining patient data, machine learning and computer simulations: A case study. Morphologie, 103(343), 169–179. https://doi.org/10.1016/j.morpho.2019.09.001

25 Vos, M., Starmans, M. P. A., Timbergen, M. J. M., Van der Voort, S. R., Padmos, G. A., Kessels, W., Niessen, W. J., Van Leenders, G. J. L. H., Grünhagen, D. J., Sleijfer, S., Verhoef, C., Klein, S., & Visser, J. J. (2019). Radiomics approach to distinguish between well differentiated liposarcomas and lipomas on MRI. British Journal of Surgery, 106(13), 1800–1809. https://doi.org/10.1002/bjs.11410

26 Mayerhoefer, M. E., Materka, A., Langs, G., Häggström, I., Szczypiński, P., Gibbs, P., & Cook, G. (2020). Introduction to Radiomics. Journal of Nuclear Medicine, 61(4), 488–495. https://doi.org/10.2967/jnumed.118.222893

27 Schofield, R., Ganeshan, B., Fontana, M., Nasis, A., Castelletti, S., Rosmini, S., Treibel, T., Manisty, C., Endozo, R., Groves, A., & Moon, J. (2019). Texture analysis of cardiovascular magnetic resonance cine images differentiates aetiologies of left ventricular hypertrophy. Clinical Radiology, 74(2), 140–149. https://doi.org/10.1016/j.crad.2018.09.016

28 Amano, Y., Yanagisawa, F., Omori, Y., Suzuki, Y., Ando, C., Yamamoto, H., & Matsumoto, N. (2020). Detection of Myocardial Tissue Alterations in Hypertrophic Cardiomyopathy Using Texture Analysis of T2-Weighted Short Inversion Time Inversion Recovery Magnetic Resonance Imaging. Journal of Computer Assisted Tomography, 44(3), 341–345. https://doi.org/10.1097/rct.0000000000001007

29 Alimadadi, A., Manandhar, I., Aryal, S., Munroe, P. B., Joe, B., & Cheng, X. (2020). Machine learning-based classification and diagnosis of clinical cardiomyopathies. Physiological Genomics, 52(9), 391–400. https://doi.org/10.1152/physiolgenomics.00063.2020

30 Starmans, M. P. A., van der Voort, S. R., Phil, T., Timbergen, M. J. M., & Vos, M. (in press). Reproducible radiomics through automated machine learning validated on twelve clinical applications. Alzheimers Disease Neuroimaging Initiative.

31 Starmans, M. P., Van der Voort, S. R., Castillo Tovar, J. M., Veenland, J. F., Klein, S., & Niessen, W. J. (2020). Radiomics. Handbook of Medical Image Computing and Computer Assisted Intervention, 429–456. https://doi.org/10.1016/b978-0-12-816176-0.00023-5

32 Lafata, K. J., Wang, Y., Konkel, B., Yin, F. F., & Bashir, M. R. (2021). Radiomics: a primer on high-throughput image phenotyping. Abdominal Radiology. https://doi.org/10.1007/s00261-021-03254-x

33 Richards, C. S., Bale, S., Bellissimo, D. B., Das, S., Grody, W. W., Hegde, M. R., Lyon, E., & Ward, B. E. (2008). ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. Genetics in Medicine, 10(4), 294–300. https://doi.org/10.1097/gim.0b013e31816b5cae

34 Van der Velde, N., Huurman, R., Hassing, H. C., Budde, R. P. J., Van Slegtenhorst, M. A., Verhagen, J. M. A., Schinkel, A. F. L., Michels, M., & Hirsch, A. (2021). Novel Morphological Features on CMR for the Prediction of Pathogenic Sarcomere Gene Variants in Subjects Without Hypertrophic Cardiomyopathy. Frontiers in Cardiovascular Medicine, 8. https://doi.org/10.3389/fcvm.2021.727405

35 Timbergen, M. J., Starmans, M. P., Padmos, G. A., Grünhagen, D. J., Van Leenders, G. J., Hanff, D., Verhoef, C., Niessen, W. J., Sleijfer, S., Klein, S., & Visser, J. J. (2020). Differential diagnosis and mutation stratification of desmoid-type fibromatosis on MRI using radiomics. European Journal of Radiology, 131, 109266. https://doi.org/10.1016/j.ejrad.2020.109266

36 Starmans, M. P. A., Buisman, F. E., Renckens, M., Willemssen, F. E. J. A., Van der Voort, S. R., Groot Koerkamp, B., Grünhagen, D. J., Niessen, W. J., Vermeulen, P. B., Verhoef, C., Visser, J. J., & Klein, S. (2021). Distinguishing pure histopathological growth patterns of colorectal liver metastases on CT using deep learning and radiomics: a pilot study. Clinical & Experimental Metastasis, 38(5), 483–494. https://doi.org/10.1007/s10585-021-10119-6

37 Van Griethuysen, J. J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G., Fillion-Robin, J. C., Pieper, S., & Aerts, H. J. (2017). Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research, 77(21), e104–e107. https://doi.org/10.1158/0008-5472.can-17-0339

38 Avanzo, M., Stancanello, J., & El Naqa, I. (2017). Beyond imaging: The promise of radiomics. Physica Medica, 38, 122–139. https://doi.org/10.1016/j.ejmp.2017.05.071

39 Van der Voort, S.R.; Starmans, M.P.A. Predict a Radiomics Extensive Differentiable

Interchangable Classification Toolkit (P ICT). Available online: https://github.com/Svdvoort/PREDICTFastr (accessed on 21 November 2021); doi:10.5281/zenodo.3854839.

40 Zwanenburg, A., Vallières, M., Abdalah, M. A., Aerts, H. J. W. L., Andrearczyk, V., Apte, A., Ashrafinia, S., Bakas, S., Beukinga, R. J., Boellaard, R., Bogowicz, M., Boldrini, L., Buvat, I., Cook, G. J. R., Davatzikos, C., Depeursinge, A., Desseroit, M. C., Dinapoli, N., Dinh, C. V., . . . Löck, S. (2020). The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology, 295(2), 328–338. https://doi.org/10.1148/radiol.2020191145

41 Frangi, A. F., Niessen, W. J., Vincken, K. L., & Viergever, M. A. (1998). Multiscale vessel enhancement filtering. Medical Image Computing and Computer-Assisted Intervention — MICCAI'98, 130–137. https://doi.org/10.1007/bfb0056195

42 Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), 971–987. https://doi.org/10.1109/tpami.2002.1017623

43 Kovesi, P. Phase Congruency Detects Corners and Edges. In Proceedings of the VIIth Digital Image Computing: Techniques and Applications, Sydney, Australia, 10-12 December 2003.

44 Symmetry and asymmetry from local phase. In: Tenth Australian joint conference on artificial intelligence; 1997: CiteSeer. p. 2–4.

45 Starmans, M. P., Miclea, R. L., Vilgrain, V., Ronot, M., Purcell, Y., Verbeek, J., Niessen, W. J., Ijzermans, J. N., de Man, R. A., Doukas, M., Klein, S., & Thomeer, M. G. (2021). Automated differentiation of malignant and benign primary solid liver lesions on MRI: an externally validated radiomics model. MedRxiv. https://doi.org/10.1101/2021.08.10.21261827

46 Lee, P. T., Dweck, M. R., Prasher, S., Shah, A., Humphries, S. E., Pennell, D. J., Montgomery, H. E., & Payne, J. R. (2013). Left Ventricular Wall Thickness and the Presence of Asymmetric Hypertrophy in Healthy Young Army Recruits. Circulation: Cardiovascular Imaging, 6(2), 262–267. https://doi.org/10.1161/circimaging.112.979294

47 Schuster, A., Chiribiri, A., Ishida, M., Morton, G., Paul, M., Hussain, S., Bigalke, B., Perera, D., & Nagel, E. (2011). End-systolic versus end-diastolic late gadolinium enhanced imaging for the assessment of scar transmurality. The International Journal of Cardiovascular Imaging, 28(4), 773–781. https://doi.org/10.1007/s10554-011-9877-3

48 Ederhy, S., Mansencal, N., Réant, P., Piriou, N., & Barone-Rochette, G. (2019). Role of multimodality imaging in the diagnosis and management of cardiomyopathies. Archives of Cardiovascular Diseases, 112(10), 615–629. https://doi.org/10.1016/j.acvd.2019.07.004

Appendix A: Morphological features

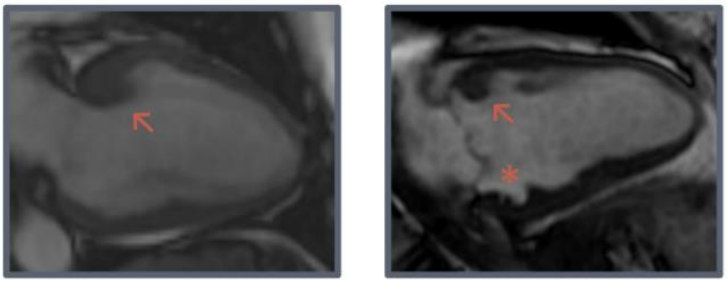

Figure A1: *Example of anterobasal hook and crypt, indicated by arrow and asterisk, respectively. Adapted from N. van der Velde et al. (2021) [34].*

Appendix B: Extracted image features

Table B1: *Overview of the 555 features used in this study: GLCM and GLCMMS features were calculated in four different directions (0, 45, 90, 135 degrees) using 16 gray levels and pixel distances of 1 and 3. LBP features were calculated using the following three parameter combinations: 1 pixel radius and 8 neighbours, 2 pixel radius and 12 neighbours, and 3 pixel radius and 16 neighbours. Gabor features were calculated using three different frequencies (0.05, 0.2, 0.5) and four different angles (0, 45, 90, 135 degrees). LoG features were calculated using three different widths of the Gaussian (1, 5 and 10 pixels). Vessel features were calculated using the full mask, the edge, and the inner region. Local phase features were calculated on the monogenic phase, phase congruency and phase symmetry. Adapted from M.P.A. Starmans (in press) [30].*

Histogram (13 features):	**LoG (13*3=39 features):**	**Vessel (12*3=39 features):**	**GLCM (MS) (6*3*4*2=144 features):**	**Gabor (13*4*3=156 features):**	**NGTDM (5 features):**	**LBP (13*3=39 features):**
min	min	min	contrast (normal, MS mean + std)	min	busyness	min
max	max	max	dissimilarity (normal, MS mean + std)	max	coarseness	max
mean	mean	mean	homogeneity (normal, MS mean + std)	mean	complexity	mean
median	median	median	angular second moment (ASM) (normal, MS mean + std)	median	contrast	median
std	std	std	energy (normal, MS mean + std)	std	strength	std
skewness	skewness	skewness	correlation (normal, MS mean + std)	skewness		skewness
kurtosis	kurtosis	kurtosis		kurtosis		kurtosis
peak	peak	peak		peak		peak
peak position	peak position	peak position		peak position		peak position
range	range	range		range		range
energy	energy	energy		energy		energy
quartile range	quartile	quartile		quartile range		quartile range
entropy	entropy	entropy		entropy		entropy

GLSZM (16 features):	**GLRM (16 features):**	**GLDM (14 features):**	**Shape (35 features):**	**Local phase (13*3=39 features):**
Gray Level Non Uniformity	Gray Level Non Uniformity	Dependence Entropy	compactness (mean + std)	min
Gray Level Non Uniformity Normalized	Gray Level Non Uniformity Normalized	Dependence Non-Uniformity	radial distance (mean + std)	max
Gray Level Variance	Gray Level Variance	Dependence Non-Uniformity Normalized	roughness (mean + std)	mean
High Gray Level Zone Emphasis	High Gray Level Run Emphasis	Dependence Variance	convexity (mean + std)	median
Large Area Emphasis	Long Run Emphasis	Gray Level Non-Uniformity	circular variance (mean + std)	std
Large Area High Gray Level Emphasis	Long Run High Gray Level Emphasis	Gray Level Variance	principal axes ratio (mean + std)	skewness
Large Area Low Gray Level Emphasis	Long Run Low Gray Level Emphasis	High Gray Level Emphasis	elliptic variance (mean + std)	kurtosis
Low Gray Level Zone Emphasis	Low Gray Level Run Emphasis	Large Dependence Emphasis	solidity (mean + std)	peak
SizeZoneNonUniformity	RunEntropy	Large Dependence High Gray Level Emphasis	area (mean, std, min + max	peak position
SizeZoneNonUniformityNormalized	RunLengthNonUniformity	Large Dependence Low Gray Level Emphasis	volume (total, mesh, volume)	range
SmallAreaEmphasis	RunLengthNonUniformityNormalized	Low Gray Level Emphasis	elongation	energy
SmallAreaHighGrayLevelEmphasis	RunPercentage	Small Dependence Emphasis	flatness	quartile
SmallAreaLowGrayLevelEmphasis	RunVariance	Small Dependence High Gray Level Emphasis	least axis length	entropy
ZoneEntropy	ShortRunEmphasis	Small Dependence Low Gray Level Emphasis	major axis length	
ZonePercentage	ShortRunHighGrayLevelEmphasis		minor axis length	
ZoneVariance	ShortRunLowGrayLevelEmphasis		maximum diameter 3D	
			maximum diameter 2D (rows, columns, slices)	
			sphericity	
			surface area	
			surface volume ratio	

Abbreviations: LoG: Laplacian of Gaussian; GLCM: gray level co-occurrence matrix; MS: multi slice; NGTDM: neighborhood gray tone difference matrix; LBP: local binary patterns; GLSZM: gray level size zone matrix; GLRLM: gray level run length matrix; GLDM: gray level dependence matrix; std: standard deviation.

Appendix C: Baseline models

Table C1: *Performance values of the 24 baseline models. Baseline models with the highest performance, which were eventually included in the radiomics models for further evaluation, are shown in bold*

Phase	View	Segmentation	AUC	Accuracy	Sensiticity	Specificity
ED	2CH	LVBloodpool	0.74 [0.64, 0.85]	0.65 [0.55, 0.75]	0.57 [0.41, 0.73]	0.77 [0.60, 0.94]
		LVMyocard	**0.81 [0.72, 0.90]**	**0.75 [0.66, 0.83]**	**0.81 [0.69, 0.93]**	**0.66 [0.49, 0.82]**
		LVBloodpool+Myocard	0.77 [0.67, 0.86]	0.69 [0.60, 0.79]	0.69 [0.53, 0.85]	0.70 [0.53, 0.87]
	3CH	LVBloodpool	**0.76 [065, 0.87]**	**0.67 [0.56, 0.77]**	**0.67 [0.52, 0.82]**	**0.67 [0.49, 0.85]**
		LVMyocard	0.65 [0.53, 0.76]	0.59 [0.49, 0.69]	0.59 [0.43, 0.75]	0.59 [0.40, 0.77]
		LVBloodpool+Myocard	0.72 [0.61, 0.82]	0.62 [0.51, 0.73]	0.58 [0.39, 0.77]	0.69 [0.50, 0.89]
	4CH	LVBloodpool	0.73 [0.63, 0.83]	0.64 [0.55, 0.73]	0.64 [0.47, 0.80]	0.65 [0.49, 0.81]
		LVMyocard	0.70 [0.57, 0.82]	0.63 [0.52, 0.75]	0.65 [0.48, 0.82]	0.60 [0.38, 0.82]
		LVBloodpool+Myocard	**0.75 [0.64, 0.86]**	**0.66 [0.56, 0.76]**	**0.64 [0.47, 0.81]**	**0.70 [0.52, 0.88]**
	SA	LVBloodpool	0.99 (0.98, 1,01)	0.95 (0.89, 1.00)	0.93 (0.85, 1.01)	0.97 (0.91, 1.03)
		LVMyocard	1.00 (0.98, 1.01)	0.97 (0.93, 1.01)	0.96 (0.90, 1.02)	0.99 (0.95, 1.03)
		LVBloodpool+Myocard	1.00 (0.99, 1.01)	0.98 (0.95, 1.01)	0.97 (0.93, 1.02)	1.00 (0.98, 1.02)
ES	2CH	LVBloodpool	0.82 [0.71, 0.93]	0.74 [0.64, 0.85]	0.81 [0.66, 0.95]	0.65 [0.47, 0.82]
		LVMyocard	0.80 [0.70, 0.90]	0.73 [0.62, 0.84]	0.76 [0.60, 0.91]	0.69 [0.52, 0.85]
		LVBloodpool+Myocard	**0.86 [0.78, 0.94]**	**0.79 [0.70, 0.88]**	**0.79 [0.66, 0.93]**	**0.78 [0.64, 0.93]**
	3CH	LVBloodpool	0.82 [0.72, 0.91]	0.73 [0.63, 0.83]	0.74 [0.60, 0.88]	0.71 [0.54, 0.88]
		LVMyocard	0.77 [0.67, 0.88]	0.69 [0.59, 0.78]	0.68 [0.53, 0.83]	0.70 [0.54, 0.86}
		LVBloodpool+Myocard	**0.83 [0.74, 0.92]**	**0.74 [0.64, 0.84]**	**0.72 [0.58, 0.85]**	**0.78 [0.63, 0.93]**
	4CH	LVBloodpool	0.84 [0.75, 0.93]	0.75 [0.66, 0.84]	0.76 [0.63, 0.90]	0.73 [0.59, 0.87]
		LVMyocard	0.86 [0.76, 0.95]	0.77 [0.67, 0.87]	0.84 [0.72, 0.96]	0.67 [0.51, 0.83]
		LVBloodpool+Myocard	**0.86 [0.77, 0.94]**	**0.77 [0.68, 0.86]**	**0.80 [0.66, 0.93]**	**0.74 [0.60, 0.88]**
	SA	LVBloodpool	0.99 (0.98, 1.01)	0.96 (0.91, 1.01)	0.95 (0.89, 1.01)	0.96 (0.88, 1.05)
		LVMyocard	0.99 (0.97, 1.01)	0.95 (0.90, 0.99)	0.95 (0.89, 1.01)	0.94 (0.86, 1.02)
		LVBloodpool+Myocard	1.00 (0.99, 1.00)	0.96 (0.92, 1.00)	0.97 (0.92, 1.02)	0.96 (0.88, 1.03)

Abbreviations: 2CH, 2-chamber; 3CH, 3-chamber; 4CH, 4-chamber; SA, short axis; ED, End-diastolic; ES, End-systolic; AUC, Area Under the Curve

** Outcomes are presented with the 95% confidence interval.*

Appendix D: Automatic segmentation

Table D1: *Performance values of the three datasets performed with automatic segmentation based on: model 1: 2CH view $LV_{Bloodpool+Myocad}$ in ES phase; model 2: 2CH, 3CH and 4CH view in $LV_{Bloodpool+Myocad}$ ES phase; model 3: 2CH, 3CH and 4CH view in $LV_{Bloodpool+Myocad}$ ES phase and 2CH view $LV_{Myocard}$, 3CH view $LV_{Bloodpool}$ and 4CH view $LV_{Bloodpool+Myocard}$ in ED phase.*

	Development dataset		
	Model 1	Model 2	Model 3
AUC	0.73 [0.62, 0.85]	0.75 [0.65, 0.84]	0.77 [0.67, 0.87]
Accuracy	0.67 [0.56, 0.77]	0.63 [0.52, 0.74]	0.66 [0.55, 0.76]
Sensiticity	0.69 [0.54, 0.85]	0.54 [0.31, 0.77]	0.57 [0.39, 0.75]
Specificity	0.62 [0.43, 0.81]	0.76 [0.58, 0.94]	0.78 [0.61, 0.96]
	Prospective validation dataset		
	Model 1	Model 2	Model 3
AUC	0.73 [0.53, 0.93]	0.77 [0.58, 0.97]	0.84 [0.66, 1.01]
Accuracy	0.72 [0.58, 0.87]	0.78 [0.64, 0.92]	0.81 [0.68, 0.93]
Sensiticity	0.64 [0.34, 0.93]	0.45 [0.15, 0.76]	0.55 [0.25, 0.84]
Specificity	0.76 [0.59, 0.93]	0.92 [0.81, 1.03]	0.92 [0.81, 1.03]
	External validation dataset		
	Model 1	Model 2	Model 3
AUC	0.64 [0.48, 0.80]	0.50 [0.32, 0.68]	0.57 [0.41, 0.73]
Accuracy	0.55 [0.42, 0.68]	0.53 [0.39, 0.67]	0.53 [0.39, 0.67]
Sensiticity	1.00 [-, -]	1.00 [-, -]	1.00 [-, -]
Specificity	0.04 [-0.03, 0.12]	0.00 [-, -]	0.00 [-, -]

Abbreviations: AUC, Area Under the Curve

** Outcomes are presented with the 95% confidence interval.*

Table D2: *The pairwise Dice Similarity Coefficient (DSC) values between manual and automatic segmentation within different datasets used in this study.*

		ED			ES		
Dataset	Segmentation	2CH	3CH	4CH	2CH	3CH	4CH
Development	$LV_{Bloodpool}$	0.97 ± 0.01	0.96 ± 0.02	0.97 ± 0.02	0.93 ± 0.03	0.92 ± 0.04	0.92 ± 0.03
	$LV_{Myocard}$	0.88 ± 0.03	0.87 ± 0.03	0.89 ± 0.03	0.91 ± 0.03	0.89 ± 0.09	0.90 ± 0.03
	$LV_{Bloodpool+Myocard}$	0.93 ± 0.05	0.92 ± 0.05	0.93 ± 0.05	0.92 ± 0.03	0.90 ± 0.07	0.91 ± 0.03
Prospective validation	$LV_{Bloodpool}$	0.97 ± 0.01	0.96 ± 0.02	0.97 ± 0.01	0.93 ± 0.02	0.92 ± 0.04	0.92 ± 0.03
	$LV_{Myocard}$	0.88 ± 0.04	0.88 ± 0.03	0.89 ± 0.02	0.91 ± 0.03	0.87 ± 0.15	0.91 ± 0.02
	$LV_{Bloodpool+Myocard}$	0.93 ± 0.05	0.92 ± 0.05	0.93 ± 0.04	0.92 ± 0.03	0.89 ± 0.11	0.92 ± 0.02
External validation	$LV_{Bloodpool}$	0.96 ± 0.01	0.95 ± 0.02	0.95 ± 0.02	0.90 ± 0.02	0.90 ± 0.06	0.89 ± 0.03
	$LV_{Myocard}$	0.86 ± 0.03	0.85 ± 0.03	0.85 ± 0.03	0.90 ± 0.03	0.88 ± 0.03	0.89 ± 0.03
	$LV_{Bloodpool+Myocard}$	0.91 ± 0.06	0.90 ± 0.05	0.90 ± 0.06	0.90 ± 0.02	0.89 ± 0.05	0.89 ± 0.03

Abbreviations: 2CH, 2-chamber; 3CH, 3-chamber; 4CH, 4-chamber; ED, End-diastolic; ES, End-systolic; AUC, Area Under the Curve

** Values are given in mean ± std*

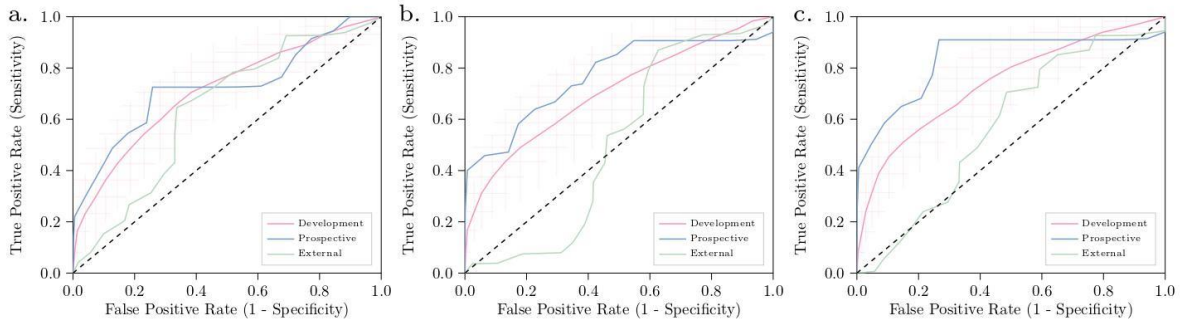

Figure D1: *ROC curves of automatic segmentation present the model on development dataset (red), prospective validation dataset (blue) and external validation dataset (green). The ROC curves show: a) model 1: 2CH view* $LV_{Bloodpool+Myocad}$ *in ES phase; b) model 2: 2CH, 3CH and 4CH view in* $LV_{Bloodpool+Myocad}$ *ES phase, and c) model 3: 2CH, 3CH and 4CH view in* $LV_{Bloodpool+Myocad}$ *ES phase and 2CH view* $LV_{Myocard}$*, 3CH view* $LV_{Bloodpool}$ *and 4CH view* $LV_{Bloodpool+Myocard}$ *in ED phase.*

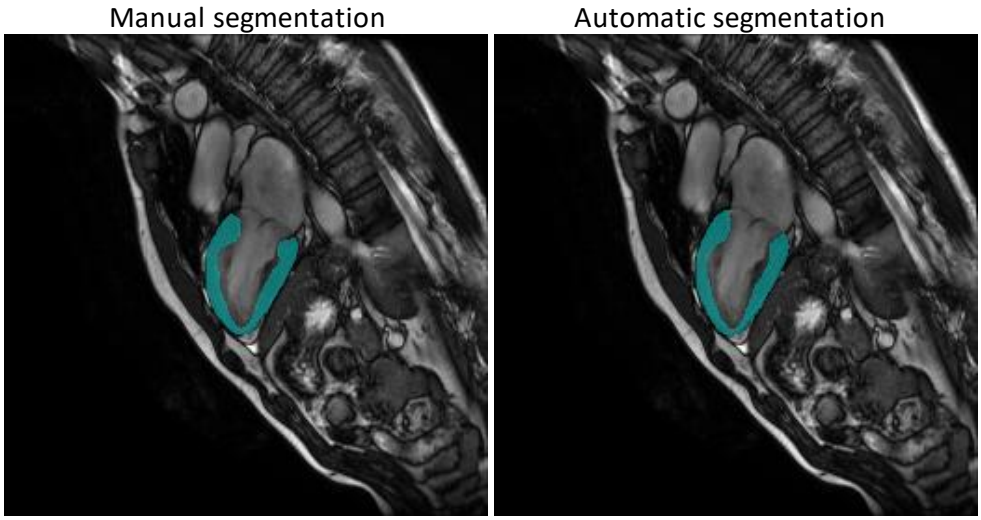

Figure D2: *Example of manual and automatic segmentation from development dataset. Difference in segmentation visible in ends of basal wall and overall smoother delineation*

Appendix E: Additional models

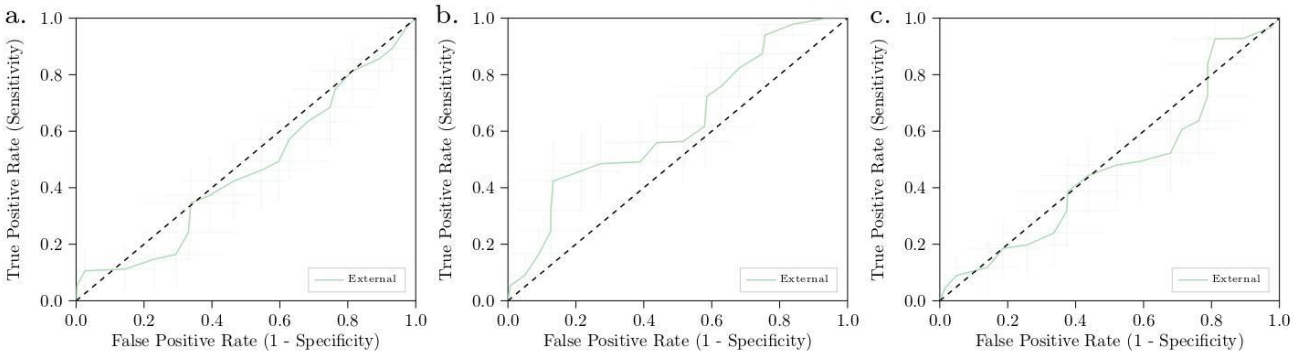

Figure G1: *ROC curves present the additional model on external validation dataset (green). The ROC curves show: a) model 1A: shape feature extraction solely; b) model 1B: manually adjusted the slice thickness to same slice thickness as training set, and c) model 1C: combined setups of model 1A and 1B.*

Table D1: *Performance values of the additional models performed on external validation dataset based on: model 1A: shape feature extraction solely; model 1B: manually adjusted the slice thickness to same slice thickness as training set, and model 1C: combined setups of model 1A and 1B.*

	External validation dataset		
	Model 1A	Model 1B	Model 1C
AUC	0.46 [0.29, 0.62]	0.63 [0.47, 0.78]	0.47 [0.31, 0.63]
Accuracy	0.53 [0.39, 0.67]	0.53 [0.40, 0.66]	0.53 [0.39, 0.67]
Sensiticity	1.00 [-, -]	1.00 [-, -]	0.93 [0.83, 1.03]
Specificity	0.00 [-, -]	0 [-, -]	0.08 [0.03, 0.20]

Abbreviations: AUC, Area Under the Curve

** Outcomes are presented with the 95% confidence interval.*

Appendix F: Significant features

Table F1: *Overview of radiomics features with their statistically significant p-values obtained from development dataset performed with manual segmentation. The statistical significance was assessed using a Mann-Whitney U test. Only the features that showed statistically significant different are included. All p-values were corrected for multiple testing by multiplying the p-values with the total number of tests. CT_0 represents features extracted from $LV_{Myocard}$ segmentation, CT_1 represents features extracted from $LV_{Bloodpool}$ segmentation.*

Features	p-value
tf_Gabor_energy_F0.05_A0.0_CT_1	7.00E-05
sf_Maximum2DDiameterRow_CT_1	4.35E-04
tf_Gabor_mean_F0.05_A0.0_CT_1	6.67E-04
sf_rad_dist_std_2D_CT_1	6.92E-04
sf_Maximum3DDiameter_CT_1	8.01E-04
sf_Maximum2DDiameterSlice_CT_1	8.01E-04
tf_Gabor_energy_F0.05_A0.79_CT_1	2.42E-03
phasef_monogenic_entropy_WL3_N5_CT_1	3.19E-03
tf_LBP_energy_R15_P36_CT_1	4.64E-03
sf_solidity_avg_2D_CT_0	4.64E-03
phasef_phasesym_entropy_WL3_N5_CT_1	5.13E-03
tf_Gabor_entropy_F0.05_A0.0_CT_1	5.67E-03
tf_Gabor_min_F0.5_A0.79_CT_1	5.67E-03
tf_Gabor_entropy_F0.05_A0.79_CT_1	6.07E-03
tf_LBP_energy_R8_P24_CT_1	7.66E-03
tf_LBP_entropy_R3_P12_CT_1	9.03E-03
sf_MajorAxisLength_CT_1	9.03E-03
tf_LBP_entropy_R15_P36_CT_1	9.96E-03
tf_Gabor_entropy_F0.5_A2.36_CT_1	1.10E-02
tf_Gabor_entropy_F0.5_A0.79_CT_1	1.10E-02
tf_Gabor_entropy_F0.2_A0.0_CT_1	1.13E-02
tf_Gabor_entropy_F0.5_A1.57_CT_1	1.17E-02
tf_Gabor_entropy_F0.5_A0.0_CT_1	1.21E-02
tf_Gabor_kurtosis_F0.2_A0.0_CT_0	1.25E-02
tf_Gabor_entropy_F0.2_A2.36_CT_1	1.29E-02
tf_Gabor_entropy_F0.2_A0.79_CT_1	1.29E-02
tf_Gabor_kurtosis_F0.2_A2.36_CT_0	1.33E-02
tf_Gabor_entropy_F0.2_A1.57_CT_1	1.56E-02
tf_Gabor_entropy_F0.05_A2.36_CT_1	1.56E-02
tf_LBP_entropy_R8_P24_CT_1	1.56E-02
tf_NGTDM_Busyness_CT_1	2.08E-02
hf_entropy_CT_1	2.15E-02
tf_Gabor_entropy_F0.05_A1.57_CT_1	2.51E-02
sf_Sphericity_CT_0	2.59E-02
sf_rad_dist_avg_2D_CT_1	2.59E-02
tf_Gabor_kurtosis_F0.05_A0.0_CT_1	2.76E-02
tf_Gabor_kurtosis_F0.5_A0.0_CT_1	2.93E-02
tf_NGTDM_Strength_CT_1	3.02E-02
tf_Gabor_energy_F0.2_A2.36_CT_0	4.11E-02

Table F2: *Overview of radiomics features with their statistically significant p-values obtained from development dataset performed with automatic segmentation. The statistical significance was assessed using a Mann-Whitney U test. Only the features that showed statistically significant different are included. All p-values were corrected for multiple testing by multiplying the p-values with the total number of tests. CT_0 represents features extracted from* $LV_{Myocard}$ *segmentation, CT_1 represents features extracted from* $LV_{Bloodpool}$ *segmentation.*

Label	p-value
sf_Maximum2DDiameterRow_1_0_CT_1	8.56E-03
sf_solidity_avg_2D_1_0_CT_0	2.67E-02
tf_Gabor_min_F0.5_A0.79_1_0_CT_1	2.93E-02