frontiers | Research Topics

Current and future role of artificial intelligence in cardiac imaging, volume II

Edited by
Steffen Erhard Petersen, Alistair A. Young, Tim Leiner
and Karim Lekadir

Published in
Frontiers in Cardiovascular Medicine

ISSN 1664-8714
ISBN 978-2-8325-3046-7
DOI 10.3389/978-2-8325-3046-7

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

July 2023

Current and future role of artificial intelligence in cardiac imaging, volume II

Topic editors

Steffen Erhard Petersen — Queen Mary University of London, United Kingdom
Alistair A. Young — King's College London, United Kingdom
Tim Leiner — Mayo Clinic, United States
Karim Lekadir — University of Barcelona, Spain

Citation

Petersen, S. E., Young, A. A., Leiner, T., Lekadir, K., eds. (2023). *Current and future role of artificial intelligence in cardiac imaging, volume II*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-3046-7

Table of contents

frontiers | Frontiers in Cardiovascular Medicine

TYPE Editorial
PUBLISHED 08 June 2023
DOI 10.3389/fcvm.2023.1220836

OPEN ACCESS

EDITED BY
Sebastian Weingärtner,
Delft University of Technology, Netherlands

REVIEWED BY
Li-Yueh Hsu,
National Institutes of Health (NIH), United States
Mehmet Akcakaya,
University of Minnesota Twin Cities,
United States

*CORRESPONDENCE
Alistair A. Young
✉ alistair.young@kcl.ac.uk

RECEIVED 11 May 2023
ACCEPTED 02 June 2023
PUBLISHED 08 June 2023

CITATION
Leiner T, Lekadir K, Petersen SE and Young AA (2023) Editorial: Current and future role of artificial intelligence in cardiac imaging, volume II.
Front. Cardiovasc. Med. 10:1220836.
doi: 10.3389/fcvm.2023.1220836

Editorial: Current and future role of artificial intelligence in cardiac imaging, volume II

Tim Leiner[1], Karim Lekadir[2], Steffen E. Petersen[3,4] and Alistair A. Young[5]*

[1]Department of Radiology, Mayo Clinic, Rochester, MN, United States, [2]Departament de Matemàtiques and Informàtica, Universitat de Barcelona, Artificial Intelligence in Medicine Lab (BCN-AIM), Barcelona, Spain, [3]Barts Heart Centre, Barts Health NHS Trust, London, United Kingdom, [4]NIHR Barts Biomedical Research Centre, William Harvey Research Institute, Queen Mary University of London, London, United Kingdom, [5]School of Biomedical Engineering and Imaging Sciences, King's College London, London, United Kingdom

KEYWORDS

artificial intelligence, cardiac imaging modalities, big data, cardiac image analysis, cardiovascular personalized medicine, AI adoption and translation

Editorial on the Research Topic
Current and future role of artificial intelligence in cardiac imaging, volume II

Introduction

Heart imaging without the need for invasive procedures plays a crucial role in various aspects of cardiology, including diagnosis, risk evaluation, treatment decisions, medical and invasive therapies, prognosis, and ongoing monitoring. As a key component in the pursuit of precision medicine, cardiac imaging enables a personalized approach to healthcare. Cardiac imaging is essential for a comprehensive understanding of cardiovascular conditions, intricate physiology, and the implications of imaging results for managing cardiovascular health and disease. Consequently, the demand for cardiac imaging continues to grow.

Artificial Intelligence (AI) methods are increasingly needed for improving imaging exams, to optimize workflow, reduce waiting times, improve accuracy and precision, reduce costs and benefit patient outcomes. The special issue on "Current and Future Role of Artificial Intelligence in Cardiac Imaging" in 2020 provided 10 comprehensive reviews of a wide range of AI applications in cardiovascular imaging, from efficient image acquisition and fast reconstruction, to structure and function analysis, to statistical atlases and imaging-genetics synergies. In the meantime the field has been advancing rapidly. The second volume of this topic focuses on original research articles which showcase new methods and applications. Applications were invited for all topics and forms of cardiac imaging, and the resulting published papers reflect a range of modalities and applications which have a common goal of improving patient outcomes through extraction of more detailed and precise information than has previously been possible.

Special issue content

The special issue attracted a greater number of submissions than the first edition, with 15 papers eventually published (Abdulkareem et al., Alabed et al., Beetz et al., Campello et al., Chen et al., Hampe et al., Li et al., Lin et al., Lin et al., Puyol-Antón et al., Suinesiaputra et al., Szabo et al., Zhai et al., Zhao et al. and Zhao et al.). **Figure 1** shows a breakdown of topics and applications.

Imaging modality was varied with CMR being the most common (Abdulkareem et al., Alabed et al., Beetz et al., Campello et al., Puyol-Antón et al. and Suinesiaputra et al.), followed closely by CT (Chen et al., Hampe et al., Li et al. and Zhai et al.) and echocardiography (Lin et al., Lin et al. and Puyol-Antón et al.). One paper provided a guide to trustworthy and responsible AI in all imaging modalities (Szabo et al.). Although not an imaging method, ECG processing is an important topic in cardiovascular disease, and can give anatomical information such as hypertrophy, which was the focus of one paper (Zhao et al.). In terms of application area, the most common was structural analysis (segmentation of chambers, vessels, and detection of hypertrophy) (Hampe et al., Lin et al., Suinesiaputra et al., Zhao et al. and Zhao et al.), followed by image motion analysis [in echo (Lin et al.) and CT (Chen et al. and Li et al.)]. These reflect the ongoing development of methods for quantifying anatomy and function in cardiac images, which are extremely important for diagnosis and prognosis. Two papers were concerned with synthesis of images or anatomy, using generative adversarial networks (Campello et al.) and variational auto encoders (Beetz et al.) respectively. This is an exciting area of research with the promise to synthesise large numbers of datasets with different pathology and characteristics, which can be used to understand the influence of disease processes on structure and function. Two papers discussed learning contrast-enhanced information from non-contrast scans, applied to late gadolinium enhancement CMR (Abdulkareem et al.) and calcium scoring CT (Zhai et al.) respectively. These highlight the ability of AI methods to leverage information present in standard scans to provide additional information, which would normally require an extra scan and administration of a contrast agent. The papers on AI fairness (Puyol-Antón et al.), and trustworthiness (Szabo et al.), illustrate another promising area of research into methods for improving AI methods, mitigating bias in the training data and explaining how inferences are made in complex networks. These were identified in the first special issue as topics of high importance. Another important topic highlighted by the first special issue is the reporting of AI methods to foster reproducibility and generalization, and this is the focus of the review article (Alabed et al.).

Future perspectives

AI methods are becoming ubiquitous in all areas of cardiac imaging. This is driven by the increasing role of imaging in guidelines for patient treatment, as well as the wealth of data generated by large cohort imaging studies such as the UK Biobank and the Multi-Ethnic Study of Atherosclerosis. The recent advent of widely available large language models which link imaging and videos with text such as GPT-4 highlight the rapidly advancing technology which will transform medical imaging in general and cardiovascular imaging in particular. The ability to generate reports from images and to combine images and reports into structured datasets will enable large amounts of data already residing in hospital databases to be reused for new applications. Generative models will enable new understanding of the ways in which longitudinal changes between imaging exams can be more precisely quantified, by comparing predicted follow-up images with the actual scans and highlighting any anomalies. Physics-based networks and neural implicit functions are also on the horizon which will enable super-resolution in fluid flow imaging, and computation of physical entities such as pressure and stiffness from imaging data. The linking of digital twins with

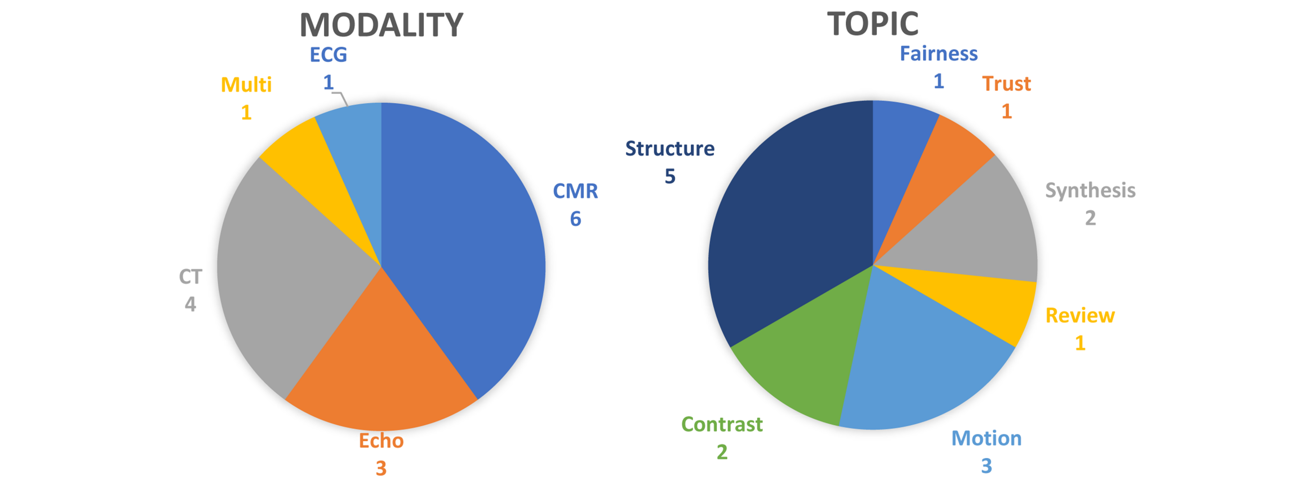


FIGURE 1
Breakdown of special issue papers by modality and topic.

imaging methods will enable surrogate AI-based methods to simulate growth and remodeling in health and disease, to estimate physiological parameters from non-invasive imaging. In summary, this second issue on AI in cardiac imaging illustrates important areas which are rapidly developing. However the next advances in this area are only limited by the imaginations of the researchers in the community. At present, there is no AI solution for this!

Author contributions

All co-authors discussed the structure and content of the special issue and editorial. AY wrote the first draft. TL, KL, and SP revised the manuscript. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Cardiovascular Medicine

ORIGINAL RESEARCH
published: 21 January 2022
doi: 10.3389/fcvm.2021.807728

Deep Learning Analysis of Cardiac MRI in Legacy Datasets: Multi-Ethnic Study of Atherosclerosis

*Avan Suinesiaputra [1,2], Charlène A. Mauger [1], Bharath Ambale-Venkatesh [3], David A. Bluemke [4], Josefine Dam Gade [5], Kathleen Gilbert [6], Markus H. A. Janse [7], Line Sofie Hald [5], Conrad Werkhoven [6], Colin O. Wu [8], Joao A. C. Lima [3] and Alistair A. Young [9]**

[1] Department of Anatomy and Medical Imaging, University of Auckland, Auckland, New Zealand, [2] Department of Biomedical Engineering, School of Biomedical Engineering and Imaging Sciences, King's College London, London, United Kingdom, [3] Johns Hopkins Medical Center, Baltimore, MD, United States, [4] Department of Radiology, University of Wisconsin School of Medicine and Public Health, Madison, WI, United States, [5] Department of Biomedical Engineering and Informatics, School of Medicine and Health, Aalborg University, Aalborg, Denmark, [6] Auckland Bioengineering Institute, University of Auckland, Auckland, New Zealand, [7] Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, Netherlands, [8] Division of Intramural Research, National Heart, Lung and Blood Institute, National Institutes of Health, Baltimore, MD, United States, [9] Faculty of Life Sciences & Medicine, School of Biomedical Engineering & Imaging Sciences, King's College London, London, United Kingdom

OPEN ACCESS

Edited by:
*Luis Lopes,
University College London,
United Kingdom*

Reviewed by:
*Rhodri Davies,
University College London,
United Kingdom
Andrew James Swift,
The University of Sheffield,
United Kingdom
Filippo Cademartiri,
Gabriele Monasterio Tuscany
Foundation (CNR), Italy*

***Correspondence:**
*Alistair A. Young
alistair.young@kcl.ac.uk*

Specialty section:
*This article was submitted to
Cardiovascular Imaging,
a section of the journal
Frontiers in Cardiovascular Medicine*

Received: *02 November 2021*
Accepted: *24 December 2021*
Published: *21 January 2022*

Citation:
*Suinesiaputra A, Mauger CA,
Ambale-Venkatesh B, Bluemke DA,
Dam Gade J, Gilbert K, Janse MHA,
Hald LS, Werkhoven C, Wu CO,
Lima JAC and Young AA (2022) Deep
Learning Analysis of Cardiac MRI in
Legacy Datasets: Multi-Ethnic Study
of Atherosclerosis.
Front. Cardiovasc. Med. 8:807728.
doi: 10.3389/fcvm.2021.807728*

The Multi-Ethnic Study of Atherosclerosis (MESA), begun in 2000, was the first large cohort study to incorporate cardiovascular magnetic resonance (CMR) to study the mechanisms of cardiovascular disease in over 5,000 initially asymptomatic participants, and there is now a wealth of follow-up data over 20 years. However, the imaging technology used to generate the CMR images is no longer in routine use, and methods trained on modern data fail when applied to such legacy datasets. This study aimed to develop a fully automated CMR analysis pipeline that leverages the ability of machine learning algorithms to enable extraction of additional information from such a large-scale legacy dataset, expanding on the original manual analyses. We combined the original study analyses with new annotations to develop a set of automated methods for customizing 3D left ventricular (LV) shape models to each CMR exam and build a statistical shape atlas. We trained VGGNet convolutional neural networks using a transfer learning sequence between two-chamber, four-chamber, and short-axis MRI views to detect landmarks. A U-Net architecture was used to detect the endocardial and epicardial boundaries in short-axis images. The landmark detection network accurately predicted mitral valve and right ventricular insertion points with average error distance <2.5 mm. The agreement of the network with two observers was excellent (intraclass correlation coefficient >0.9). The segmentation network produced average Dice score of 0.9 for both myocardium and LV cavity. Differences between the manual and automated analyses were small, i.e., $<1.0 \pm 2.6$ mL/m^2 for indexed LV volume, 3.0 ± 6.4 g/m^2 for indexed LV mass, and $0.6 \pm 3.3\%$ for ejection fraction. In an independent atlas validation dataset, the LV atlas built from the fully automated pipeline showed similar statistical relationships to an atlas built from the manual analysis. Hence, the proposed pipeline is not only a promising framework to automatically assess additional measures of ventricular function, but also to study relationships between cardiac morphologies and future cardiac events, in a large-scale population study.

Keywords: cardiac anatomy, machine learning, left ventricle, MRI, deep learning

INTRODUCTION

Cardiovascular magnetic resonance (CMR) is widely used for the non-invasive assessment of cardiac function, and has excellent accuracy and reproducibility for clinical evaluation of cardiac mass and volume (1). The ability of CMR to evaluate all regions of the heart with high signal to noise ratio without harmful radiation exposure has led to its use in several large cohort studies investigating the development of cardiac disease in general populations, including the Multi-Ethnic Study of Atherosclerosis (MESA) (2) and the UK Biobank (3). MESA was the first large epidemiological study to utilize CMR to evaluate pre-clinical characteristics of participants before the onset of clinical symptoms of cardiovascular disease (CVD). The baseline MESA CMR exam was performed between 2000 and 2002 using the common imaging method prevalent at that time: gradient echo cine imaging. However, this imaging method has been largely replaced by steady-state free precession cine imaging in subsequent studies and in clinical practice (4). Due to differences in fundamental properties that comprise image contrast as well as spatial resolution (5), image analysis tools designed for modern steady-state free precession images are likely to have poor performance when applied to 20-year-old gradient echo imaging.

Three-dimensional (3D) atlas-based analysis methods have been developed to quantify subtle differences in heart shape (remodeling) and function associated with CVD risk factors such as hypertension, smoking and diabetes (6–10). To date, these methods have only been applied to a limited subset of MESA cases, due to the need for additional image analysis which was not performed as part of the original CMR analysis. This is a recurring problem in large cohort legacy datasets, since a limited amount of annotations are available and manual analysis is unfeasible due to time and resource constraints. A fully automated processing pipeline is therefore necessary to enable more comprehensive analysis and make better use of the large amount of image data acquired.

Deep learning methods, particularly convolutional neural networks (CNN), have demonstrated high accuracy and reproducibility for fully automated image analysis when sufficient training images and high computational power is available (11, 12). CNN can automatically learn optimal weights for convolutional operations in each layer to extract image features. It has been applied and adapted for image classification (13), object recognition (14), segmentation (15), and image registration (16). However, CNN solutions trained on modern steady-state free precession images fail when applied to the old gradient echo images. Transfer learning approaches, such as pre-training or layer-wise fine tuning, have been proposed to adapt a network to different domain, but when large amount of labeled data is available, full training from scratch is the best option to train a CNN (17).

In this study, we developed an automated CMR preprocessing pipeline, shown in **Figure 1**. In order to automatically construct 3D LV shape models and a statistical shape atlas, anatomical landmarks were required to orient the model and contours were required to customize the shape models. Custom CNNs were used to detect anatomical landmarks and to segment

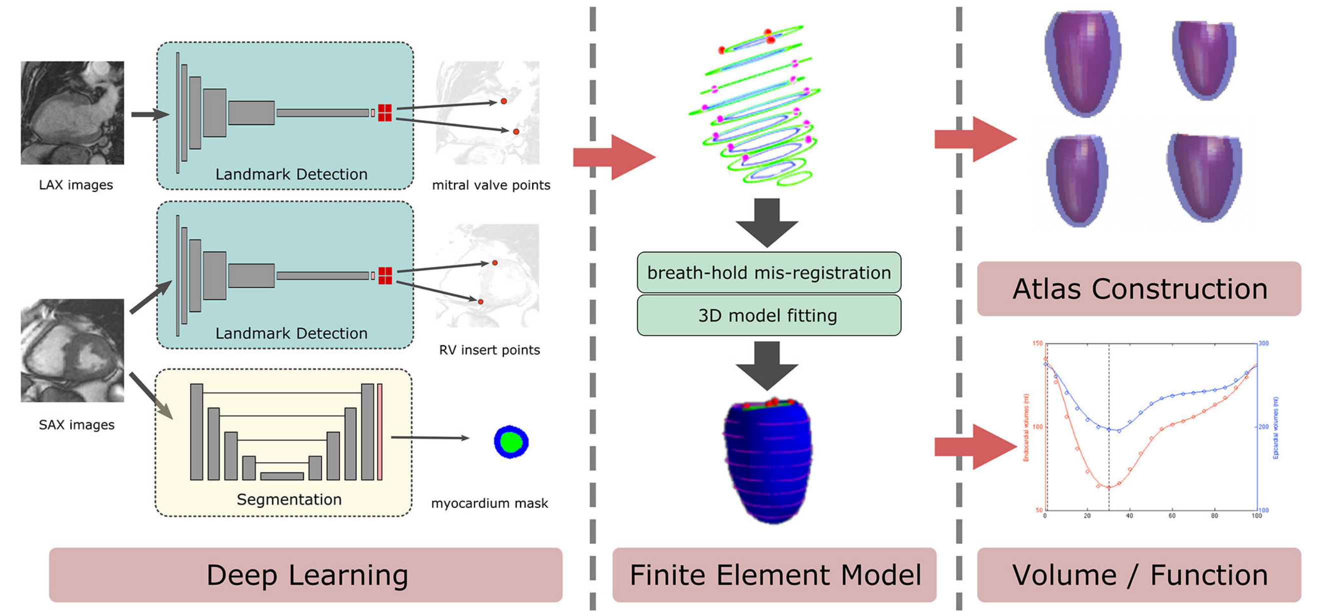


FIGURE 1 | Fully-automated atlas generation pipeline of cardiac MRI analyses. Three deep learning networks were trained to perform: (1) detection of mitral valve points from long-axis (LAX) images, from both two-chamber or four-chamber views, (2) detection of right ventricular (RV) insert points from short-axis (SAX) images, and (3) segmentation of myocardium mask from SAX images. Landmark points and contours from myocardium mask images were converted into 3D patient coordinates to guide the customization of a left ventricle (LV) model. Breath-hold mis-registration of SAX slices were corrected. The final model was used to construct a statistical shape LV atlas.

TABLE 1 | Patient demographics from the MESA cohort.

		MESA CMR	Landmark detection	Segmentation	Atlas validation
N		5,003	2,372	1,545	1,052
Age (years)		61.5 (10.1)	61.3 (10.1)	61.0 (10.2)**	60.1 (9.8)***
Gender	Female	2,622 (52.4)	1,230 (51.9)	814 (52.7)	430 (40.9)***
	Male	2,381 (47.6)	1,142 (48.1)	731 (47.3)	622 (59.1)
SBP (mmHg)		125.4(21.3)	126.2 (21.9)*	126.4 (22.0)*	124.8 (20.2)
DBP (mmHg)		71.8 (10.30)	71.6 (10.3)	71.7 (10.3)	73.6 (10.1)***
Heart Rate (bpm)		62.8 (9.5)	62.7 (9.5)	62.9 (9.5)	62.1 (9.6)**
Diabetes	Yes	459 (9.2)	232 (9.8)	162 (10.5)*	74 (7.0)**
	No	4,544 (90.8)	2,140 (90.2)	1,383 (89.5)	978 (93.0)
Hypertension	Yes	1,766 (35.3)	805 (34.0)	539 (34.9)	373 (35.5)
	No	3,234 (64.7)	1,566 (66.0)	1,005 (65.1)	677 (64.5)
Smoking status	Never	2,569 (51.5)	1,237 (52.3)	805 (52.4)	511 (48.6)
	Former	1,786 (35.8)	824 (34.9)	521 (33.9)	394 (37.5)
	Current	634 (12.7)	302 (12.8)	209 (13.6)	146 (13.9)
Framingham score		13.9 (9.5)	14.1 (9.5)	14.0 (9.6)	13.7 (9.2)

Two sub-cohorts were defined to train and validate deep learning networks for landmark detection and segmentation. Another sub-cohort, disjoint from the two training datasets, was defined for validation of the atlas generated from automated compared with core lab manual analysis. Continuous variables are written as mean (standard deviation), while categorical variables are written as count (percentage). Statistical tests were performed between a sub-cohort against its complement with one-way ANOVA for continuous variables and χ^2 *test for categorical variables.* *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$ *for difference between a particular sub-cohort and the rest of the MESA CMR cohort.*

myocardium from the MESA gradient echo CMR images. We demonstrate that these networks provide robust and consistent contours and landmarks compared with manual annotations. We also show that an LV atlas built from the proposed pipeline produced similar associations with CVD risk factors to an atlas built from manual analyses.

MATERIALS AND METHODS

Dataset

The MESA study has been described previously in (2). Briefly, the CMR exam consisted of 5,098 participants who were initially free from clinically recognized CVD at the time of enrollment (18). Images were acquired with 1.5T MR scanners at six different institutions across the United States using Siemens and General Electric scanners between July 2000 and July 2002. All images were acquired during breath-holding at resting lung volume. From each CMR examination, we only included short- and long-axis cine images for this study. The cine CMR images consist of 10–12 short-axis slices (SAX), single four-chamber (4CH) and single two-chamber (2CH) long-axis (LAX) views. All cine images were acquired using fast gradient echo pulse sequence, with typical parameters of slice thickness 6, 4 mm gap, field of view 360–400 mm, 256 × 160 image matrix (smallest 192 × 160), flip angle 20°, echo time 3–5 ms, repetition time 8–10 ms with 20–30 frames per slice (temporal resolution <50 ms) and pixel size from 1.4 to 2.5 mm/pixel depending on patient size. All participants gave informed consent, and the institutional review board at each site approved the study protocol.

The MESA Core Lab provided 2D contour points drawn manually by trained technologists. The Core Lab analysis protocol for MESA study has been described previously (18), including inter- and intra-observer variability. Briefly, endocardial and epicardial borders were traced on short-axis slices at end-diastole (ED) and end-systole (ES) frames using Q-MASS software (version 4.2, Medis, the Netherlands). Papillary muscles were included in the blood pool. All image contours were reviewed and corrected by a cardiac MR physician.

In total 5,003 exams had adequate MRI data for analysis (**Table 1**). Of these, 2,496 cases (49.9%) were available from the Cardiac Atlas Project (19), while the remaining 2,507 cases (50.1%) were provided by the MESA Core Lab at the Johns Hopkins Medical Center, Baltimore, USA. In this study, we used cases from the Cardiac Atlas Project for training, testing and validating the deep learning networks, while the remaining cases were used for an independent LV atlas validation. **Figure 2** shows detail divisions of the baseline MESA cohort for the automated CMR analysis pipeline development.

Of the 2,496 cases for the training data, 2,273 cases had manual contours. We further excluded 728 cases due to misalignment of contours with the image slices, unmatched contours with DICOM images or missing DICOM header information. This resulted in 1,545 cases to train the segmentation network, which were randomly split into 1,236 training cases (80%), 154 validation cases (10%), and 155 test cases (10%). Contour points were converted into mask images consisting of three disjoint areas: myocardium, LV cavity, and background pixels.

As anatomical cardiac landmark points were not part of the MESA Core Lab protocol, we employed two experienced analysts (both had >5 years of fulltime experience in CMR exams) to manually place cardiac landmarks by using Cardiac Image Modeler software (version 6.2; Auckland MR Research Group, University of Auckland, New Zealand). Of the 2,496 cases for the training data, 2,372 cases had adequate annotations to train the landmark detection network.

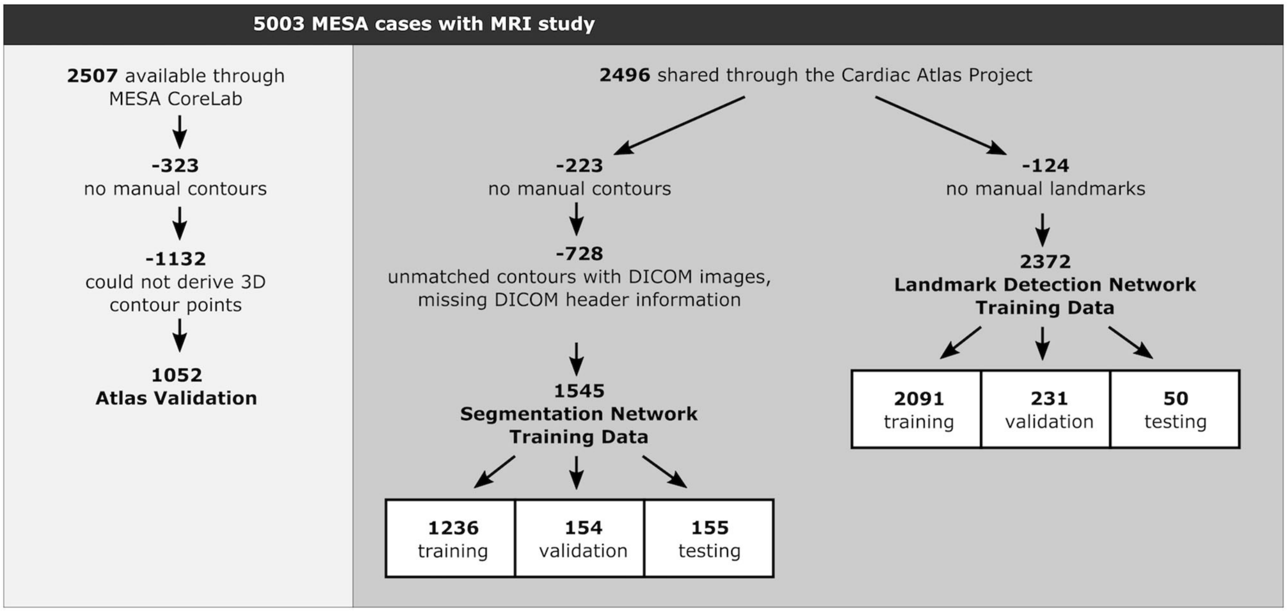

5003 MESA cases with MRI study
2507 available through MESA CoreLab
-323 no manual contours
-1132 could not derive 3D contour points
1052 Atlas Validation
2496 shared through the Cardiac Atlas Project
-223 no manual contours
-728 unmatched contours with DICOM images, missing DICOM header information
1545 Segmentation Network Training Data
1236 training
154 validation
155 testing
-124 no manual landmarks
2372 Landmark Detection Network Training Data
2091 training
231 validation
50 testing

FIGURE 2 | Division of MESA cases into two independent sets of Atlas Validation and Training sub-cohorts. Within the Training sub-cohort, cases were divided into training, validation and testing sub-groups for the different deep learning networks (Segmentation Network and Landmark Detection Network).

These were randomly split into 2,091 training cases (88%), 231 validation cases (10%), and 50 test cases (2%). The test cases were also used for inter-observer variability study, where landmark points from both analysts are available for each case.

For the LV atlas validation, we need cases where we can derive 3D points from the manual contours. Unfortunately, information about 3D image positions and orientations were not stored in the Q-MASS contour files available from the Core Lab. We therefore developed a simple matching algorithm to align Q-MASS contours with the DICOM image headers. This consisted of ordering the images and contours from apex to base, followed by alignment based on image position and orientation. The alignment results were manually reviewed to confirm correct matching of contours and images. This process resulted in 1,052 cases with manually verified DICOM image matching, sufficient to validate the automated pipeline developed in this study (see **Figure 2**).

Cardiac MRI Analysis Pipeline

As shown in **Figure 1**, the proposed automated CMR analysis combines two types of CNNs (myocardial segmentation and landmark detection) with LV finite element shape modeling. Cardiac landmark points were needed to determine the initial pose and orientation of the LV model, but were not part of the original MESA analysis protocol, hence further annotation was required to provide training data. The LV contours were required to guide the patient-specific customization of the LV model, and training data could be provided from the original MESA CMR analyses.

Landmark Detection Network

The landmark detection network was based on the VGGNet architecture (20), which has been successfully used to classify images and to recognize objects. It consists of 16 layers of CNN blocks that gradually extract features into smaller tensor size. The input is 256 × 256 MR image and the output is a feature vector of 2,048 elements. The final layer reduces this feature vector into four neurons corresponding to two points on the input image in $[x_1, y_1, x_2, y_2]$ format. Details of this landmark detection architecture are given in **Appendix A**.

Two types of anatomical landmarks are predicted for the proposed pipeline. The first landmark is the position of mitral valve hinge points at the intersection between the left atrium and the left ventricle from two long-axis MR images: two-chamber (2CH) and four-chamber (4CH) views. The other landmarks are the position of the intersection points between the right ventricle and the interventricular septum (RV insert points) from short-axis MR images. Mitral valve points were used to determine the basal extent of the heart, whereas RV insertion points were used to estimate the position of the septum.

Although sharing the same architecture, we trained three separate landmark detection networks to detect the different types of cardiac landmark points and image views: 2CH mitral valve points, 4CH mitral valve points and short-axis RV insert points. We developed a novel transfer learning scheme between these networks during training, which was designed to exploit similarities in the images, yet allowing for differences in the spatial relationships. First, an initial network for one view was trained from scratch with random weight initialization until convergence. Then, the network was retrained for one of the remaining two views. However, instead of using a random

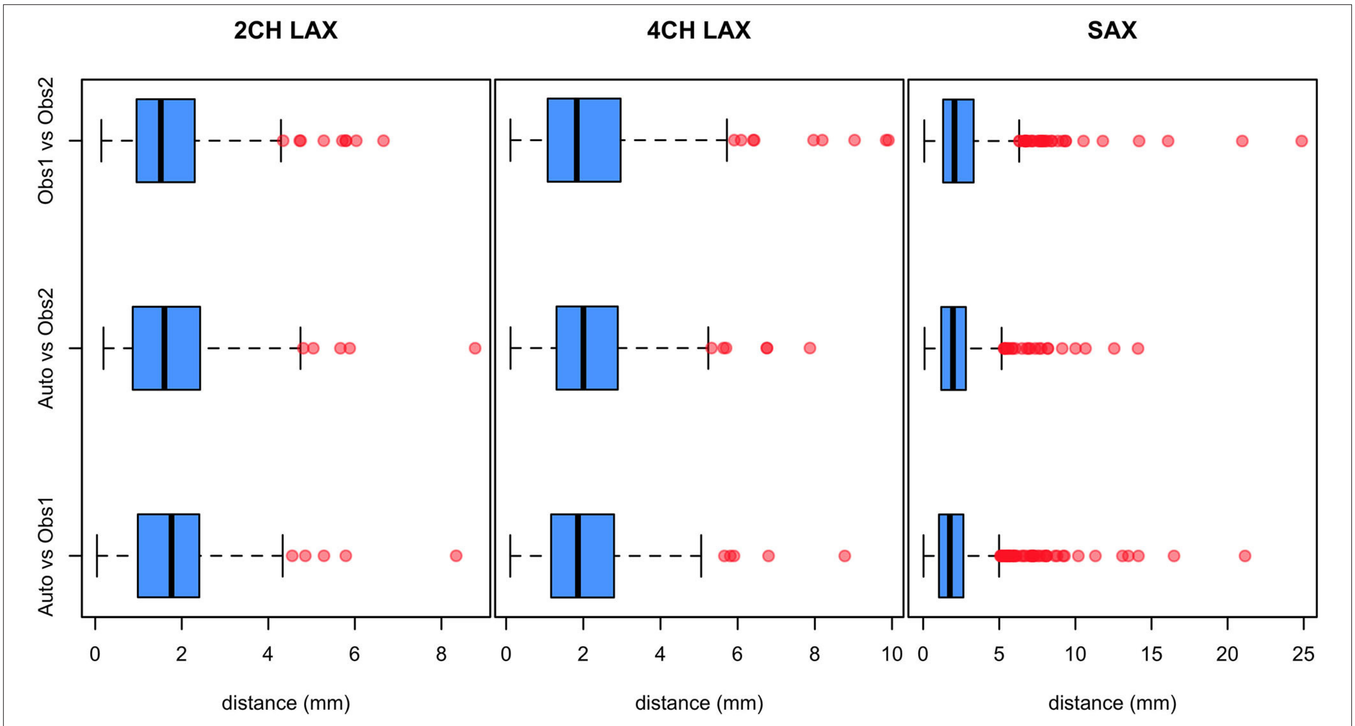


FIGURE 3 | Distributions of distances between landmark points identified by the landmark detection method (Auto) and the two analysts (Obs1 and Obs2). Median (solid line), quartiles (thin lines) outliers (red points).

TABLE 2 | Landmark distance errors from neural networks trained independently compared to networks trained with our training strategy.

	System trained with independent neural networks	System trained with our training strategy
Two-chamber	2.98 (1.44)	1.53 (0.74)
Four-chamber	3.24 (1.55)	1.44 (0.74)
Short axis	2.94 (1.6)	2.07 (1.11)

Error values were measured on 232 validation cases and shown as mean (standard deviation). All values are in millimeters.

initialization, the weights from the previous training step were used as initial weights. After the new network was converged, its weights were used as initialization for the third view. The order in which the three different views were trained was random. This sequence was repeated until convergence (e.g., the performance of the two-chamber network compared to the previously trained two-chamber network was not improved). An advantage of this sequence is that it allows for maximum freedom when training the neural networks for the different kinds of image views whilst still being able to infer features learned from other images. **Table 2** shows the improvement of performance using the transfer learning scheme in the validation set, where the landmark distance errors significantly decreased.

On average, we included five frames per case for the mitral valve points on each of the four-chamber and two-chamber views, and five short-axis slices per case for the RV inserts on the end-diastolic frame. In total, there were 11,604 images for the two-chamber view, 11,670 images for the four-chamber view, and 13,402 images for the short-axis view. Images were whitened by subtracting the mean pixel intensity and divided by standard deviation, on a per-image basis. Zero-padded cropping was performed to create 256 × 256 input images as needed.

We validated the predicted landmark points by the Euclidean distance (in mm) on the image space. The strength of agreement between the landmark detection and the two analysts was measured using the intraclass correlation coefficient (ICC) with a two-way random effects model (21). A high ICC (close to 1) indicates a high similarity between landmark point locations from all observers.

Segmentation Network

To segment the myocardium, we used the U-Net architecture (22), which has been successfully used in a wide range area of medical image analysis (12). The input is 256 × 256 short-axis MR image and the output is a mask image of the same size that consists of either myocardium, cavity or background pixel. The short-axis image was segmented individually; no temporal or other spatial multi-slice information was learned for this segmentation network. During training, data augmentation was performed by image flipping, zoom, brightness, and contrast variations. Input images were zero-padded and cropped into 256 × 256 image size as needed. More details about the segmentation network architecture and its training results are given in **Appendix B**.

We validated the accuracy of the segmentation network by using the Dice score (23), for both myocardium and LV cavity. We also validated standard clinical measurements for post-processing CMR exams (1), which include LV volumes at end-diastole and end-systole, ejection fraction and LV mass. Volumes were estimated by the LV cavity areas times the slice thickness (and slice gaps) for all short-axis slices where endocardial contours were available. LV masses were calculated from the myocardial volume (defined between endocardial and epicardial contours) multiplied by a density of 1.05 g/mL. All volumes and masses were indexed by body surface area, resulted in LV end-diastolic volume index (LVEDVi), LV end-systolic volume index (LVESVi), LV mass index (LVMi). Ejection fraction (LVEF) was measured by (LVEDVi – LVESVi) / LVEDVi * 100. We compared all these values from the test cases ($n = 155$) using the Bland-Altman plot analysis (24) to identify if there is a systematic error from the mean offset of the differences, inconsistent variability from the limits of agreement (mean ± 1.96 × standard deviation), and any trend of proportional error.

LV Atlas Construction

After landmark detection and segmentation (**Figure 1**), a finite element LV model was automatically customized to each set of myocardial contours and landmark points, as described previously in (25). Briefly, the LV model was first fitted to the landmark and contour points by a least squares optimization. The extent of the LV was defined from landmarks on mitral valve points and an LV apex point obtained from the contours. The septum area was located using the RV insertion landmark points. After orienting the model according to the landmarks, the endocardial and epicardial surfaces were fitted to the short axis contours by minimizing the distance between the surfaces and the contour points.

One advantage of using this LV model customization is that we can automatically correct image slice shifting due to breathing motion. In **Figure 1**, an example of this shifting artifact can be seen from the 3D contour points. The automatic breath-hold misregistration correction was based on (6). Briefly, a highly regularized customization of the LV was performed first to align a smooth LV model with the data. This model preserves the overall shape but is robust to breath-hold misalignments. Intersections between the LV model with short-axis image slices were then calculated and the contours were aligned with the model. The alignment movement was performed in-plane allowing only two degrees of freedom during shifting (no shift in the longitudinal direction). The shifting direction was calculated from the centroid of the intersection of the model with the image slice, based on the area-weighted average of the mesh barycenter. Then the LV model was re-customized to the data with a low regularization weight, minimizing the distance between the model and the contours.

After model fitting, an LV atlas was constructed by concatenating LV models from end-diastolic (ED) and end-systolic (ES) frames to capture both shape and motion information. In our previous study (7), concatenating ED and ES surface sample points yielded better performance to extract cardiac shape remodeling features compared to points from individual frames alone. Let N be the number of points sampled from the finite element model, and P_{endo_ED}, P_{epi_ED}, P_{endo_ES}, $P_{epi\_ES} \in \mathbb{R}^{Nx3}$ be 3D surface sampling points from the endocardium at ED, epicardium at ED, endocardium at ES and epicardium at ES, respectively. A single shape vector is defined by flattening each point matrix into $S = \left[x_1, y_1, z_1, \ldots, x_N, y_N, z_N\right]^T$vector and concatenating all of the four surfaces, resulting in $4 \times 3 \times N = 12N$ points. We removed position and orientation variations between shape vectors by using Procrustes alignment (26). The mean shape was then calculated and the principal component analysis (PCA) can be applied to the registered shape vectors.

Association With Cardiovascular Risk Factors

To demonstrate the clinical efficacy of the predicted LV atlas, we analyzed associations between LV shape and cardiovascular risk factors, i.e., hypertension, diabetes, smoking status, cholesterol level, and calcium score, and compared atlas associations obtained from the automatic pipeline with atlas associations obtained from manual contours and landmarks. For this evaluation, we evaluated 1,052 MESA cases independent of the sub-cohorts used to train the landmark and segmentation networks (the atlas validation dataset, **Table 1** and **Figure 2**).

Our hypothesis was that there is no significant differences in the strength of risk factor associations between the automatically generated LV atlas and the atlas derived from manual analyses. Logistic regression (LR) models were used to evaluate the strength of the risk factor associations. A separate LR model was generated for each risk factor using that factor as a binary univariate dependent variable and the first 20 principal component scores (90% total variance explained) derived from the atlas as the independent variables. Visual comparisons between modes of shape variations from LV Atlas derived from manual analyses and from the proposed cardiac MRI pipeline are available in the **Supplementary Files**. The strength of the association between shape and risk factor was quantified using the area under the curve of the receiver operating characteristic (AUC). To avoid overfitting, a ten-fold cross validation scheme was employed. At each cross validation iteration, we rebuilt the PCA from scratch to show that the associations were not dependent to a fixed orientation of the principal axes.

RESULTS

Landmark Detection

The total training time for three landmark detection networks was 14 h on NVidia Titan X Pascal GPU. Typically, five iterations of transfer learning between 2CH, 4CH, and SAX networks were required for overall convergence. The performance of the landmark detection networks was tested on 50 independent cases, which were annotated by two expert analysts independently. Only images where both analysts identified all landmark points were included. These resulted in 111 2CH, 107 4CH, and 286 SAX images for comparisons. Since two points are identified from each

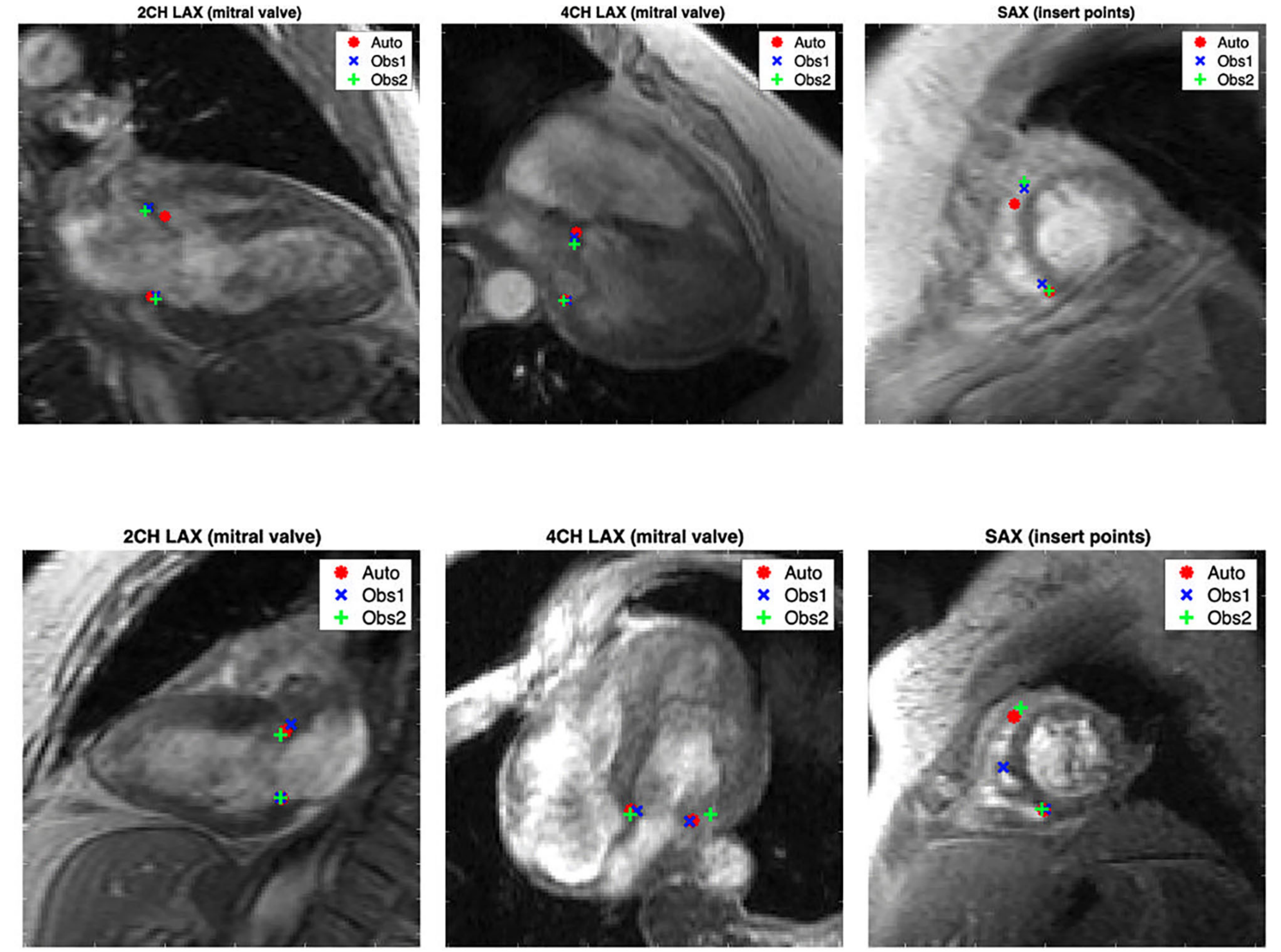


FIGURE 4 | Examples of automated landmark detection (red markers) compared with manually defined placements by two observers (blue and green markers). The top row shows cases with the maximum distance of automated detection to one of the observers while interobserver distances are small. The bottom row shows cases with the largest interobserver distances.

image, the total number of points during the test was 222, 214 and 572 points for 2CH, 4CH, and SAX, respectively.

The distributions of Euclidean distances between automated methods and the observers are shown in **Figure 3**. Mean, standard deviation, and maximum distances are given in **Table 3**. The results show that the automated landmark detection errors are within the inter-observer variabilities with no significant differences in the location of landmark points (all $p < 0.001$). ICC between the automated method and the two analysts were all excellent, i.e., 0.998, 0.996, and 0.995 for 2CH, 4CH, and SAX respectively.

Examples of landmark detections are shown in **Figure 4** together with manual expert observer placements. The top row images show the largest distance of the automated detection method where the distance between observers was low (< 3 pixels). Even in these cases, the automated method could identify the landmarks very close to the observers. The bottom row images in **Figure 4** showcase the largest distances between expert observers. The automated method was able to identify landmark points in these cases with the position very close to one of the observers. These cases show the difficulty of visually identifying landmark points where image contrast is low and high image noise is present.

TABLE 3 | Differences and intraclass correlation (ICC) values in detecting landmarks on 50 validation cases.

	2CH LAX *N* = 222	4CH LAX *N* = 214	SAX *N* = 572
Auto vs. Obs1	1.86 (1.19)	2.09 (1.32)	2.29 (2.15)
Auto vs. Obs2	1.81 (1.21)	2.19 (1.28)	2.27 (1.61)
Obs1 vs. Obs2	1.78 (1.16)	2.24 (1.68)	2.67 (2.29)
ICC value	0.998	0.996	0.995

All difference values are expressed mean (standard deviation) from the Euclidean distance between annotations in millimeters. N is the number of cases.

Segmentation

Quartiles, means, and standard deviations of the Dice score from the test dataset are presented in **Table 4**. Median and mean Dice scores were high (>0.8) for myocardium and LV cavity masks, both at ED and ES frames. Typical segmentation results are shown in **Figure 5** with cases of best, mean, and worst results. **Figure 5** also demonstrates the difficulty of segmenting basal slices near the LV outflow tract.

Table 5 shows comparisons of volumes (LVEDVi and LVESVi), mass (LVMi) and ejection fraction (LVEF) from the test cases. The segmentation network achieved excellent correlation coefficients for all clinical measurements (all Pearson's coefficients are >0.9, $p < 0.001$). The mean offset of differences are also small, i.e., <1 mL/m^2 for volumes, only 0.7% for ejection fraction, and 3 g/m^2 for mass. As shown in **Figure 6**, the differences are consistent within the limit of agreement lines without any visible trend for proportional error.

Atlas Validation

Finally, we compared cardiovascular risk factor associations from the LV atlas from the automated analysis pipeline with an atlas formed from the manual analyses using a similar analysis method to (25). **Table 6** shows the comparison of the area under the receiver operating characteristic curves (AUC) from risk factor association results (test cases from the cross validation). From all risk factors (hypertension, diabetes, smoking status, cholesterol, and calcium score), none of them have significant differences between the two methods except for cholesterol ($p = 0.02$) which

TABLE 4 | Dice score results of the segmentation network from the test dataset with 2,465 images.

Mask	Frame	Q1	Median	Q3	Mean	Std dev
Cavity	ED	0.92	0.95	0.97	0.93	0.07
	ES	0.86	0.91	0.94	0.88	0.11
Myocardium	ED	0.85	0.89	0.91	0.87	0.07
	ES	0.89	0.92	0.94	0.90	0.08

Frames indicate end-diastole (ED) and end-systole (ES). The 25th quartile (Q1), median, and 75th quartile (Q3) are shown, together with means and standard deviations.

TABLE 5 | Comparisons of indexed LV volumes, ejection fraction and mass from the 155 test cases between the predicted segmentation results with manual contours.

LV function	Correlation coefficient	Differences
LVEDVi (mL/m^2)	0.98 ($p < 0.001$)	−0.02 (2.6)
LVESVi (mL/m^2)	0.95 ($p < 0.001$)	−0.46 (2.3)
LVEF (%)	0.92 ($p < 0.001$)	0.69 (3.3)
LVMi (g/m^2)	0.92 ($p < 0.001$)	3.0 (6.4)

The differences are written as mean (standard deviation).

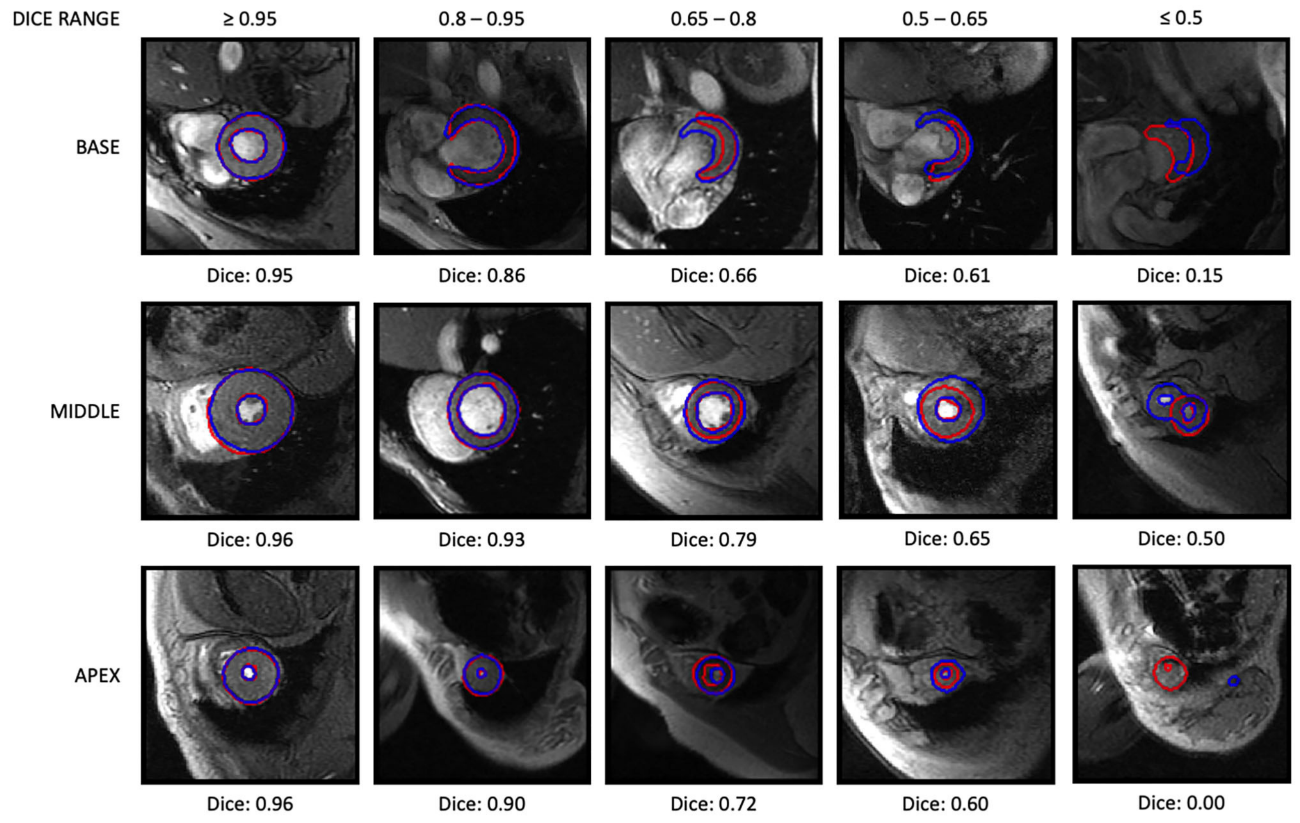


FIGURE 5 | Examples of short axis segmentation network results. Top row, base; middle row, mid-ventricle; bottom row, apex. Manual contours are in red while automated contours are in blue. A range of Dice score results are shown.

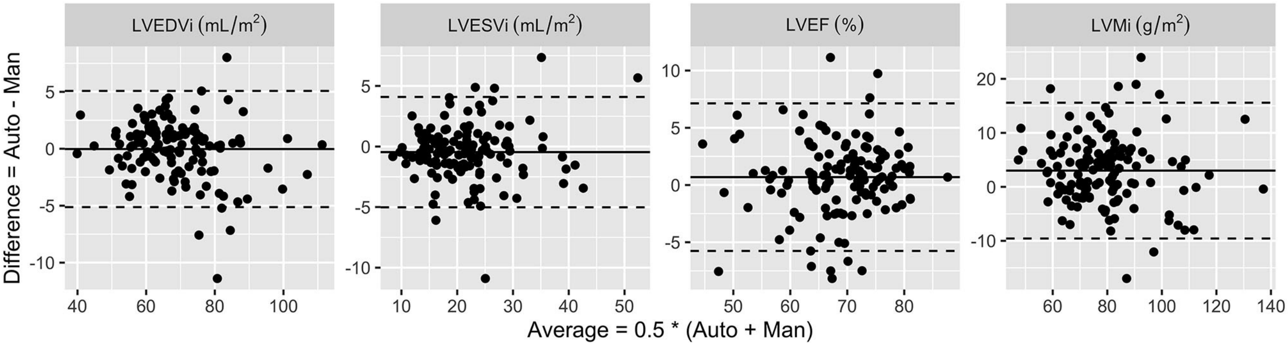


FIGURE 6 | Differences between automated analysis (Auto) and manually drawn contours (Man). Solid lines are mean differences and dashed lines are the limits of agreement within ±1.96 × standard deviation from the mean. The mean difference values are shown in **Table 5**.

TABLE 6 | Area under the ROC curve (AUC) comparisons from the 1,052 LV shape association studies using different contours: manual (Man) and deep learning (Auto).

	AUC		*P*-value
	Man	Auto	
Hypertension	0.69	0.71	0.22
Diabetes	0.56	0.53	0.34
Smoking status	0.59	0.61	0.33
Cholesterol	0.50	0.54	0.02
Calcium score	0.61	0.61	0.99

showed a stronger association with the automated analysis than with the manual analysis.

DISCUSSION

In this study, we present methods for the automated analysis of large cohort data from a legacy dataset obtained in the MESA study, aided by deep learning methods. These methods enable a more complete analysis of large cohort datasets, augmenting the parameter set available from these valuable studies. In addition to the end-diastolic and end-systolic volumes computed in the original study, these methods enable the analysis of 3D shapes, facilitating a fully automated 3D model-based atlas analysis method. Almost all risk factors showed similar strength of relationships with atlas scores, except for cholesterol level in which the automated method showed a stronger relationship (**Table 5**). However with AUC around 0.50, the elevated cholesterol association was essentially random. The slightly higher AUC for the automated contours may indicate that some signal may be available in the automated analysis which was lost in the manual analysis. This requires more research using a larger cohort.

The automated landmark detection method was successfully applied to GRE images, which are known to have lower signal-to-noise ratio and lower contrast compared to the current standard steady state free precession CMR imaging methods (5). The agreements with two expert analysts were all excellent (ICC > 0.9). Since signal-to-noise ratio is low in some gradient echo images, the analysts had noticeable disagreements between them in some cases, as shown in **Figure 4** (bottom row). However, the automated detection method could identify the location of the landmark point in agreement with one of the observers. This ability was achieved by our approach to transfer learning weight parameters between image views iteratively. We exploited features between different domains to make the detection robust to noise and other artifacts.

Other machine learning methods have reported good results with landmark detection in cardiac MRI data, as well. For instance, Tarroni et al. (27) applied a hybrid random forest approach integrating both regression and structured classification networks and reported mean errors of 3.2–3.9 mm in mitral valve landmark detection. Although it is difficult to determine which methods give the "best performance" in this application, our results show that the CNN-based method is powerful enough in the applications where legacy datasets provide sufficient annotated cases.

For the segmentation task, we demonstrated that the popular U-Net architecture (22) without any major modifications is capable of providing acceptable segmentation of the myocardium in gradient echo cine images. The segmentation network, which was trained based only on individual SAX images (without temporal information), has already achieved excellent performance. The first quartiles of the Dice score were all above 0.85 (**Table 4**), and 92% of the Dice scores were above 0.80. From the test dataset, the network only failed to segment one slice and only 8 images with Dice scores <0.5. All of these slices were the apical slices, where blood cavity is hardly recognizable even by visual inspection. Other problematic slices were at the base around the outflow tract, where there are more variability of the contours at the aortic root. **Figure 5** shows some examples of the segmentation results at different levels of the LV (base, middle and apex) with variations of the Dice scores. Although apical and basal slices were more difficult for the network, the LV shape customization method was relatively robust to segmentation mask outliers, as evidenced by the agreement in statistical relationships with common risk

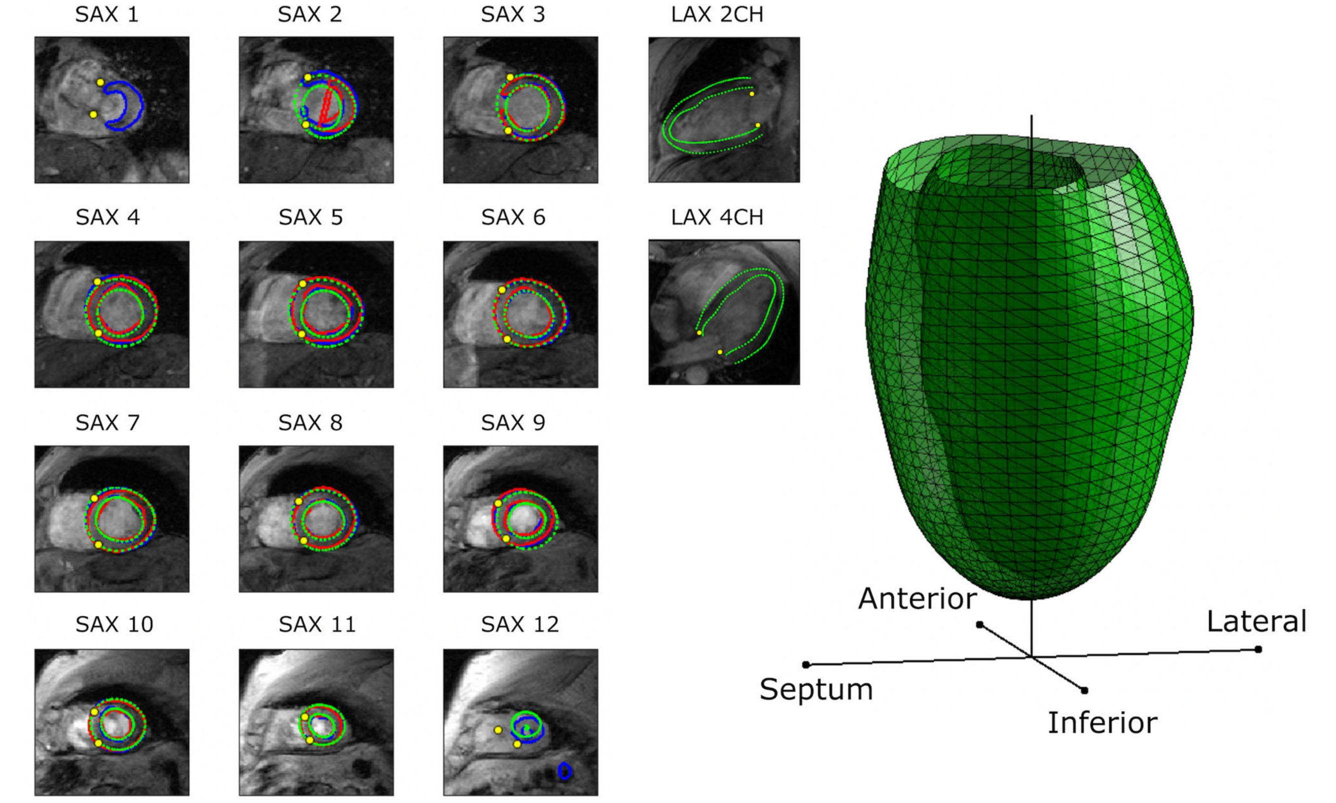


FIGURE 7 | An example of fully automated CMR pipeline result as a patient-specific LV model. Intermediate predictions of the myocardial contours (in blue) and landmark points (yellow circles) are shown in each corresponding DICOM image. Manual contours are shown in red. The intersection contours between the 3D LV model with the images are shown in green. This particular example demonstrates how failed segmentation contours (in apex and base slices) do not affect the final LV model, which are clearly shown in the LAX intersection contours.

factors, since the model customization process used data from all slices. **Figure 7** demonstrates the benefit of the LV model customization over large errors predicted in some problematic slices. This is shown by the intersection contours of the LV model with the SAX images that are well aligned with the manual contours. **Figure 7** also shows the intersection contours on LAX images where the alignment of the contours at the myocardium can be visually assessed.

It is known that different groups annotate cardiac MRI data differently (28). For this study, the manual contours were performed by a single core lab, whereas the landmarks were performed in another core lab, so both the landmark detection and segmentation networks will reflect the core lab standard operating procedures on the gradient echo images. Differences in local shape are expected when comparing the shape models generated with gradient echo imaging with those generated from other protocols, and these can be corrected using atlas-based methods (29). Alternatively, the training data distribution can be made richer to include more pathologies, images from different centers and multiple observers, as has been demonstrated by Tao et al. (30) and Bhuva et al. (31).

A common approach to train a complex deep learning network is by end-to-end training (32, 33), where a combined loss function is defined for multiple tasks as the global cost function to optimize. In this work, landmarks and contours were only available on separated image views, so we decided to train the landmark detection network separately to the segmentation network to make each network capable of predicting unseen images independently. The ability to identify mitral valve points therefore does not need to depend on the segmentation masks or vice versa.

The problem of missing information is common to legacy datasets such as MESA. In this study, information linking contours with the corresponding 3D image position was not available. Since most cases were able to be matched with a simple algorithm, leading to sufficient training data, we did not invest more time in developing more sophisticated image-contour matching algorithms. The 3D conversions failed mainly due to missing 3D position information in the DICOM header or missing trigger time information needed to sort the images temporally. To investigate whether there was any bias due to poor image quality, we examined the image quality score given by the original Core Lab readers. This was a three-level subjective rating: 1 for good, 2 for moderate and 3 for poor. There were no significant differences between included and excluded cases ($p = 0.4$, Fisher's Exact test), with 85.4% vs. 86.3% having score

1, 14.6% vs. 13.5% for score 2 and 0% vs. 0.2% for score 3, for included vs. excluded cases, respectively. LV wall motion was also scored on a three point scale and there were no differences between included and excluded cases.

Although this study specifically trained deep learning networks for old legacy gradient echo (GRE) cine images, there are some clinical applications employing GRE imaging. In a recent guideline (34), the image quality of GRE images is better than that of the current steady-state free precession (SSFP) cine images for patients with cardiac implantable electronic devices (35). GRE images are also preferred for T1 and T2-weighted images particularly for patients with suspected iron overload (36). Hence the proposed CMR analysis pipeline has a wider application in other cardiac imaging studies as well, albeit transfer learning is needed to adapt the learned weight parameters to specific pathology. Note that the pipeline does not depend only on GRE; it can be applied directly to other types of CMR images, particularly where legacy datasets can provide valuable additional data.

CONCLUSIONS AND FUTURE WORK

We have shown that deep learning networks can be used for automatically finding LV landmarks and segmentations on legacy MESA CMR images, in order to automate the construction of LV models, which can be used to build an atlas and evaluate associations between LV shape and risk factors. The final prediction of the LV model based on deep learning networks had similar power to evaluate associations with cardiovascular risk factors compared to manual analysis. This has greatly reduced the amount of time to analyze large-scale collections of cardiac MRI study. In future work, the automated atlas will be used to derive associations between LV shape and outcomes. In addition, analysis of all frames in the cine will allow the calculation of ejection and filling rates and other dynamic information.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study are available on request from the Cardiac Atlas Project (www.cardiacatlas.org). Codes will be available from https://github.com/orgs/CardiacAtlasProject/repositories.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Johns Hopkins University School of Medicine (NA 00031350) and New Zealand Multiregion Ethics Committee (MEC/08/04/052). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AS, CM, BA-V, KG, and AY designed the overall study and performed the final analysis. MJ developed, trained, and validated the landmark detection network. LSH, JDG, and CWe developed, trained, and validated the segmentation network. KG and CM developed and validated the left ventricular fitting method. AS, KG, CM, and AY processed the pipeline, applied the trained networks into the remaining MESA cohort, and performed the independent atlas validation. BA-V, JL, CWu, and DB assessed the final validation results. All authors participated in the analysis, interpretation of data, drafting of the manuscript, revising it critically, and final approval of the submitted manuscript.

FUNDING

This research was funded by the Health Research Council of New Zealand (17/608 and 17/234). MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA was provided by contracts N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, and CTSA UL1-RR-024156.

ACKNOWLEDGMENTS

We would like to thank Benjamin Wen and Augustin Okamura for their analysis of the data. We also acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU and the Center for eResearch at the University of Auckland for facilitating a GPU virtual machine used for this research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcvm.2021.807728/full#supplementary-material

REFERENCES

1. Schulz-Menger J, Bluemke DA, Bremerich J, Flamm SD, Fogel MA, Friedrich MG, et al. Standardized image interpretation and post-processing in cardiovascular magnetic resonance-—2020 update: Society for Cardiovascular Magnetic Resonance (SCMR): Board of Trustees Task Force on Standardized Post-Processing. *J Cardiovasc Magn Reson*. (2020) 22:19. doi: 10.1186/s12968-020-00610-6
2. Bild DE. Multi-ethnic study of atherosclerosis: objectives and design. *Am J Epidemiol.* (2002) 156:871–81. doi: 10.1093/aje/kwf113
3. Petersen SE, Matthews PM, Bamberg F, Bluemke DA, Francis JM, Friedrich MG, et al. Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of UK Biobank—rationale, challenges and approaches. *J Cardiovasc Magn Reson.* (2013) 15:46. doi: 10.1186/1532-429X-15-46

4. Kramer CM, Barkhausen J, Bucciarelli-Ducci C, Flamm SD, Kim RJ, Nagel E. Standardized cardiovascular magnetic resonance imaging (CMR) protocols: 2020 update. *J Cardiovasc Magn Reson*. (2020) 22:17. doi: 10.1186/s12968-020-00607-1
5. Malayeri AA, Johnson WC, Macedo R, Bathon J, Lima JAC, Bluemke DA. Cardiac cine MRI: Quantification of the relationship between fast gradient echo and steady-state free precession for determination of myocardial mass and volumes. *J Magn Reson Imaging.* (2008) 28:60–6. doi: 10.1002/jmri.21405
6. Medrano-Gracia P, Cowan BR, Ambale-Venkatesh B, Bluemke DA, Eng J, Finn JP, et al. Left ventricular shape variation in asymptomatic populations: the multi-ethnic study of atherosclerosis. *J Cardiovasc Magn Reson*. (2014) 16:57. doi: 10.1186/s12968-014-0056-2
7. Zhang X, Cowan BR, Bluemke DA, Finn JP, Fonseca CG, Kadish AH, et al. Atlas-based quantification of cardiac remodeling due to myocardial infarction. *PLoS ONE.* (2014) 9:e110243. doi: 10.1371/journal.pone.0110243
8. Suinesiaputra A, Dhooge J, Duchateau N, Ehrhardt J, Frangi AF, Gooya A, et al. Statistical shape modeling of the left ventricle: myocardial infarct classification challenge. *IEEE J Biomed Health Inform.* (2018) 22:503–15. doi: 10.1109/JBHI.2017.2652449
9. Piras P, Teresi L, Puddu PE, Torromeo C, Young AA, Suinesiaputra A, et al. Morphologically normalized left ventricular motion indicators from MRI feature tracking characterize myocardial infarction. *Sci Rep.* (2017) 7:12259. doi: 10.1038/s41598-017-12539-5
10. Albà X, Lekadir K, Pereañez M, Medrano-Gracia P, Young AA, Frangi AF. Automatic initialization and quality control of large-scale cardiac MRI segmentations. *Med Image Anal.* (2018) 43:129–41. doi: 10.1016/j.media.2017.10.001
11. Leiner T, Rueckert D, Suinesiaputra A, Baeßler B, Nezafat R, Išgum I, Young AA. Machine learning in cardiovascular magnetic resonance: basic concepts and applications. *J Cardiovasc Magn Reson.* (2019) 21:61. doi: 10.1186/s12968-019-0575-y
12. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. Survey on deep learning in medical image analysis. *Med Image Anal.* (2017) 42:60–88. doi: 10.1016/j.media.2017.07.005
13. Shin H-C, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging.* (2016) 35:1285–98. doi: 10.1109/TMI.2016.2528162
14. Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE Trans Med Imaging.* (2016) 35:1207–16. doi: 10.1109/TMI.2016.2535865
15. Minaee S, Boykov YY, Porikli F, Plaza AJ, Kehtarnavaz N, Terzopoulos D. Image segmentation using deep learning: a survey. *IEEE Trans Pattern Anal Mach Intell*. (2021). doi: 10.1109/TPAMI.2021.3059968. [Epub ahead of print].
16. Fu Y, Lei Y, Wang T, Curran WJ, Liu T, Yang X. Deep learning in medical image registration: a review. *Phys Med Biol.* (2020) 65:20TR01. doi: 10.1088/1361-6560/ab843e
17. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, et al. Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Trans Med Imaging.* (2016) 35:1299–312. doi: 10.1109/TMI.2016.2535302
18. Natori S, Lai S, Finn JP, Gomes AS, Hundley WG, Jerosch-Herold M, et al. Cardiovascular function in multi-ethnic study of atherosclerosis: normal values by age, sex, and ethnicity. *Am J Roentgenol.* (2006) 186:S357–65. doi: 10.2214/AJR.04.1868
19. Fonseca CG, Backhaus M, Bluemke DA, Britten RD, Chung JD, Cowan BR, et al. The Cardiac Atlas Project—an imaging database for computational modeling and statistical atlases of the heart. *Bioinformatics.* (2011) 27:2288–95. doi: 10.1093/bioinformatics/btr360
20. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Bengio Y, LeCun Y, editors. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Available at: http://arxiv.org/abs/1409.1556 (accessed November 10, 2021).
21. Koo TK, Li MY. A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med.* (2016) 15:155–63. doi: 10.1016/j.jcm.2016.02.012
22. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. Cham: Springer International Publishing. p. 234–41. doi: 10.1007/978-3-319-24574-4_28
23. Eelbode T, Bertels J, Berman M, Vandermeulen D, Maes F, Bisschops R, et al. Optimization for medical image segmentation: theory and practice when evaluating with dice score or jaccard index. *IEEE Trans Med Imaging.* (2020) 39:3679–90. doi: 10.1109/TMI.2020.3002417
24. Bland JM, Altman DG. Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. *Lancet.* (1986) 327:307–10. doi: 10.1016/S0140-6736(86)90837-8
25. Gilbert K, Bai W, Mauger C, Medrano-Gracia P, Suinesiaputra A, Lee AM, et al. Independent left ventricular morphometric atlases show consistent relationships with cardiovascular risk factors: a UK Biobank Study. *Sci Rep.* (2019) 9:1130. doi: 10.1038/s41598-018-37916-6
26. Dryden IL, Mardia KV. *Statistical Shape Analysis*. Chichester, New York: John Wiley & Sons (1998).
27. Tarroni G, Bai W, Oktay O, Schuh A, Suzuki H, Glocker B, et al. Large-scale quality control of cardiac imaging in population studies: application to UK Biobank. *Sci Rep.* (2020) 10:2408. doi: 10.1038/s41598-020-58212-2
28. Suinesiaputra A, Bluemke DA, Cowan BR, Friedrich MG, Kramer CM, Kwong R, et al. Quantification of LV function and mass by cardiovascular magnetic resonance: multi-center variability and consensus contours. *J Cardiovasc Magn Reson*. (2015) 17:63. doi: 10.1186/s12968-015-0170-9
29. Medrano-Gracia P, Cowan BR, Bluemke DA, Finn JP, Kadish AH, Lee DC, et al. Atlas-based analysis of cardiac shape and function: correction of regional shape bias due to imaging protocol for population studies. *J Cardiovasc Magn Reson.* (2013) 15:80. doi: 10.1186/1532-429X-15-80
30. Tao Q, Yan W, Wang Y, Paiman EHM, Shamonin DP, Garg P, et al. Deep learning–based method for fully automatic quantification of left ventricle function from cine MR images: a multivendor, multicenter study. *Radiology.* (2019) 290:81–8. doi: 10.1148/radiol.2018180513
31. Bhuva AN, Bai W, Lau C, Davies RH, Ye Y, Bulluck H, et al. A Multicenter, scan-rescan, human and machine learning CMR study to test generalizability and precision in imaging biomarker analysis. *Circ Cardiovasc Imaging.* (2019) 12:e009214. doi: 10.1161/CIRCIMAGING.119.009214
32. Song L, Lin J, Wang ZJ, Wang H. An end-to-end multi-task deep learning framework for skin lesion analysis. *IEEE J Biomed Health Inform.* (2020) 24:2912–21. doi: 10.1109/JBHI.2020.2973614
33. Yap MH, Goyal M, Osman FM, Martí R, Denton E, Juette A, et al. Breast ultrasound lesions recognition: end-to-end deep learning approaches. *J Med Imaging Bellingham Wash.* (2019) 6:011007.
34. Paterson ID, White JA, Butler CR, Connelly KA, Guerra PG, Hill MD, et al. 2021 Update on safety of magnetic resonance imaging: joint statement from Canadian Cardiovascular Society/Canadian Society for Cardiovascular Magnetic Resonance/Canadian Heart Rhythm Society. *Can J Cardiol.* (2021) 37:835–47. doi: 10.1016/j.cjca.2021.02.012
35. Schwitter J, Gold MR, Al Fagih A, Lee S, Peterson M, Ciuffo A, et al. Image quality of cardiac magnetic resonance imaging in patients with an implantable cardioverter defibrillator system designed for the magnetic resonance imaging environment. *Circ Cardiovasc Imaging.* (2016) 9:e004025. doi: 10.1161/CIRCIMAGING.115.004025
36. Serai SD, Trout AT, Fleck RJ, Quinn CT, Dillman JR. Measuring liver T2* and cardiac T2* in a single acquisition. *Abdom Radiol N Y.* (2018) 43:2303–8. doi: 10.1007/s00261-018-1477-4

Author Disclaimer: The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 07 April 2022
doi: 10.3389/fcvm.2022.859310

Fairness in Cardiac Magnetic Resonance Imaging: Assessing Sex and Racial Bias in Deep Learning-Based Segmentation

Esther Puyol-Antón[1], Bram Ruijsink[1,2,3], Jorge Mariscal Harana[1], Stefan K. Piechnik[4], Stefan Neubauer[4], Steffen E. Petersen[5,6,7,8], Reza Razavi[1,2], Phil Chowienczyk[1,9] and Andrew P. King[1]*

[1] School of Biomedical Engineering and Imaging Sciences, King's College London, London, United Kingdom, [2] Department of Adult and Paediatric Cardiology, Guy's and St Thomas' NHS Foundation Trust, London, United Kingdom, [3] Division of Heart and Lungs, Department of Cardiology, University Medical Centre Utrecht, Utrecht, Netherlands, [4] Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, United Kingdom, [5] National Institute for Health Research (NIHR) Barts Biomedical Research Centre, William Harvey Research Institute, Queen Mary University London, London, United Kingdom, [6] Barts Heart Centre, St Bartholomew's Hospital, Barts Health NHS Trust, London, United Kingdom, [7] Health Data Research UK, London, United Kingdom, [8] Alan Turing Institute, London, United Kingdom, [9] British Heart Foundation Centre, King's College London, London, United Kingdom

OPEN ACCESS

Edited by:
Chayakrit Krittanawong,
NYU Grossman School of Medicine,
United States

Reviewed by:
Vivek Pinakin Jani,
The Johns Hopkins Hospital, Johns Hopkins Medicine, United States
Feng Yang,
National Institutes of Health (NIH),
United States
Anna Baritussio,
University of Padua, Italy
Yuki Mori,
University of Copenhagen, Denmark

***Correspondence:**
Esther Puyol-Antón
esther.puyol_anton@kcl.ac.uk

Specialty section:
This article was submitted to Cardiovascular Imaging, a section of the journal Frontiers in Cardiovascular Medicine

Received: *21 January 2022*
Accepted: *02 March 2022*
Published: *07 April 2022*

Citation:
Puyol-Antón E, Ruijsink B, Mariscal Harana J, Piechnik SK, Neubauer S, Petersen SE, Razavi R, Chowienczyk P and King AP (2022) Fairness in Cardiac Magnetic Resonance Imaging: Assessing Sex and Racial Bias in Deep Learning-Based Segmentation. Front. Cardiovasc. Med. 9:859310. doi: 10.3389/fcvm.2022.859310

Background: Artificial intelligence (AI) techniques have been proposed for automation of cine CMR segmentation for functional quantification. However, in other applications AI models have been shown to have potential for sex and/or racial bias. The objective of this paper is to perform the first analysis of sex/racial bias in AI-based cine CMR segmentation using a large-scale database.

Methods: A state-of-the-art deep learning (DL) model was used for automatic segmentation of both ventricles and the myocardium from cine short-axis CMR. The dataset consisted of end-diastole and end-systole short-axis cine CMR images of 5,903 subjects from the UK Biobank database (61.5 ± 7.1 years, 52% male, 81% white). To assess sex and racial bias, we compared Dice scores and errors in measurements of biventricular volumes and function between patients grouped by race and sex. To investigate whether segmentation bias could be explained by potential confounders, a multivariate linear regression and ANCOVA were performed.

Results: Results on the overall population showed an excellent agreement between the manual and automatic segmentations. We found statistically significant differences in Dice scores between races (white ∼94% vs. minority ethnic groups 86–89%) as well as in absolute/relative errors in volumetric and functional measures, showing that the AI model was biased against minority racial groups, even after correction for possible confounders. The results of a multivariate linear regression analysis showed that no covariate could explain the Dice score bias between racial groups. However, for the Mixed and Black race groups, sex showed a weak positive association with the Dice score. The results of an ANCOVA analysis showed that race was the main factor that can explain the overall difference in Dice scores between racial groups.

Conclusion: We have shown that racial bias can exist in DL-based cine CMR segmentation models when training with a database that is sex-balanced but not race-balanced such as the UK Biobank.

Keywords: cardiac magnetic resonance, deep learning, fair AI, segmentation, inequality fairness in deep learning-based CMR segmentation

INTRODUCTION

Artificial intelligence (AI) is a rapidly evolving field in medicine, especially cardiology. AI has the potential to aid cardiologists in making better decisions, improving workflows, productivity, cost-effectiveness, and ultimately patient outcomes (1). Deep learning (DL) is a recent advance in AI which allows computers to learn a task using data instead of being explicitly programmed. Several studies in cardiology and other applications have shown that DL methods can match or even exceed human experts in tasks such as identifying and classifying disease (2–4).

In cardiology, cardiovascular imaging has a pivotal role in diagnostic decision making. Cardiac magnetic resonance (CMR) is the established non-invasive gold-standard modality for quantification of cardiac volumes and ejection fraction (EF). For decades, clinicians have been relying on manual or semi-automatic segmentation approaches to trace the cardiac chamber contours. However, manual expert segmentation of CMR images is tedious, time-consuming and prone to subjective errors. Recently, DL models have shown remarkable success in automating many medical image segmentation tasks. In cardiology, human-level performance in segmenting the main structures of the heart has been reported (5, 6), and researchers have proposed to use these models for tasks such as automating cardiac functional quantification (7). These methods are now starting to move toward broader clinical translation.

In the vast majority of cardiovascular diseases (CVDs), there are known associations between sex/race and epidemiology, pathophysiology, clinical manifestations, effects of therapy, and outcomes (8–10). Furthermore, in clinically asymptomatic individuals the Multi-Ethnic Study of Atherosclerosis (MESA) study showed that men had greater right ventricular (RV) mass and larger RV volumes than women, but had lower RV ejection fraction; African-Americans had lower RV mass than whites, whereas Hispanics had higher RV mass (11); and the LV was more trabeculated in African-American and Hispanic participants than white participants, and smoothest in Chinese-American participants (12), but the greater extent of LV trabeculation was not associated with an absolute decline in LVEF during the approximately 10 years of the MESA study. Similarly, the Coronary Artery Risk Development in Young Adults (CARDIA) study (13) showed differences between races (African American and white) and sexes in LV systolic and diastolic function, which persist after adjustment for established cardiovascular risk factors.

Although these physiological differences are associations and not proven causative links with race/gender, their presence raises a potential concern about the performance of AI models in cardiovascular imaging. Although AI has great potential in this area, no previous work has investigated the fairness of such models. In AI, the concept of "fairness" refers to assessing AI algorithms for potential bias based on demographic characteristics such as race and sex. In general, AI models are trained agnostic to demographic characteristics, and they assume that if the model is unaware of these characteristics while making decisions, the decisions will be fair. However, we have recently shown, for the first time, that using this assumption there exists racial bias in DL-based cine CMR segmentation models when trained using racially imbalanced data (14). The previous study aimed to identify the presence of bias and the technical development of different bias mitigation strategies, in order to reduce the bias effect between different racial groups. The object of this study is to investigate in more detail the origin and the effect of this bias on cardiac structure and function and to assess whether the bias could be explained by any confounder and therefore be linked with changes in subject characteristics, anatomy or cardiovascular risk factors.

MATERIALS AND METHODS

Participants

The UK Biobank is a prospective cohort study with more than 500,000 participants aged 40–69 years of age conducted in the United Kingdom (15). This study complies with the Declaration of Helsinki; the work was covered by the ethical approval for UK Biobank studies from the NHS National Research Ethics Service on 17th June 2011 (Ref 11/NW/0382) and extended on 18th June 2021 (Ref 21/NW/0157) with written informed consent obtained from all participants. The present study was performed using a sub-cohort of the UK Biobank imaging database, for whom CMR imaging and ground truth manual segmentations were available. In this study, in order to minimize the effects of physiological differences due to cardiovascular and other related diseases, we only focus on the healthy population of the UK Biobank database and analyze possible confounders that can explain racial and sex bias.

Therefore, we excluded any subjects with known cardiovascular disease, respiratory disease, hematological disease, renal disease, rheumatic disease, malignancies, symptoms of chest pain, respiratory symptoms or other diseases impacting the cardiovascular system, except for diabetes mellitus, hypercholesterolemia and hypertension (see all exclusion criteria in **Supplementary List 1**). We included these cardiovascular risk factors to evaluate if or to what degree different cardiovascular risk in otherwise healthy patients could explain a potential bias in segmentation performance. We used the ICD-9 and ICD-10 codes and self-reported detailed health questionnaires and medication history for the selection process.

In this paper, race was assumed to align with self-reported ethnicity, which was the data collected in the UK Biobank. From the total UK Biobank database (N = 501,642), the race distribution is as follows: White 94.3%, Mixed 0.6%, Asian, 1.9%, Black 1.6%, Chinese: 0.9%, Other: 0.4%. The UK Biobank cohort has a similar ethnic distribution to the national population of the same age range in the 2011 UK Census (16). The imaging cohort used in this study (N = 5,660) has a slightly different racial distribution (White 81%, Mixed 3%, Asian, 7%, Black 4%, Chinese: 2%, Other: 3%), but it is still predominantly White race, in line with the full cohort of the UK Biobank database. Imaging centers of the UK Biobank are in Newcastle upon Tyne, Stockport, Reading and Bristol. The same imaging protocol was used in all imaging centers and no racial distribution difference was found between them. More details of the image acquisition protocol can be found in Petersen et al. (17).

Subject characteristics obtained were age, binary sex category, race, body measures (height; weight; body mass index, BMI; and body surface area, BSA), and smoker status (smoker was defined as a subject smoking or smoked daily for over 25 years in the previous 35 years). We also obtained the average heart rate (HR) and brachial systolic and diastolic blood pressure (SBP and DBP) measured during the CMR exam. These subject characteristics were considered as possible confounders in the statistical analysis, as they are directly or indirectly related to the measurements made and therefore plausibly associated with the accuracy of the measurements.

Automated Image Analysis

A state-of-the-art DL based segmentation model, the "nnU-Net" framework (18), was used for automatic segmentation of the left ventricle blood pool (LVBP), left ventricular myocardium (LVMyo) and right ventricle blood pool (RVBP) from cine short-axis CMR slices at end-diastole (ED) and end-systole (ES). This model was chosen as it has performed well across a range of segmentation challenges and was the top-performing model in the "ACDC" CMR segmentation challenge (6). For training and testing the segmentation model, we used a random split of 4,410 and 1,250 subjects, respectively, each with similar sex and racial distributions. We refer the reader to our previous paper (14) for further details of the model architecture and training.

Evaluation of the Method

For quantitative assessment of the image segmentation model, we used the Dice similarity coefficient (DSC), which quantifies the overlap between an automated segmentation and a ground truth segmentation. DSC has values between 0 and 100%, where 0 denotes no overlap, and 100% denotes perfect agreement. From the manual and automated image segmentations, we calculated the LV end-diastolic volume (LVEDV) and end-systolic volume (LVESV), and RV end-diastolic volume (RVEDV) and end-systolic volume (RVESV) by summing the number of voxels belonging to the corresponding label classes in the segmentation and multiplying this by the volume per voxel. The LV myocardial mass (LVmass) was calculated by multiplying the LV myocardial volume by a density of 1.05 g/mL. Derived from the LV and RV volumes, we also computed LV ejection fraction (LVEF) and RV ejection fraction (RVEF). We evaluated the accuracy of these volumetric and functional measures by computing the absolute and relative differences between automated and manual measurements. We define the absolute and relative error as $\varepsilon_{absolute} = |v_{manual} - v_{auto}|)$ and $\varepsilon_{relative}(\%) = 100*|v_{manual} - v_{auto}|/v_{manual}$, where v corresponds to each clinical measure.

Analysis of the Influence of Confounders

To investigate whether a true bias between racial and/or sex groups exists for automated DL-based cine CMR segmentation, we conducted a statistical analysis to investigate if the observed bias could be explained by the most common confounders. In this study, we use as possible confounders age, sex, body measures (i.e., height, weight and BMI), HR, SBP, DBP, CMR-derived parameters (LVEDV, LVESV, RVEDV, RVESV, LVmass), cardiovascular risk factors (i.e., hypertension, hypercholesteremia, diabetes and smoking) and center (i.e., core lab where most of the segmentations were performed vs. additional lab).

Statistical Analysis

Data analysis was performed using SPSS Statistics (version 27, IBM, United States). Continuous variables are reported as mean ± standard deviation (SD) and tested for normal distributions with the Shapiro–Wilk test. Log transformations were applied to the (1-DSC) values to obtain an approximately normal distribution. After transformation, all continuous variables were normally distributed. Categorical data are presented as absolute counts and percentages. Comparison of variables between groups (i.e., races and sexes) was carried out using an independent Student's *t*-test.

Independent association between log-transformed DSC values and race was performed using univariate linear regression followed by multivariate adjustment for confounders. All variables in the regression models were standardized by computing the z-score for individual data points.

Finally, the differences in DSC values among different racial groups were initially assessed by a 1-way ANOVA (Model 4) followed by an analysis of covariance—ANCOVA (Model 5) to statistically control the effect of covariates. In addition, we check the assumption concerning regression residuals (19) as follows: (1) Homoscedasticity tested by a Levene's Test of quality of error variance; (2) Normality of residuals tested by the Kolmogorov–Smirnov and Shapiro–Wilk test; (3) Multicollinearity tested by the Durbin Watson Test. For all statistical analysis, the threshold for statistical significance was $p < 0.01$ and confidence intervals (%) were calculated by non-parametric bootstrapping with 1,000 resamples.

Pairwise *post hoc* testing was carried out using Bonferroni correction and Scheffé correction for multiple comparisons on the *t*-test and ANOVA analysis, respectively.

MATERIALS

Subject Characteristics

The dataset used consisted of ED and ES short-axis cine CMR images of 5,660 healthy subjects (with or without cardiovascular risk factors). Subject characteristics for all participants were obtained from the UK Biobank database and are provided in **Table 1**.

For all subjects, the LV endocardial and epicardial borders and the RV endocardial border were manually traced at ED and ES frames using the cvi42 software (version 5.1.1, Circle Cardiovascular Imaging Inc., Calgary, Alberta, Canada). 4,975 subjects were previously analyzed by two core laboratories based in London and Oxford (20), the remaining 685 subjects were analyzed by two experienced CMR cardiologists at Guy's and St Thomas' Hospital following the same standard operating procedures described in Petersen et al. (20). For all CMR examinations that underwent manual image analysis, any case with insufficient quality (i.e., presence of artifacts or slice location problems, operator error or evidence of pathology, such as significant shunt or valve regurgitation) were rejected (21). All experts performing the segmentations were blinded to subject characteristics such as race and sex. From our database, 4,410 subjects were used to train and validate the DL-based CMR segmentation model, and 1,250 subjects were used as a test set for the validation of the model and the statistical analysis (split 70/10/20 for training/validation/test set). The train and test sets were stratified to contain approximately the same percentage of samples for each racial group and sex. **Supplementary Figure 1** shows the flow chart for selection of cases for this study.

RESULTS

Deep Learning-Based Image Segmentation Pipeline

Table 2 reports the DSC values between manual and automated segmentations evaluated on the test set of 1,250 subjects which the segmentation model had never seen before. The table shows the mean DSC for LVBP, LVMyo and RVBP for both the full test set and stratified by sex and race. Overall, the average (AVG) DSC was 93.03 ± 3.83% (94.40 ± 2.61% for the LVBP, 88.78 ± 3.08% for the LVMyo and 90.77 ± 3.96% for the RVBP). **Table 2** shows that the CMR segmentation model had a racial bias for all comparisons but no sex bias (independent Student's *t*-test between each racial group and rest of the population; $p < 0.001$ for LVBP, LVMyo, RVBP and AVG for all races).[1] **Supplementary Figure 2** shows in the first-row visual examples of frames from a cine CMR sequence and their associated ground truth segmentations, and in the two last rows some sample segmentation results (on different frames) for different racial groups with both high and low DSC.

Next, we evaluate the accuracy of the volumetric and functional measures (LVEDV, LVESV, LVEF, LVmass, RVEDV, RESV, RVEF). **Table 3A** reports the mean values based on the manual segmentations, and **Tables 3B,C** report the mean absolute differences and relative differences between automated and manual measurements, respectively. The Bland-Altman plots for agreement between the pipeline and manual analysis are shown in **Supplementary Figure 3**. For the overall population, results are in line with previous reported values (5, 22) and within the inter-observability range (20).

These results show that for sex there is a statistically significant difference in the absolute error for LVEF, LVmass and RVEF

[1] **Table 2** differs from **Table 1** of our previous work (14), as in the present study we have excluded any case with cardiovascular disease.

TABLE 1 | Population characteristics for the train/validation and test sets.

			Train/validation	Test
Continuous variables	Patients, *n*		4,410	1,250
	Age (years; mean, *SD*)		62 (8)	61 (8)
	Height (cm; mean, *SD*)		169 (9)	169 (9)
	Weight (kg; mean, *SD*)		76 (15)	75 (14)
	BMI (kg/m^2; mean, *SD*)		27 (4)	26 (4)
	BSA (m^2; mean, *SD*)		1.86 (0.21)	1.85 (0.20)
	Systolic blood pressure (mmHg; mean, *SD*)		136 (20)	136 (18)
	Diastolic blood pressure (mmHg; mean, SD)		79 (11)	80 (10)
	Heart rate (bpm; mean, SD)		63 (20)	63 (10)
Categorical variables	Sex (males; *n*, %)		2,299 (52)	655 (52)
	Racial group	White (*n*, %)	3,570 (81)	1,025 (81)
		Mixed (*n*, %)	136 (3)	34 (3)
		Asian (*n*, %)	313 (7)	83 (7)
		Black (*n*, %)	190 (4)	47 (4)
		Chinese (*n*, %)	87 (2)	27 (2)
		Other (*n*, %)	144 (3)	34 (3)

All continuous values are reported as mean(SD), while categorical variables are reported as number (percentage). SD, standard deviation.

TABLE 2 | Dice similarity coefficient (DSC) values for the overall test set and by sex and race.

***N* = 1,250**	**LVBP**	**LVMyo**	**RVBP**	**AVG**
Total	94.39 (2.61)	88.68 (3.06)	90.77 (3.86)	91.28 (3.18)
Male	94.35 (2.55)	89.10 (2.84)	90.61 (3.96)	91.35 (3.12)
Female	94.44 (2.67)	88.59 (3.26)	90.94 (3.94)	91.32 (3.29)
White	95.13 (1.98)***	89.81 (1.48)***	92.24 (2.11)***	92.39 (1.86)***
Mixed	89.79 (1.34)**	80.72 (2.38)**	82.95 (2.53)**	84.49 (2.08)**
Asian	92.15 (2.48)**	86.46 (2.18)*	86.27 (2.63)**	88.29 (2.43)**
Black	91.41 (1.53)***	85.78 (1.73)***	80.88 (2.10)***	86.02 (1.79)***
Chinese	88.98 (2.43)*	79.75 (2.21)*	82.58 (2.32)*	83.77 (2.32)*
Others	90.46 (2.53)*	82.64 (5.44)*	84.77 (3.46)	85.96 (3.81)*

*DSC reported for the LV blood pool (LVBP), LV myocardium (LVMyo) and RV blood pool (RVBP), and average DSC values across LVBP, LVM and RVBP (AVG column). DSC is reported as mean and standard deviation (in parentheses). The first row reports the DSC for the full database, the second and third rows report DSC by sex and the remaining rows report DSC by racial group. Values are reported as mean(SD). Comparison of variables between groups (i.e., male vs. female, white vs. non-white, mixed vs. non-mixed, etc.) was carried out using an independent Student's t-test. Pairwise post hoc testing was carried out using Bonferroni correction for multiple comparisons. Asterisks indicate statistically significant differences between each group and the rest of the test set after correction (28 tests), where *$p < 0.01/28$, **$p < 0.001/28$, ***$p < 0.0001/28$. Exact p-values are reported in **Supplementary Table 3**. SD, standard deviation.*

(independent Student's t-test $p < 0.001$). For different racial groups, they show that the White and Mixed groups have for all clinical parameters a statistically significant difference in absolute and relative error (except Mixed LVmass $p = 0.66$ and $p = 0.15$ for absolute and relative error, respectively). They also show that there is a statistically significant difference in the absolute and relative errors for LVEDV, LVESV, LVEF (except for absolute error for Black and Other LVESV $p = 0.25$ and $p = 0.01$, respectively, and Black LVEF $p = 0.17$; and relative error for Black LVEDV $p = 0.03$, LVESV $p = 0.53$ and LVEF $p = 0.20$). Interestingly, there is no statistically significant difference in absolute or relative error for RV clinical parameters for the Chinese and Other racial groups.

Multivariable Analysis

To analyze if there is any other factor (i.e., risk factors, patient characteristics) that could explain the bias in DSC between races, we performed a multivariate linear regression between the DSC and race adjusted for patient size, cardiac parameters and cardiovascular risk factors and taking the white group as control. **Table 4** shows the unadjusted [model 1—4(a)] and adjusted [model 2—4(b)] standardized regression beta coefficients [with 95% confidence interval (CI)] for the association between DSC and racial groups. **Supplementary Table 1** shows the full list of standardized regression beta-coefficients from the multivariate analysis for each racial group (model 3), representing the z-score change in variables with the associated factors. Our results show that all associations remained significant after multivariate adjustment and that there is no covariate that can explain the DSC bias between racial groups (see **Table 4B**). For the Mixed and Black race groups, sex shows a weak positive association with DSC (see **Supplementary Table 1**), however, race remains the main factor.

Analysis of Variance

We also compared change of marginal means of DSC between different racial groups using a 1-way ANOVA ($F = 219.43$, $p < 0.0001$, $\eta^2 = 0.47$) and an ANCOVA adjusted for patient size, cardiac parameters and cardiovascular risk factors ($F = 196.237$, < 0.0001, $\eta^2 = 0.44$, see **Supplementary Table 2**). Estimated marginal means are given in **Table 5**, before and after adjustment for the mean of covariates. The results show that there is an overall difference between racial groups, and after adjustment for covariates race still remains the main factor.

Effect of Bias on Heart Failure Diagnosis

The previous experiments have demonstrated that racial bias exists in the DL-based CMR segmentation model. This final experiment aims to provide an example of how this racial bias could potentially have an effect on the diagnosis and characterization of heart failure (HF). To this end, we trained another nnU-Net segmentation model using both healthy and cardiomyopathy subjects from the UK Biobank (training and validation: 4,410 healthy subjects/200 cardiomyopathy subjects and test: 1,250 healthy subjects/150 cardiomyopathy subjects). For the cardiomyopathy test cases, we computed the misclassification rate (MCR, %) between the manual LVEF and the automated LVEF based on the standard classification of HF according to LVEF (23, 24), i.e., HF with reduced EF (HFrEF): HF with an LVEF of $\leq$ 40%; HF with mildly reduced EF (HFmrEF): HF with an LVEF of 41–49%; HF with preserved EF (HFpEF): HF with an LVEF of $\geq$ 50%. The results are presented in **Table 6**. Overall, although the number of subjects in the minority racial groups was relatively small, the misclassification rate using the AI-derived segmentations for White subjects was low, with generally much higher rates for minority races.

DISCUSSION

We have demonstrated for the first time the existence of racial bias in DL-based cine CMR segmentation. The results show that after adjustment for possible confounders such as cardiovascular risk factors the bias persists, suggesting that it is related to the balance of the database used to train the DL model. This conclusion is supported by our earlier work (14), where a model trained with a (much smaller) racially balanced database had

TABLE 3 | Manual clinical measurements (top table) and absolute (middle table) and relative (bottom table) differences in volumetric and functional measures between automated and manual segmentations, overall and by sex and race.

(A) Manual

	iLVEDV (mL/mm^2)	**iLVESV (mL/mm^2)**	**LVEF (%)**	**iLVmass (g/mm^2)**	**iRVEDV (mL/mm^2)**	**iRVESV (mL/mm^2)**	**RVEF (%)**
Total	79 (20)	33 (12)	60 (7)	51 (14)	86 (22)	38 (13)	57 (7)
Male	82 (20)*	36 (12)	59 (7)*	50 (12)	95 (21)*	45 (13)	54 (7)*
Female	72 (14)*	29 (8)	61 (7)*	42 (9)	77 (14)*	32 (8)	58 (6)*
White	83 (20)	35 (12)	59 (6)	51 (14)*	87 (22)*	39 (13)*	56 (6)
Mixed	76 (20)*	27 (9)*	64 (8)*	47 (14)	83 (20)*	35 (10)*	58 (8)*
Asian	70 (18)*	25 (10)*	65 (8)*	48 (12)*	76 (19)*	32 (11)	58 (6)
Black	87 (21)	33 (11)	63 (6)	59 (13)	94 (27)*	41 (14)	56 (6)
Chinese	66 (12)*	22 (7)*	66 (7)*	46 (11)*	75 (16)	32 (8)	58 (6)
Others	77 (19)*	28 (9)	64 (6)*	53 (15)	86 (23)	36 (13)	59 (7)

(B) Absolute difference

	iLVEDV (mL/mm^2)	**iLVESV (mL/mm^2)**	**LVEF (%)**	**iLVmass (g/mm^2)**	**iRVEDV (mL/mm^2)**	**iRVESV (mL/mm^2)**	**RVEF (%)**
Total	2.6 (1.7)	2.1 (1.8)	2.5 (2.4)	3.8 (3.9)	3.5 (2.6)	3.0 (2.2)	3.6 (3.0)
Male	2.7 (1.7)	2.1 (1.7)	2.1 (1.9)*	4.1 (4.2)	3.4 (2.6)	3.0 (2.1)	3.1 (2.7)*
Female	2.6 (1.7)	2.1 (1.8)	2.9 (2.8)*	3.5 (3.4)	3.5 (2.6)	4.6 (2.2)	4.1 (3.3)*
White	2.3 (1.5)	1.9 (1.5)*	2.1 (2.1)*	4.0 (3.3)*	3.2 (2.6)*	2.8 (2.2)	3.4 (2.9)*
Mixed	3.9 (2.1)*	3.4 (1.7)*	4.1 (2.7)	1.9 (1.7)*	4.6 (1.8)*	3.9 (1.8)*	4.9 (2.5)*
Asian	3.4 (1.9)*	2.8 (2.3)*	4.0 (2.9)	2.0 (2.3)*	4.4 (2.4)*	3.4 (1.9)	4.4 (3.3)
Black	3.6 (1.8)*	2.9 (2.8)*	3.3 (3.0)*	2.0 (2.2)*	4.4 (1.6)*	3.5 (1.9)	3.9 (2.6)
Chinese	4.4 (2.2)*	3.4 (2.1)*	4.7 (2.8)*	4.1 (3.6)*	4.8 (2.4)	4.0 (2.9)*	6.4 (5.4)*
Others	3.7 (1.9)	3.1 (2.0)*	4.3 (3.2)	2.3 (2.5)	4.6 (3.4)	3.6 (1.8)*	4.3 (2.8)

(C) Relative difference

	iLVEDV (mL/mm^2)	**iLVESV (mL/mm^2)**	**LVEF (%)**	**iLVmass (g/mm^2)**	**iRVEDV (mL/mm^2)**	**iRVESV (mL/mm^2)**	**RVEF (%)**
Total	3.4 (2.5)	7.1 (7.4)	4.1 (3.9)	8.7 (8.3)	4.3 (3.4)	8.8 (7.5)	6.4 (5.2)
Male	3.0 (2.3)*	6.2 (6.3)*	3.6 (3.1)*	7.8 (6.5)*	3.7 (3.0)*	7.3 (5.9)*	5.8 (5.0)*
Female	3.7 (2.7)*	7.9 (8.2)*	4.6 (4.4)*	9.6 (9.6)*	4.9 (3.7)*	10.2 (8.4)*	7.0 (5.4)*
White	3.0 (2.1)*	6.0 (6.1)*	3.7 (3.6)	8.4 (8.7)	4.0 (3.4)*	8.2 (7.3)	6.0 (5.1)*
Mixed	5.7 (3.1)*	14.1 (8.2)*	6.5 (4.2)*	10.3 (6.1)*	6.2 (2.4)*	13.3 (6.8)*	9.2 (5.1)*
Asian	5.1 (3.2)*	11.8 (11.6)*	5.8 (4.2)*	10.5 (5.4)*	6.1 (3.4)*	11.5 (6.8)	7.2 (4.9)
Black	4.1 (2.3)	7.7 (6.8)	5.1 (4.8)*	7.3 (4.1)	5.1 (2.2)	9.3 (5.9)	7.3 (4.7)
Chinese	7.0 (4.3)*	16.5 (10.6)*	6.9 (3.7)*	13.6 (7.1)*	6.2 (3.2)	13.8 (11.4)*	10.4 (9.4)*
Others	5.0 (2.9)*	12.6 (10.2)	7.7 (5.5)*	8.9 (4.2)	5.2 (3.9)	11.9 (7.0)*	8.1 (4.9)

Clinical measurements for the LV and RV end diastolic volume (EDV), end systolic volume (ESV), ejection fraction (EF), and left ventricular mass (LVmass). All cardiac volumes were indexed to body surface area using the Dubois and Dubois formula (32). We define the absolute and relative errors as$\varepsilon_{absolute} = |v_{manual} - v_{auto}|$) *and*$\varepsilon_{relative}(\%) = 100 * |v_{manual} - v_{auto}|/v_{manual}$, *where*$v$ *corresponds to each clinical measure. Clinical measures are reported as mean and standard deviation (in parentheses). The first row reports the clinical measurements for the full database, the second and third rows report the clinical measurements by sex and the remaining rows report the clinical measurements by racial group. Values are reported as mean(SD). Comparison of variables between groups (i.e., male vs. female, white vs. non-white, mixed vs. non-mixed, etc.) was carried out using an independent Student's t-test. Pairwise post hoc testing was carried out using Bonferroni correction for multiple comparisons. Asterisks indicate statistically significant differences between each group and the rest of the test set after correction (49 tests), i.e., $p < 0.01/49$. Exact p-values are reported on* ***Supplementary Table 4****. SD, standard deviation.*

much reduced bias (although poorer performance overall due to the smaller training database).

Assessment of the Bias in the Deep Learning-Based Cardiac Magnetic Resonance Segmentation Model

For the overall population, the DSC values are in line with previous reported values (5, 22) and with the inter-observer variability range (20). DSC as well as absolute differences and relative differences show a higher bias on the RV, however, this is expected as previous studies have highlighted the difficulty in manual contouring of the RV and the higher variability between observers (20).

The bias we found in segmentation model performance was near-exclusively based on race. Statistically significant differences in some derived volumetric/functional measures (see **Table 3**) were found by sex but these differences were small

TABLE 4 | Associations between average DSC and racial group.

(A) Univariate linear regression

	Standardized beta-coefficients (95% CI)		
	N	**Model 1**	***p*-value**
Mixed	1,250	0.34 (0.30, 0.38)***	6.30E-16
Asian	1,250	0.33 (0.29, 0.37)***	1.57E-12
Black	1,250	0.36 (0.32, 0.40)***	1.30E-19
Chinese	1,250	0.32 (0.28, 0.36)***	1.08E-8
Other	1,250	0.30 (0.26, 0.34)***	4.43E-14

(B) Multivariate linear regression

	Standardized beta-coefficients (95% CI)		
	N	**Model 2**	***p*-values**
Age	1,250	0.03 (–0.02, 0.08)	0.210
Sex	1,250	0.02 (–0.03, 0.08)	0.364
Weight	1,250	0.10 (–0.36, 0.51)	0.699
Height	1,250	0.00 (–0.28, 0.29)	0.972
BMI	1,250	-0.02 (–0.36, 0.36)	0.944
HR	1,250	0.03 (–0.01, 0.07)	0.114
SBP	1,250	-0.01 (–0.07, 0.04)	0.579
DBP	1,250	-0.04 (–0.08, 0.01)	0.114
LVEDV	1,250	-0.02 (–0.21, 0.17)	0.855
LVESV	1,250	-0.07 (–0.20, 0.06)	0.284
RVEDV	1,250	0.12 (–0.09, 0.31)	0.235
RVESV	1,250	-0.11 (–0.24, 0.04)	0.127
Lvmass	1,250	-0.04 (–0.11, 0.02)	0.174
Diabetes	1,250	0.10 (–0.07, 0.27)	0.273
Hypertension	1,250	0.05 (0.00, 0.10)	0.034
Hyper cholesterolemia	1,250	0.00 (–0.04, 0.05)	0.860
Smoking	1,250	0.00 (–0.05, 0.03)	0.812
Center	1,250	0.15 (0.09, 0.21)	9.99E-02
Mixed	1,250	0.38 (0.36, 0.41)**	9.99E-04
Asian	1,250	0.37 (0.34, 0.41)**	9.99E-04
Black	1,250	0.40 (0.38, 0.43)**	9.99E-04
Chinese	1,250	0.36 (0.34, 0.39)**	9.99E-04
Other	1,250	0.34 (0.30, 0.38)**	9.99E-04

*Standardized regression beta-coefficients and CI are shown, representing the z-score change in variables with increasing DSC. The White racial group was selected as control. LV, left ventricle, EDV, end-diastolic volume, ESV, end-systolic volume, SBP, systolic blood pressure, DBP, diastolic blood pressure, CI, confidence interval. Model 1 is unadjusted; Model 2 is adjusted for sex, height, weight, blood pressure at scan-time, heart rate at scan-time, LVEDV, LVESV, RVEDV, RVESV, LVmass, diabetes, hypertension, hypercholesterolemia, smoking and center. *p < 0.01, **p < 0.001, ***p < 0.00001.*

compared to the differences observed in both DSC (**Table 2**) and volumetric/functional measures (**Table 3**) by race. Therefore, none of the confounders used in this study could explain the differences by race. Results from the ANCOVA analysis show that one factor that contributed more to the model was the center where the segmentations were performed. This could be explained by differences in CMR reporting between the core lab and the additional lab. Similarly to the complete UK Biobank database, the subcohort that we used is approximately sex-balanced but not race-balanced, and the highest errors were found for relatively underrepresented racial groups. This phenomenon has been observed before in applications in computer vision (25) and medical imaging (26, 27), but never before reported in CMR image analysis.

We believe that this bias is due to the imbalanced nature of the training data. Combined with previous studies that have shown race-based associations with differences in cardiac physiology using diverse databases (10, 11), the imbalance causes the performance of the DL model to be biased toward the physiology of the majority group (i.e., white subjects), to the detriment of performance on minority racial groups.

Our last experiment showed that using the AI-based predicted EF values will result in higher misclassification rates for the minority races compared to the White subjects, which is in line with the other experiments showing a higher bias for the minority groups.

Consistent Reporting of Sex and Racial Subgroups in Artificial Intelligence Models

It is envisioned that AI will dramatically change the way doctors practice medicine. In the short term, it will assist physicians with easy tasks, such as automating measurements, making predictions based on big data, and putting clinical findings into an evidence-based context. In the long term, it has the potential to significantly optimize patient care, reduce costs, and improve outcomes. With AI models now starting to be deployed in the real world it is essential that the benefits of AI are shared equitably according to race, sex and other demographic characteristics. It has long been known that current medical guidelines have the potential for sex/racial bias due to the imbalanced nature of the cohorts upon which they were based (28, 29). One might think that AI can solve such problems, as they are "neutral" or "blind" to characteristics such as sex and race. However, as we have shown in this paper, when AI models are used naively, they can inherit the bias present in clinical databases. It is important to highlight

TABLE 5 | The comparison of adjusted mean between racial groups based on one-way ANOVA and ANCOVA.

	Mean (95% CI)		
	N	**Model 4**	**Model 5**
White	1,025	0.93 (0.93, 0.93)	0.93 (0.93, 0.93)
Mixed	34	0.84 (0.86, 0.82)	0.83 (0.85, 0.80)
Asian	83	0.89 (0.90, 0.88)	0.88 (0.89, 0.88)
Black	47	0.86 (0.87, 0.85)	0.85 (0.86, 0.83)
Chinese	27	0.84 (0.86, 0.81)	0.82 (0.84, 0.78)
Other	34	0.86 (0.88, 0.85)	0.85 (0.87, 0.83)

Model 4 is unadjusted; Model 5 is adjusted for sex, height, weight, blood pressure at scan-time, heart rate at scan-time, LVEDV, LVESV, RVEDV, RVESV, LVmass, diabetes, hypertension, hypercholesterolemia, smoking, and center. CI, confidence interval. For model 4 and model 5, pairwise post hoc testing was carried out using Scheffé's method.

TABLE 6 | Misclassification rate for HF diagnosis.

		HFrEF		**HFmrEF**		**HFpEF**	
		LVEF < 40%		**LEF 40–49%**		**LVEF ≥ 50%**	
	n	**n GT**	**MCR (%)**	**n GT**	**MCR**	**n GT**	**MCR (%)**
White	107	5	3.74	14	5.61	88	7.48
Mixed	11	3	45.45	0	–	8	36.36
Black	8	0	–	4	12.05	4	25.00
Asian	14	4	21.43	2	7.14	8	14.29
Chinese	4	0	–	2	25.00	2	50.00
Other	6	1	33.33	5	16.67	0	–
Minority groups	43	8	23.26	13	9.30	22	23.26

The table summarizes numbers of subjects in each racial group and HF diagnosis (i.e., HFrEF, HFmrEF and HFpEF), as well as the misclassification rate (MCR,%) for each racial group and diagnosis. The row Minority groups combines data from the Mixed, Black, Asian, Chinese and Other groups. The left column (n overall) shows the number of subjects for each racial group used to compute the MCRs. For each HF diagnosis, the first column shows the number of ground truth positive subjects in that group, and the second column shows the MCR. When computing the MCRs, the ground truth negative subjects were all subjects from the other HF diagnoses for that racial group. HFrEF, HF with reduced EF; HFmrEF, HF with mildly reduced EF; HFpEF, HF with preserved EF. Blank cells show regions with missing data.

the potential shortcomings of AI at this stage before AI models become more widely deployed in clinical practice.

For these reasons, we believe that it is necessary that new standards are established to ensure equality between demographic groups in AI model performance, and that there is consistent and rigorous reporting of performance for new AI models that are intended to be integrated into clinical practice. Similar to Noseworthy et al. (30), we would recommend that any new AI-based publication include a report of performance across a range of demographic subgroups, particularly race/sex.

Strategies to Reduce Racial Bias

The obvious way to mitigate bias due to imbalanced datasets (whether in current clinical guidelines or AI models) is to use more balanced datasets. However, this is a multifactorial problem and is associated with many challenges, such as historical discrimination, research design and accessibility (22). We note that AI has the potential to address/mitigate bias without requiring such balanced datasets. A range of bias mitigation strategies have been proposed that either pre-process the dataset to make it less imbalanced, alter the training procedure or post-process the model outputs to reduce bias (31). We have recently proposed three algorithms to mitigate racial bias in CMR image segmentation: (1) train a CMR segmentation algorithm that ensures racial balance during training; (2) add an AI race classifier that helps the segmentation model to capture racial variations; and (3) train a different CMR segmentation model for each racial group. For more detail of these models, we refer to the reader to our previous work (14). All three proposed algorithms result in a fairer segmentation model that aims to ensure that no racial group will be disadvantaged when segmentations of their CMR data are used to inform clinical management. Note that, compared to our previous work (14), in this paper we have excluded all subjects with cardiovascular disease to ensure that racial bias was not influenced by this factor.

Limitations

This study utilizes the imaging cohort from the UK Biobank. UK Biobank is a long-term prospective epidemiology study of over 500,000 persons aged 40–69 years across England, Scotland, and Wales. Therefore, the data are geographically limited to the UK population, which might not reflect geographic, socioeconomic or healthcare differences among other populations. This work uses the UK Biobank participants' self-reported ethnicity, which corresponds to them self-identifying as belonging to ethnic groups based on shared culture and heritage. A possible limitation is that ethnic groups are socially constructed and thus may not serve as reliable proxies for analysis. Future work should aim to perform a similar study using genetic ancestry data, which will make the analysis more generalizable. In addition, Mixed Race was considered to be a single category, whereas in reality this encompasses many different subcategories.

Manual analysis of CMR scans was performed by three independent centers using the same operating procedures for analysis. For the three centers, inter- and intra-observer variability between analysts was assessed by analysis of fifty, randomly selected CMR examinations (20). However, one limitation of this study is that inter- and intra-observer variability was not assessed individually by race and sex. Also, this study is limited by the lack of diversity and relatively small sample sizes for certain racial groups and by the exclusion criteria for comorbid and pre-morbid conditions. The study only includes the following cardiovascular risk factors as confounders: hypertension, hypercholesteremia, diabetes and smoking. However, there are other clinically relevant risk factors such as sedentarism, alcohol consumption or stress that could potentially explain the bias found in our study. For instance, a previous study showed an association between RV size and living in a high traffic area (7). Another limitation is that current analysis does not adjust for any measures of ventricular function, which could explain the structural differences. Future work will aim to extract echocardiographic measures of relaxation to assess whether the current bias could be explained by changes in subclinical diastolic dysfunction.

CONCLUSION

We have demonstrated that a DL-based cine CMR segmentation model derived from an imbalanced database has poor generalizability across racial groups and has the potential to lead to inequalities in early diagnosis, treatments and outcomes. Therefore, for best practice, we recommend reporting of performance among diverse groups such as those based on sex and race for all new AI tools to ensure responsible use of AI technology in cardiology.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The UK Biobank datasets are publicly available for approved research projects. Requests to access these datasets should be directed to https://www.ukbiobank.ac.uk/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the NHS National Research Ethics Service on 17th June 2011 (Ref 11/NW/0382). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

EP-A designed, developed the method, and analyzed the data. AK, RR, BR, JM, PC, and EP-A conceived the study. BR, RR, SKP, SN, and SEP provided the manual segmentation used for the implementation of the method. PC, RR, and AK were part of the supervision of EP-A. AK and EP-A wrote the manuscript with input from all authors.

FUNDING

EP-A and AK were supported by the EPSRC (EP/R005516/1) and by core funding from the Wellcome/EPSRC Centre for Medical Engineering (WT203148/Z/16/Z). This research was funded in whole, or in part, by the Wellcome Trust WT203148/Z/16/Z. For the purpose of open access, the author has applied a CC BY public copyright license to any author accepted manuscript version arising from this submission. SEP, AK, and RR acknowledge funding from the EPSRC through the Smart Heart Programme grant (EP/P001009/1). EP-A, BR, JM, AK, and RR acknowledged support from the Wellcome/EPSRC Centre for Medical Engineering at King's College London (WT 203148/Z/16/Z), the NIHR Cardiovascular MedTech Co-operative award to the Guy's and St Thomas' NHS Foundation Trust and the Department of Health National Institute for Health Research (NIHR) comprehensive Biomedical Research Centre award to Guy's & St Thomas' NHS Foundation Trust in partnership with King's College London. SEP, SN, and SKP acknowledged the British Heart Foundation for funding the manual analysis to create a cardiovascular magnetic resonance imaging reference standard for the UK Biobank imaging resource in 5,000 CMR scans (www.bhf.org.uk; PG/14/89/31194). SEP acknowledged support from the National Institute for Health Research (NIHR) Biomedical Research Centre at Barts. SEP has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 825903 (euCanSHare project). SEP also acknowledged support from the CAP-AI Programme, London's First AI Enabling Programme focused on stimulating growth in the capital's AI Sector. CAP-AI was led by Capital Enterprise in partnership with Barts Health NHS Trust and Digital Catapult and was funded by the European Regional Development Fund and Barts Charity. SEP acknowledged support from the Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. SN and SKP were supported by the Oxford NIHR Biomedical Research Centre and the Oxford British Heart Foundation Centre of Research Excellence.

ACKNOWLEDGMENTS

This research has been conducted using the UK Biobank Resource (application numbers 17,806 and 2,964) on a GPU generously donated by NVIDIA Corporation. The UK Biobank data are available for approved projects from https://www.ukbiobank.ac.uk/.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcvm.2022.859310/full#supplementary-material

REFERENCES

1. Constantinides P, Fitzmaurice DA. Artificial intelligence in cardiology: applications, benefits and challenges. *Br J Cardiol.* (2018) 7:25–86.
2. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* (2017) 542:115–8. doi: 10.1038/nature21056
3. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat Med.* (2018) 24:1559–67. doi: 10.1038/s41591-018-0177-5
4. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial intelligence in cardiology. *J Am Coll Cardiol.* (2018) 71:2668–79.
5. Bai W, Sinclair M, Tarroni G, Oktay O, Rajchl M, Vaillant G, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *J Cardiovasc Magn Reson.* (2018) 20:65. doi: 10.1186/s12968-018-0471-x
6. Bernard O, Lalande A, Zotti C, Cervenansky F, Yang X, Heng P-A, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Trans Med Imaging.* (2018) 37:2514–25. doi: 10.1109/TMI.2018.2837502

7. Yoneyama K, Venkatesh BA, Bluemke DA, McClelland RL, Lima JAC. Cardiovascular magnetic resonance in an adult human population: serial observations from the multi-ethnic study of atherosclerosis. *J Cardiovasc Magn Reson.* (2017) 19:52. doi: 10.1186/s12968-017-0367-1
8. Holmes MD. Racial inequalities in the use of procedures for ischemic heart disease. *JAMA.* (1989) 261:3242–3. doi: 10.1001/jama.1989.03420220056014
9. Regitz-Zagrosek V, Oertelt-Prigione S, Prescott E, Franconi F, Gerdts E, Foryst-Ludwig A, et al. Gender in cardiovascular diseases: impact on clinical manifestations, management, and outcomes. *Eur Heart J.* (2016) 37:24–34. doi: 10.1093/eurheartj/ehv598
10. Oertelt-Prigione S, Regitz-Zagrosek V. *Sex and Gender Aspects in Clinical Medicine*. London: Springer (2012).
11. Kawut SM, Lima JAC, Barr RG, Chahal H, Jain A, Tandri H, et al. Sex and race differences in right ventricular structure and function. *Circulation.* (2011) 123:2542–51. doi: 10.1161/CIRCULATIONAHA.110.985515
12. Captur G, Zemrak F, Muthurangu V, Petersen SE, Li C, Bassett P, et al. Fractal analysis of myocardial trabeculations in 2547 study participants: multi-ethnic study of atherosclerosis. *Radiology.* (2015) 277:707–15. doi: 10.1148/radiol.2015142948
13. Kishi S, Reis JP, Venkatesh BA, Gidding SS, Armstrong AC, Jacobs DR, et al. Race–ethnic and sex differences in left ventricular structure and function: the coronary artery risk development in young adults (CARDIA) study. *J Am Heart Assoc.* (2015) 4:e001264. doi: 10.1161/JAHA.114.001264
14. Puyol-Antón E, Ruijsink B, Piechnik SK, Neubauer S, Petersen SE, Razavi R, et al. Fairness in cardiac MR image analysis: an investigation of bias due to data imbalance in deep learning based segmentation. In: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021*. Cham: Springer (2021). p. 413–23.
15. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, 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.* (2015) 12:e1001779. doi: 10.1371/journal.pmed.1001779
16. Office for National Statistics, National Records of Scotland, Northern Ireland Statistics and Research Agency. *2011 Census Aggregate Data* (Edition: February 2017). UK Data Service (2017). doi: 10.5257/census/aggregate-2011-2
17. Petersen SE, Matthews PM, Francis JM, Robson MD, Zemrak F, Boubertakh R, et al. UK Biobank's cardiovascular magnetic resonance protocol. *J Cardiovasc Magn Reson.* (2015) 18:8. doi: 10.1186/s12968-016-0227-4
18. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods.* (2021) 18:203–11. doi: 10.1038/s41592-020-01008-z
19. Barker LE, Shaw KM. Best (but oft-forgotten) practices: checking assumptions concerning regression residuals. *Am J Clin Nutr.* (2015) 102:533–9. doi: 10.3945/ajcn.115.113498
20. Petersen SE, Aung N, Sanghvi MM, Zemrak F, Fung K, Paiva JM, et al. Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK biobank population cohort. *J Cardiovasc Magn Reson.* (2017) 19:1–19. doi: 10.1186/s12968-017-0327-9
21. Carapella V, Jiménez-Ruiz E, Lukaschuk E, Aung N, Fung K, Paiva J, et al. Towards the semantic enrichment of free-text annotation of image quality assessment for UK biobank cardiac cine MRI scans. In: Carneiro G, Mateus D, Peter L, Bradley A, Tavares JMR, Belagiannis V, et al. editors. *Deep Learning and Data Labeling for Medical Applications. DLMIA 2016, LABELS 2016. Lecture Notes in Computer Science*. (Vol. 10008), Cham: Springer (2016). doi: 10.1007/978-3-319-46976-8_25
22. Ruijsink B, Puyol-Antón E, Oksuz I, Sinclair M, Bai W, Schnabel JA, et al. Fully automated, quality-controlled cardiac analysis from CMR. *JACC Cardiovasc Imaging.* (2020) 13:684–95. doi: 10.1016/j.jcmg.2019.05.030
23. Bozkurt B, Coats AJ, Tsutsui H, Abdelhamid M, Adamopoulos S, Albert N, et al. Universal definition and classification of heart failure. *J Card Fail.* (2021) 27:387–413.
24. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS, et al. 2016 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure. *Eur Heart J.* (2016) 37:2129–200.
25. Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. In: Friedler SA, Wilson C editors. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. New York, NY (2018). p. 77—-91.
26. Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. *CheXclusion: Fairness Gaps in Deep Chest X-Ray Classifiers*. Singapore: World Scientific (2020).
27. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc Natl Acad Sci USA.* (2020) 117:12592–4. doi: 10.1073/pnas.1919012117
28. Institute of Medicine (US) Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*. Smedley BD, Stith AY, Nelson AR editors. Washington, DC: National Academies Press (2003).
29. Smith Taylor J. Women's health research: progress, pitfalls, and promise. *Health Care Women Int.* (2011) 32:555–6. doi: 10.17226/12908
30. Noseworthy PA, Attia ZI, Brewer LC, Hayes SN, Yao X, Kapa S, et al. Assessing and mitigating bias in medical artificial intelligence. *Circ Arrhythm Electrophysiol.* (2020) 13:e007988. doi: 10.1161/CIRCEP.119.007988
31. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. In: *Proceedings of the ACM Computing Surveys (CSUR)*. (Vol. 54), New York, NY: Association for Computing Machinery (2019), 1–35. doi: 10.1145/3457607
32. Du Bois D, Du Bois EF. A formula to estimate the approximate surface area if height and weight be known. *Arch Intern Med.* (1916) 17:863–71. doi: 10.1001/archinte.1916.00080130010002

Author Disclaimer: The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, EPSRC, or the Department of Health

Conflict of Interest: SEP provided consultancy to and is shareholder of Circle Cardiovascular Imaging, Inc., Calgary, Alberta, Canada.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Cardiovascular Medicine

TYPE Systematic Review
PUBLISHED 15 July 2022
DOI 10.3389/fcvm.2022.956811

Check for updates

OPEN ACCESS

EDITED BY
Steffen Erhard Petersen,
Queen Mary University of London,
United Kingdom

REVIEWED BY
Lennart Tautz,
Fraunhofer Institute for Digital
Medicine MEVIS, Germany
Esther Puyol-Antón,
King's College London,
United Kingdom

*CORRESPONDENCE
Samer Alabed
s.alabed@nhs.net

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Cardiovascular Imaging, a section of the journal Frontiers in Cardiovascular Medicine

RECEIVED 30 May 2022
ACCEPTED 30 June 2022
PUBLISHED 15 July 2022

CITATION
Alabed S, Maiter A, Salehi M, Mahmood A, Daniel S, Jenkins S, Goodlad M, Sharkey M, Mamalakis M, Rakocevic V, Dwivedi K, Assadi H, Wild JM, Lu H, O'Regan DP, van der Geest RJ, Garg P and Swift AJ (2022) Quality of reporting in AI cardiac MRI segmentation studies – A systematic review and recommendations for future studies. *Front. Cardiovasc. Med.* 9:956811. doi: 10.3389/fcvm.2022.956811

Quality of reporting in AI cardiac MRI segmentation studies – A systematic review and recommendations for future studies

Samer Alabed[1,2,3*†], Ahmed Maiter[1,2†], Mahan Salehi[1], Aqeeb Mahmood[4], Sonali Daniel[4], Sam Jenkins[4], Marcus Goodlad[1], Michael Sharkey[1], Michail Mamalakis[1,3], Vera Rakocevic[4], Krit Dwivedi[1,2], Hosamadin Assadi[5], Jim M. Wild[1,3], Haiping Lu[3,6], Declan P. O'Regan[7], Rob J. van der Geest[8], Pankaj Garg[5] and Andrew J. Swift[1,3]

[1]Department of Infection, Immunity and Cardiovascular Disease, The University of Sheffield, Sheffield, United Kingdom, [2]Department of Clinical Radiology, Sheffield Teaching Hospitals, Sheffield, United Kingdom, [3]INSIGNEO, Institute for *in silico* Medicine, The University of Sheffield, Sheffield, United Kingdom, [4]Medical School, The University of Sheffield, Sheffield, United Kingdom, [5]University of East Anglia, Norwich Medical School, Norwich, United Kingdom, [6]Department of Computer Science, The University of Sheffield, Sheffield, United Kingdom, [7]MRC London Institute of Medical Sciences, Imperial College London, London, United Kingdom, [8]Leiden University Medical Center, Leiden, Netherlands

Background: There has been a rapid increase in the number of Artificial Intelligence (AI) studies of cardiac MRI (CMR) segmentation aiming to automate image analysis. However, advancement and clinical translation in this field depend on researchers presenting their work in a transparent and reproducible manner. This systematic review aimed to evaluate the quality of reporting in AI studies involving CMR segmentation.

Methods: MEDLINE and EMBASE were searched for AI CMR segmentation studies in April 2022. Any fully automated AI method for segmentation of cardiac chambers, myocardium or scar on CMR was considered for inclusion. For each study, compliance with the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) was assessed. The CLAIM criteria were grouped into study, dataset, model and performance description domains.

Results: 209 studies published between 2012 and 2022 were included in the analysis. Studies were mainly published in technical journals (58%), with the majority (57%) published since 2019. Studies were from 37 different countries, with most from China (26%), the United States (18%) and the United Kingdom (11%). Short axis CMR images were most frequently used (70%), with the left ventricle the most commonly segmented cardiac structure (49%). Median compliance of studies with CLAIM was 67% (IQR 59–73%). Median compliance was highest for the model description domain (100%, IQR 80–100%) and lower for the study (71%, IQR 63–86%), dataset (63%, IQR 50–67%) and performance (60%, IQR 50–70%) description domains.

Conclusion: This systematic review highlights important gaps in the literature of CMR studies using AI. We identified key items missing—most strikingly poor description of patients included in the training and validation of AI models and inadequate model failure analysis—that limit the transparency, reproducibility and hence validity of published AI studies. This review may support closer adherence to established frameworks for reporting standards and presents recommendations for improving the quality of reporting in this field.

Systematic Review Registration: [www.crd.york.ac.uk/prospero/], identifier [CRD42022279214].

KEYWORDS

artificial intelligence, machine learning, cardiac MRI, segmentation, systematic review, quality, reporting

Introduction

Cardiac MRI (CMR) is the gold standard for non-invasive assessment of cardiac structures. Quantitative measurement of cardiac volumes can be achieved with CMR and relies on accurate segmentation of structures on CMR images. Manual segmentation is routinely performed by cardiac imaging experts but suffers from a number of drawbacks. In addition to being laborious and time-intensive, manual segmentation is operator-dependent, potentially impacting interobserver agreement. As the demand for cardiac imaging continues to grow and outpaces the supply of trained readers, there is an increasing need for automation (1, 2).

Artificial intelligence (AI) is changing medical imaging through the automation of complex and repetitive tasks, including the segmentation of anatomical structures (3). Machine learning is a subfield of AI that is commonly used for image analysis and processing in medical applications. Machine learning algorithms learn by experience, typically in a supervised manner: the algorithm is trained on labeled data, such as a set of manually segmented CMR images, where the manual segmentation provides the reference standard or ground truth. The algorithm identifies discriminative features and patterns in this image data, which are incorporated to generate a model that can perform the task—such as segmentation of the cardiac chambers—on new unlabeled data without the need for explicit programming. Machine learning itself encompasses a diverse range of techniques, including deep learning, which can be applied to the segmentation of structures in imaging (4).

A growing number of studies have reported the use of AI methods for segmentation in CMR. The manner in which these studies are reported is important. Transparent reporting of methods and results facilitates reproducibility and allows proper evaluation of validity. Equally, a consistent standard of reporting aids comparison between studies and may improve accessibility of the literature, which may be of particular benefit in a rapidly expanding field such as AI. The need for consistency in reporting medical research is well recognized and reflected in various guidelines and checklists for different study types. The Checklist for Artificial Intelligence in Medical Imaging (CLAIM), (5) has adopted the validated and widely used Standards for Reporting of Diagnostic Accuracy Studies (STARD) guidelines and incorporated domains specific to AI studies, including detailed descriptions of data sources, model design and performance evaluation. This systematic review aimed to evaluate the quality of reporting of studies involving AI CMR segmentation by assessing compliance with CLAIM.

Materials and methods

The study protocol was registered with The International Prospective Register of Systematic Reviews (PROSPERO; registry number CRD42022279214). The study was undertaken and is presented in accordance with the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines (6). No ethical approval was required.

Inclusion and exclusion criteria

Studies reporting the use of AI for segmentation of structures in CMR were considered for inclusion. Studies were deemed eligible if they reported: (1) any type of fully

Abbreviations: AI, artificial intelligence; CLAIM, checklist for artificial intelligence in medical imaging; CMR, cardiac MRI; CT, computer tomography; DSC, dice similarity coefficient; PRISMA, preferred reporting items for systematic reviews and meta-analyses; PROSPERO, international prospective register of systematic reviews; MICCAI, medical image computing and computer-assisted intervention.

automated AI method (including machine learning, deep learning and neural networks), (2) segmentation of cardiac chambers, myocardium or scar tissue and (3) use of adult human CMR images, regardless of acquisition methods (such as use of intravenous contrast), parameters, post-processing methods and software. Exclusion criteria were as follows: absence of a newly developed segmentation model (e.g., studies assessing existing methods), use of semi-automated AI methods (where the segmentation process required manual input), multiorgan segmentation, combined segmentation of multiple imaging modalities (e.g., CMR and CT), segmentation of cardiac vessels (e.g., aorta, pulmonary artery, coronary arteries) or pericardial tissue, use of non-human or *ex vivo* images, and conference publications. **Figure 1** shows an example of automatic biventricular (7) (**Figure 1A**) and four-chamber (8) (**Figure 1B**) segmentation on CMR.

Search method

The MEDLINE and EMBASE databases were searched for relevant studies on April 20 2022. The search strategy is outlined in the **Supplementary Material**. Non-English language publications were excluded.

Study selection

Figure 2 indicates the flow of study identification and inclusion. Duplicate studies were removed following the initial database search. The titles and abstracts of the remaining studies were screened for relevance. The full texts of all potentially relevant studies were retrieved and assessed for eligibility against the inclusion and exclusion criteria. Conference abstracts and studies lacking sufficient information for evaluation were excluded at this point. Screening was performed independently by (SA) and by (SD, AM2, MS2) and full texts were assessed for eligibility by SA, AM1 and MS, with SA acting as an arbitrator.

Data extraction

Three authors extracted data from the included studies (SA, AM1, MS1) according to a standardized checklist. Half of the included studies were also evaluated independently by an additional five authors (SD, AM2, SJ, MG, HA) for the purpose of quality control. All discrepancies were resolved with discussion, with SA acting as an arbitrator, and the final extracted data confirmed. Descriptive information about each study was recorded, including publication details (type, source, country, year), data used (type of data set, type of CMR image, segmented structures) and AI model (validation and performance evaluation methods). The studies were assessed for compliance against the 42 criteria of CLAIM, which were grouped into four domains: study description (9 criteria), dataset description (17 criteria), model description (6 criteria) and model performance (10 criteria). **Supplementary Table 2** indicates all CLAIM criteria and their assignment to the domains. For each criterion, compliance was marked as yes, no or not applicable (N/A). Studies deemed N/A were excluded when evaluating the proportion of studies compliant with CLAIM criteria. For studies using solely public datasets, the following criteria were marked as N/A, as they can be considered implicit in the use of publicly available data sources: retrospective or prospective study, source of ground truth annotations, annotation tools, de-identification methods and inter- and intra-rater variability. Additionally, the following criteria were marked as N/A for all studies: rationale for choosing the reference standard (as manual expert contouring is the standard in the field) and registration number and name of registry. Descriptive data and the number of studies compliant with CLAIM criteria are presented as proportional values (%).

Results

Search results

The database search yielded 2,855 hits from which the title and abstract screening identified 364 relevant studies. The subsequent full-text assessment deemed 209 eligible for inclusion in the analysis (**Figure 2**).

Included studies

Descriptive information for all of the 209 included studies are provided in **Supplementary Table 1**. Selected metrics are highlighted in **Figure 3**. The majority of studies (57%) were published since 2019 (**Figure 3A**). Most studies were published in technical journals (58%), with a minority published in medical (31%) or hybrid (11%) journals. The studies were undertaken in 37 different countries (**Figure 3B**), with just over half coming from China (26%), the United States (18%) and the United Kingdom (11%).

Publicly available datasets were used in 49% of studies, and single or multicenter non-public datasets used in 61%, 17% of studies used multiple or combined datasets (including multiple public datasets and a combination of public and non-public datasets). A minority of studies (6%) did not report their data source (**Figure 3C**). Of the public datasets used, the majority (86%) had been made available through Medical Image Computing and Computer-Assisted Intervention (MICCAI) challenges or the Cardiac Atlas Project (9) (**Figure 3D**). Most studies reported the number of cases used (95%), with a range of 3–12,984 and a median of 78. Short axis CMR images were most

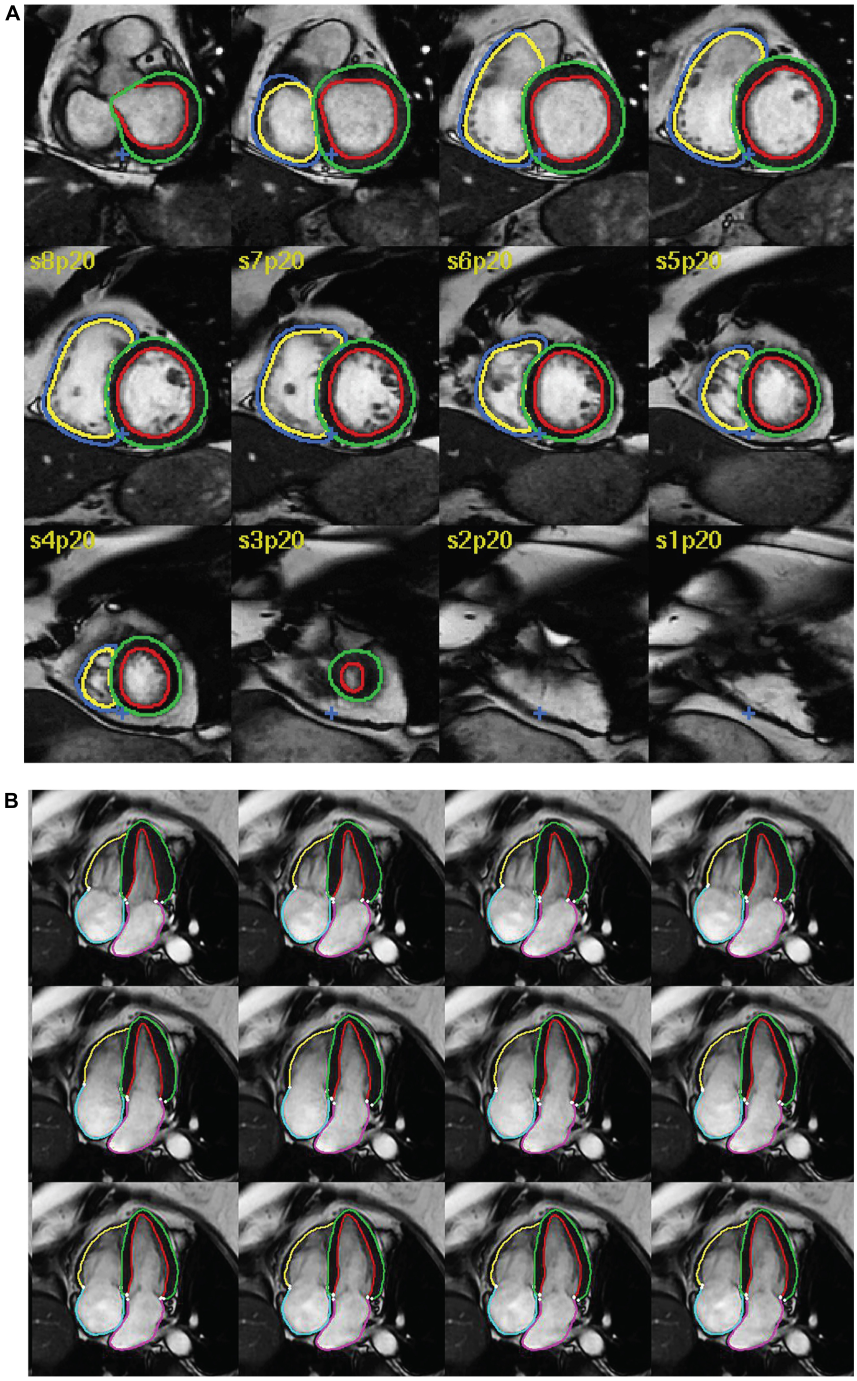


FIGURE 1
Examples of AI cardiac MRI segmentation. Examples of automatic **(A)** biventricular and **(B)** four-chamber segmentation. The colored contours in green and red show the left ventricular epi- and endocardium, respectively. The contours in dark blue and yellow show the right ventricular epi- and endo- cardium, respectively. The pink and turquoise contours outline the left and right atria, respectively.

frequently used (70%), while 14% of studies did not report the specific type of CMR image used for segmentation (Figure 3E). The left ventricle was the most commonly segmented structure, either alone or in combination (49%, Figure 3F). Segmentation of multiple structures was reported in 23% of studies.

Model validation was mostly reported using internal holdout methods (78%), such as cross-validation. A minority reported testing on external and mainly public datasets (22%, Figure 3G). The Dice similarity coefficient (DSC) was used to assess model performance in 79% of studies,

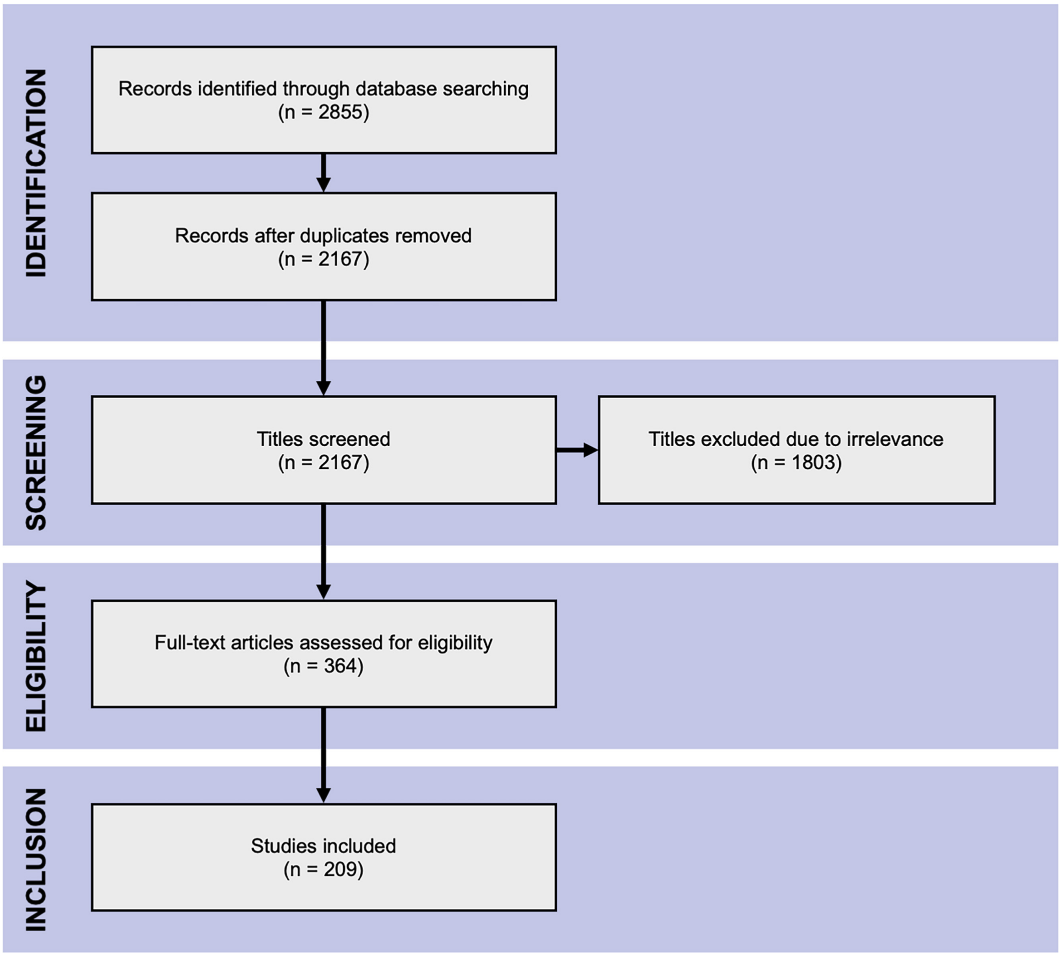

IDENTIFICATION
Records identified through database searching (n = 2855)
Records after duplicates removed (n = 2167)
SCREENING
Titles screened (n = 2167)
Titles excluded due to irrelevance (n = 1803)
ELIGIBILITY
Full-text articles assessed for eligibility (n = 364)
INCLUSION
Studies included (n = 209)

FIGURE 2
PRISMA flow chart. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow chart of literature search.

either alone or in combination with other metrics such as the Hausdorff distance or the Jaccard index (Figure 3H). Few studies (10%) provided working links to publicly available code, with a further 1% indicating that code was available on request.

Compliance with CLAIM

Results for compliance with the domains and selected individual criteria of CLAIM are summarized in Figure 4. The complete results are presented in Supplementary Table 2. The median compliance of all studies with all 42 criteria of CLAIM was 67% (IQR 59–73%). Notable results excluding non-applicable criteria are as follows.

Study description

Median compliance with the study description domain was 71% (IQR 63–86%). Almost all studies clearly indicated the use of AI methods (91%) and their objectives (94%). Where non-public datasets were used, only a minority of studies (36%) indicated whether these were prospective or retrospective. No studies provided access to a full study protocol. Sources of funding were declared in 82% of studies.

Dataset description

Median compliance with the dataset description domain was 63% (IQR 50–67%), the lowest of the four domains. The source of the dataset was reported in most studies (94%). While most studies provided eligibility criteria for included cases (74%), few studies reported their demographic and clinical characteristics (18%) or indicated the flow of these cases (10%) in sufficient detail. Details regarding the calculation of the intended sample size (4%) and how missing data were handled (9%) were also infrequently reported. The definition of the ground truth reference standard was provided in 68% of studies. Where non-public datasets were used, the source of ground truth annotations and annotation tools were stated in 55%

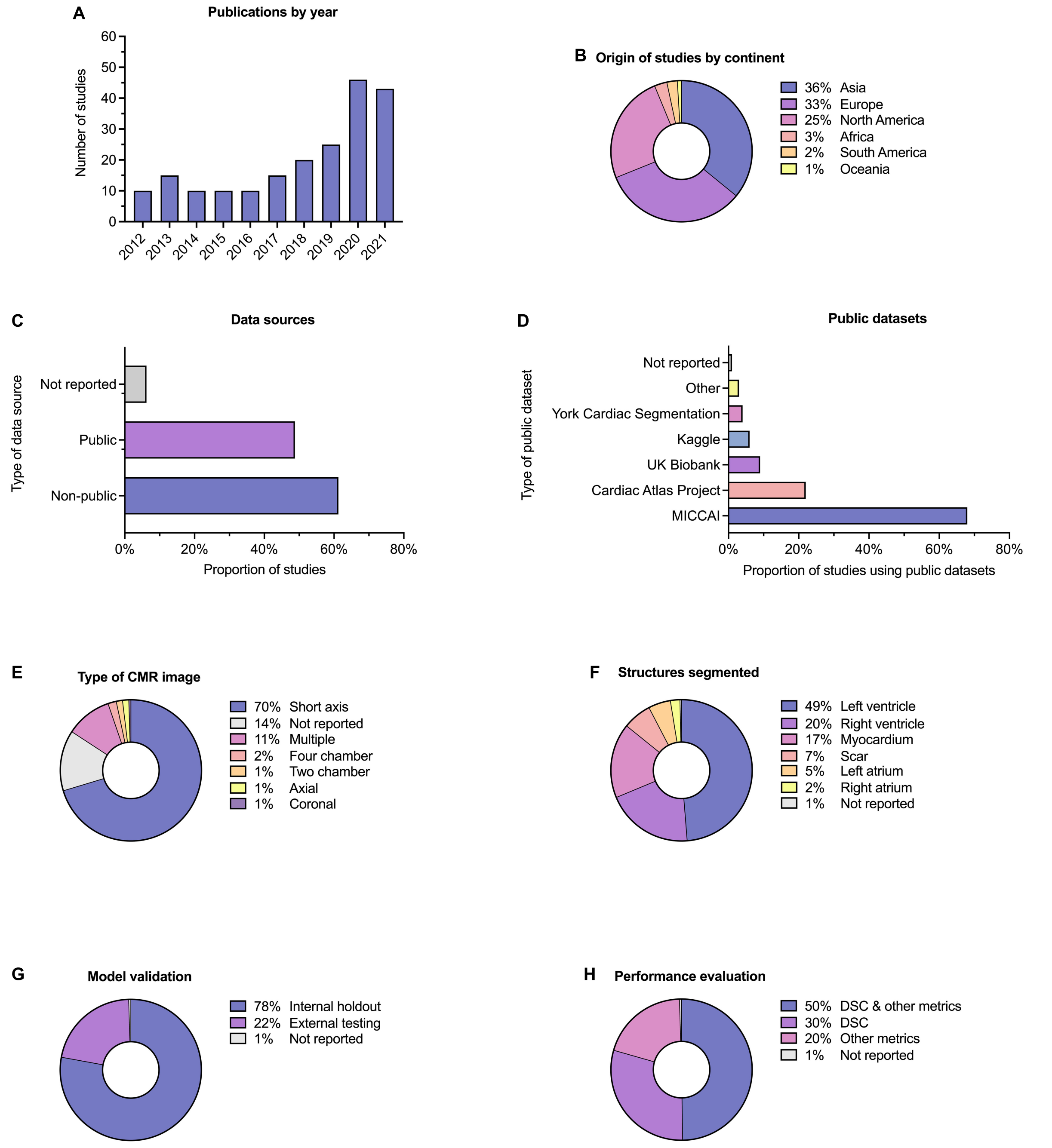


FIGURE 3
Descriptive information. Descriptive information for the 209 included studies. **(A)** Publication dates; five studies (2.4%) were included from early 2022 and are not indicated here. **(B)** Location of origin of studies. **(C)** Data sources; the proportion of studies which used public and non-public datasets is shown, with some studies having used multiple or combined datasets. **(D)** Public datasets used by studies, where relevant. **(E)** Type of CMR images used. **(F)** Cardiac structures segmented; some studies performed segmentation on multiple structures. **(G)** Method of model validation. **(H)** Method of model performance evaluation.

and 31% of studies respectively, with inter- and intra-rater variability reported in 42%. The majority of studies reported data preprocessing steps (94%), definitions of data elements (99.5%), how data were assigned to partitions (89%) and the level at which partitions were disjoint (87%).

Model description

Median compliance with the model description domain was 100% (IQR 80–100%), the highest of the four domains. The majority of studies provided details about the model used (95%), initialization of model parameters (92%), training

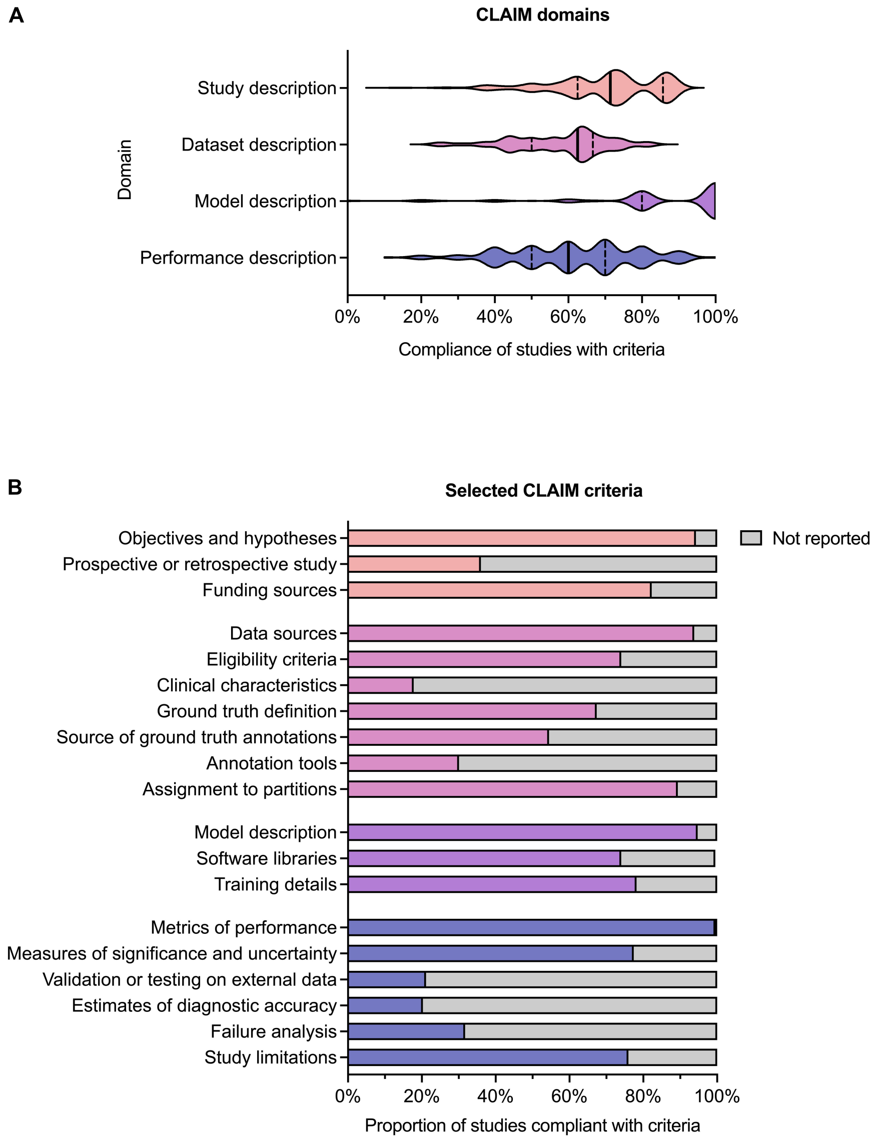


FIGURE 4
Compliance with CLAIM. **(A)** Violin plot showing compliance of the 209 included studies with the CLAIM criteria, grouped into domains of study, dataset, model and performance description. Median (*solid line*) and 1st and 3rd quartile (*dashed lines*) values are indicated. **(B)** Proportion of studies compliant with selected CLAIM criteria, grouped by domain (the titles of the individual criteria have been shortened for ease of reading).

approach (78%) and method of selecting the final model (92%). The software libraries, frameworks and packages used were reported in 74%.

Model performance

Median compliance with the performance description domain was 60% (IQR 50–70%). A minority of studies reported testing on external data (22%) Almost all studies provided metrics of model performance (99.5%). Most studies provided statistical measures of significance and uncertainty when reporting results (78%). Many studies provided forms of robustness or sensitivity analysis (61%) and methods for explainability and interpretability (64%). A minority of studies reported failure analysis for incorrectly classified cases (32%). Most studies discussed their limitations (76%) and implications for practice (76%).

Discussion

Poor reporting is a major source of research waste (10, 11) and ultimately may hinder advancement of AI research in the medical field. This systematic review evaluated the quality of reporting in AI studies involving automatic segmentation of structures on cardiac MRI. 209 studies were included from 2012 to early 2022. Each study was assessed for compliance with CLAIM, a checklist that attempts to provide a "best practice" framework for the reporting and publication of AI research in medical imaging (5). We identified major gaps in reporting and make a number of recommendations in order for this to be addressed (**Table 1**).

Accurate and sufficiently detailed descriptions of study materials and methods are of particular importance for AI studies in medical imaging to allow the assessment of reproducibility and reliability of results. Overall compliance with CLAIM was highest for the model description domain, with most studies providing a description of the model and details of training approaches. However, this was lowest for the dataset description domain, which indicated variable reporting of the data sources used to train and evaluate models.

A good understanding of data sources is a prerequisite for evaluating the validity of AI models. Although most studies identified their data sources, this was a significant omission in the studies that did not and one which greatly limits their interpretability. Public datasets were used in almost half of the studies, with the majority of these made available through segmentation challenges hosted by MICCAI (**Supplementary Table 3**). Public datasets contain previously de-identified and expertly contoured images, making them attractive to researchers. The proportion of studies using datasets from MICCAI challenges underlines its role as a driver for advancing the field. Importantly, the use of public datasets facilitates reproducibility and aids comparison between segmentation methods. However, public sources are not without their limitations. Public datasets consist of entirely retrospective data, which may place constraints on study design and model training. They are often small in size with limited demographic and clinical diversity, and therefore have inherent selection bias. Systematic biases affecting patient demographics are of serious concern in the application of AI methods to clinical practice. For example, a previous analysis of AI-based segmentation in CMR using a large-scale database found systematic bias for both sex and race (12) and similar biases have been reported for AI in radiographic imaging (13). The use of diverse datasets when training, validating and testing models is essential for generalizability and translation to clinical practice. A model trained on a dataset from one population does not guarantee equal performance on another. Multiple data sets, such as both retrospective and prospective, could be used in combination to improve the generalizability of AI models being trained. Even accounting for the use of public datasets, we found that few studies reported the intended sample size (which influences statistical power and reliability of results) or the demographic and clinical characteristics of the cases in each partition, (which indicates selection bias, confounders and generalizability). Providing summary information about the age and sex of cases is important, but may be insufficient in isolation. We noted that studies often lacked details about the proportions of cases with different pathologies, and the demographics for these groups. Furthermore, studies should not assume that readers are familiar with public datasets, and if these are used then detailed demographics and clinical characteristics should still be reported. The performance and validity of any model depend on the data on which it is trained and the data sources, including the rationale behind their choice and the intended sample size, should be clearly indicated. Study methodology must be reported in sufficient detail to enable accurate reproduction of results. Notably, the

TABLE 1 Recommendations for study reporting. Main recommendations for AI study reporting are based on the gaps in the literature identified in this systematic review.

	Recommendation	Importance
General	Utilize a reporting framework (e.g., CLAIM).	Comparability of studies.
	Use of consistent and descriptive terminology.	Accessibility and comparability of studies.
Data sources	Describe the source of data, including patients' eligibility criteria, their numbers and demographic and clinical characteristics.	Contextualizing model performance and generalizability.
	Clarify the number of scans and the flow of both patients and scans into different datasets (e.g., training, validation, and testing).	Understanding model performance and generalizability.
	Use publicly available datasets.	Comparability of models against a common benchmark.
Model training and evaluation	Describe the neural network, software packages and libraries in sufficient detail.	Study reproducibility.
	Define how the reference contours were generated, the experience of the annotator and annotation tools used.	Understanding model performance and generalizability.
	Explain the method of model training and performance.	Understanding model performance and generalizability.
	Test the model performance on external data with different characteristics to the training data.	Study and model reliability. Understanding model generalizability. Implementation in clinical practice.
	Perform failure analysis and report the limitations of the model.	Understanding model performance and generalizability.
	Publication of open-source code.	Understanding model performance and generalizability.

definition of the ground truth reference standard, the source of ground truth annotations and the annotation tools used were absent in a substantial number of studies. Understanding the structures included in the ground truth contours and the expertise of the annotator is essential in evaluating the training process and ultimately contextualizing the model's performance. The proportions of studies that provided sufficiently detailed descriptions of the ground truth and its source were lower than expected for the field. For example, judging from the figures present in the included studies, ventricular trabeculations were usually included in the blood pool contours, although few studies described this process. Similarly, many studies failed to report the specific type of image used for ground truth annotation and model training and testing. While this could be inferred from figures, it remains essential information for understanding models and their generalizability. Finally, only a handful of studies indicated how missing data were handled and no studies indicated where a full study protocol could be accessed.

Detailed description of model training and performance is expected in this field. Testing model performance on external data was performed in less than a quarter of all studies. Model generalizability can only be fully evaluated when performance is assessed in demographic and clinical populations different from the original training cohort. The reported external datasets were small and captured only limited variations in imaging appearances. This represents a major hurdle to overcome before AI models can be implemented in clinical practice. We also noted subjectively that many publications used the terms "validation" and "test" interchangeably, or failed to distinguish these methods clearly. Regarding the use of data in AI studies, a validation set is used to optimize hyperparameters and performance between training epochs, while a testing set is used to assess the performance of the final model. The lack of consistent terminology in studies can limit the interpretability of their models and blur the distinction between internal holdout and external testing methods. Additionally, few studies reported failure analysis of incorrectly classified cases, suggesting that most did not explore the reasons for model underperformance. Furthermore, the vast majority of studies did not discuss the limitations of their methods, limiting their transparency. Open publishing of source code is a contentious topic in AI research and was only provided in one in ten of all studies. The public availability of code aids transparency, assists peer review and facilitates the development of new models, but bears important implications for ownership and rights.

The use of reporting frameworks, such as CLAIM, can be beneficial. For example, they may help to inform study design and highlight areas that may require rectification prior to dissemination of results. Frameworks assist standardization in reporting, improving comparability and interpretability by the wider scientific community. Study accessibility is also an important consideration in advancing the field. Regardless of journal type, AI studies in medical imaging need to cater for a broad potential readership, from clinicians to computer scientists. More standardized reporting and the use of consistent and accessible terminology are important in this regard.

We acknowledge limitations in this systematic review. Firstly, this review focused solely on AI segmentation in CMR studies. However, these findings are likely to apply to AI studies in other cardiac imaging modalities, such as echocardiogram, CT coronary angiography or nuclear myocardial perfusion studies. Furthermore, given that AI studies in chest imaging have shown similar shortcomings in reporting quality (14), our findings may be more broadly relevant to AI studies in medical imaging. Secondly, while our systematic search aimed to identify all published AI CMR segmentation studies, the body of unpublished, pre-print or technical conference literature is vast. A Github or arxiv.org search reveals numerous segmentation attempts of varying levels of reporting quality and beyond the scope of this review to capture. Thirdly, even despite the use of structured tools such as CLAIM, there remains an element of subjectivity in determining report quality, such as the amount of information required for a study to be deemed reproducible.

Conclusion

This systematic review highlights the variability in reporting and identifies gaps in the existing literature of studies using AI segmentation of CMR images. We identified several key items that are missing in publications—most strikingly poor description of patients included in the training and validation of AI models and inadequate model failure analysis—which may limit study transparency, reproducibility and validity. This review supports closer adherence to established frameworks for reporting standards, such as CLAIM. In light of these findings, we have presented a number of recommendations for improving the quality of reporting of AI studies in both CMR and the wider field of cardiac imaging.

Data availability statement

The original contributions presented in this study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

Author contributions

SA and AS conceived the idea and need for the systematic review and contributed to the study conception and design. SD and SA performed the protocol registration in PROSPERO. SA created the search strategy and performed the literature search. SA, SD, MaS, MiS, and AqM performed screening and eligibility assessments independently. SA, SD, AhM, AqM, MaS, MiS, SJ, MG, VR, and HA evaluated the included studies and collected relevant data. SA, AhM, MaS, AqM, and SJ performed the material preparation and analysis. SA, AhM, SJ, and MaS drafted the manuscript, figures, and tables. SA, AhM, AqM, VR, and SD wrote the first draft of the manuscript. SA and AhM wrote the final draft, taking into account comments and suggestions from experts in the field AS, DO'R, HL, RG, MM, and PG. All authors contributed to the interpretation of data, commented on previous versions of the manuscript, read and approved the final manuscript, took part in the critical review and drafting of the manuscript, and have read and approved the final manuscript.

Funding

This study was supported by the NIHR grant AI_AWARD01706, Wellcome Trust grants 215799/Z/19/Z and 205188/Z/16/Z, Medical Research Council grant MC-A658-5QEB0, and British Heart Foundation grant RG/19/6/34387. The funders did not have any role in the design and conduct of the study; in the collection, analysis, and interpretation of the data; or in the preparation, review, and approval of the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcvm.2022.956811/full#supplementary-material

References

1. O'Regan DP. Putting machine learning into motion: applications in cardiovascular imaging. *Clin Radiol.* (2020) 75:33–7. doi: 10.1016/j.crad.2019.04.008

2. Reardon S. Rise of robot radiologists. *Nature.* (2019) 576:S54–8. doi: 10.1038/d41586-019-03847-z

3. Chen C, Qin C, Qiu H, Tarroni G, Duan J, Bai W, et al. Deep learning for cardiac image segmentation: a review. *Front Cardiovasc Med.* (2020) 7:25. doi: 10.3389/fcvm.2020.00025

4. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer.* (2018) 18:500–10. doi: 10.1038/s41568-018-0016-5

5. Mongan J, Moy L, Kahn CE Jr. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell.* (2020) 2:e200029. doi: 10.1148/ryai.2020200029

6. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ.* (2021) 372:n71. doi: 10.1136/bmj.n71

7. Alabed S, Alandejani F, Dwivedi K, Karunasaagarar K, Sharkey M, Garg P, et al. Validation of Artificial Intelligence Cardiac MRI Measurements: relationship to Heart Catheterization and Mortality Prediction. *Radiology.* (2022):212929. [Online ahead of print]. doi: 10.1148/radiol.212929

8. Alandejani F, Alabed S, Garg P, Goh ZM, Karunasaagarar K, Sharkey M, et al. Training and clinical testing of artificial intelligence derived right atrial cardiovascular magnetic resonance measurements. *J Cardiovasc Magnetic Resonan.* (2022) 24:25. doi: 10.1186/s12968-022-00855-3

9. Fonseca CG, Backhaus M, Bluemke DA, Britten RD, Chung JD, Cowan BR, et al. The Cardiac Atlas Project—an imaging database for computational modeling and statistical atlases of the heart. *Bioinformatics.* (2011) 27:2288–95. doi: 10.1093/bioinformatics/btr360

10. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. *Obstet Gynecol.* (2009) 114:1341–5. doi: 10.1097/AOG.0b013e3181c3020d

11. Glasziou P, Chalmers I. Research waste is still a scandal—an essay by Paul Glasziou and Iain Chalmers. *BMJ.* (2018) 363:k4645. doi: 10.1136/bmj.k4645

12. Puyol-Antón E, Ruijsink B, Mariscal Harana J, Piechnik SK, Neubauer S, Petersen SE, et al. Fairness in cardiac magnetic resonance imaging: assessing sex and racial bias in deep learning-based segmentation. *Front Cardiovasc Med.* (2022) 9:859310. doi: 10.3389/fcvm.2022.859310

13. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc Natl Acad Sci USA.* (2020) 117:12592–94. doi: 10.1073/pnas.1919012117

14. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Machine Intellig.* (2021) 3:199–217. doi: 10.1038/s42256-021-00307-0

TYPE Original Research
PUBLISHED 27 July 2022
DOI 10.3389/fcvm.2022.894503

Predicting post-contrast information from contrast agent free cardiac MRI using machine learning: Challenges and methods

Musa Abdulkareem[1,2,3]*, Asmaa A. Kenawy[1,2], Elisa Rauseo[1,2], Aaron M. Lee[1,2], Alireza Sojoudi[4], Alborz Amir-Khalili[4], Karim Lekadir[5], Alistair A. Young[6], Michael R. Barnes[7], Philipp Barckow[4], Mohammed Y. Khanji[1,2,8], Nay Aung[1,2] and Steffen E. Petersen[1,2,3,9]

[1]Barts Heart Centre, Barts Health National Health Service (NHS) Trust, London, United Kingdom, [2]National Institute for Health Research (NIHR) Barts Biomedical Research Centre, William Harvey Research Institute, Queen Mary University of London, London, United Kingdom, [3]Health Data Research UK, London, United Kingdom, [4]Circle Cardiovascular Imaging Inc., Calgary, AB, Canada, [5]Artificial Intelligence in Medicine Lab (BCN-AIM), Faculty of Mathematics and Computer Science, University of Barcelona, Barcelona, Spain, [6]Department of Biomedical Engineering, King's College London, London, United Kingdom, [7]Centre for Translational Bioinformatics, William Harvey Research Institute, Faculty of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom, [8]Newham University Hospital, Barts Health National Health Service (NHS) Trust, London, United Kingdom, [9]The Alan Turing Institute, London, United Kingdom

Objectives: Currently, administering contrast agents is necessary for accurately visualizing and quantifying presence, location, and extent of myocardial infarction (MI) with cardiac magnetic resonance (CMR). In this study, our objective is to investigate and analyze pre- and post-contrast CMR images with the goal of predicting post-contrast information using pre-contrast information only. We propose methods and identify challenges.

Methods: The study population consists of 272 retrospectively selected CMR studies with diagnoses of MI (n = 108) and healthy controls (n = 164). We describe a pipeline for pre-processing this dataset for analysis. After data feature engineering, 722 cine short-axis (SAX) images and segmentation mask pairs were used for experimentation. This constitutes 506, 108, and 108 pairs for the training, validation, and testing sets, respectively. We use deep learning (DL) segmentation (UNet) and classification (ResNet50) models to discover the extent and location of the scar and classify between the ischemic cases and healthy cases (i.e., cases with no regional myocardial scar) from the pre-contrast cine SAX image frames, respectively. We then capture complex data patterns that represent subtle signal and functional changes in the cine SAX images due to MI using optical flow, rate of change of myocardial area, and radiomics data. We apply this dataset to explore two supervised learning methods, namely, the support vector machines (SVM) and the decision tree (DT) methods, to develop predictive models for classifying pre-contrast cine SAX images as being a case of MI or healthy.

OPEN ACCESS

EDITED BY
Ali Yilmaz,
University Hospital Münster, Germany

REVIEWED BY
Lennart Tautz,
Fraunhofer Institute for Digital Medicine MEVIS, Germany
Fahmi Khalifa,
Mansoura University, Egypt

*CORRESPONDENCE
Musa Abdulkareem
musa.abdulkareem@nhs.net

SPECIALTY SECTION
This article was submitted to Cardiovascular Imaging, a section of the journal Frontiers in Cardiovascular Medicine

RECEIVED 11 March 2022
ACCEPTED 27 June 2022
PUBLISHED 27 July 2022

CITATION
Abdulkareem M, Kenawy AA, Rauseo E, Lee AM, Sojoudi A, Amir-Khalili A, Lekadir K, Young AA, Barnes MR, Barckow P, Khanji MY, Aung N and Petersen SE (2022) Predicting post-contrast information from contrast agent free cardiac MRI using machine learning: Challenges and methods. *Front. Cardiovasc. Med.* 9:894503. doi: 10.3389/fcvm.2022.894503

Results: Overall, for the UNet segmentation model, the performance based on the mean Dice score for the test set (*n* = 108) is 0.75 (±0.20) for the endocardium, 0.51 (±0.21) for the epicardium and 0.20 (±0.17) for the scar. For the classification task, the accuracy, F1 and precision scores of 0.68, 0.69, and 0.64, respectively, were achieved with the SVM model, and of 0.62, 0.63, and 0.72, respectively, with the DT model.

Conclusion: We have presented some promising approaches involving DL, SVM, and DT methods in an attempt to accurately predict contrast information from non-contrast images. While our initial results are modest for this challenging task, this area of research still poses several open problems.

KEYWORDS

CMR, contrast, contrast-free, deep learning, machine learning, support vector machines, decision tree

Introduction

Background and objectives

Cardiovascular diseases (CVDs) are a major cause of death in the world in 2022, causing approximately 18.6 million deaths (31% of all deaths) annually according to the World Heart Federation (1). Ischemic heart disease (IHD) was responsible for almost half of all cardiac deaths in 2019 (2). A common and important consequence of IHD is myocardial infarction (MI) defined pathologically as myocardial cell death due to prolonged ischemia which can lead to the loss of contraction of that damaged portion of the heart muscle.

Cardiac magnetic resonance (CMR) is an imaging modality that has proven to be very effective in diagnosing MI through visualization of the regional myocardial scar allowing the determination of the presence, location, and extent. Currently, administering contrast agent, using gadolinium-based chelates (gadolinium), is necessary for diagnosing MI with CMR. This technique relies on the relative gadolinium accumulation in areas of necrosis and fibrosis following myocardial damage. The presence and pattern of the gadolinium contrast can vary, with subendocardial or transmural late gadolinium enhancement (LGE) images usually indicating fibrosis caused by previous coronary ischemic events or MI. Eliminating the need of contrast administration could in several ways benefit many patients, such as, patients who cannot be safely given a contrast agent due to allergies or severe kidney disease, and could improve safety and patient experience (avoiding need for intravenous cannulation) and costs of cardiovascular healthcare. Moreover, a typical contrast CMR scan takes approximately 35–45 min, whereas without contrast it could take approximately half the time, leading to shorter times in the scanner. **Figure 1** shows some examples of some LGE images.

Machine learning (ML) is a set of techniques in artificial intelligence (AI) which refers to computer algorithms with human-like intelligence developed to accomplish specific tasks. Deep learning (DL) algorithms, which are a set of ML techniques based on neural networks, are useful for medical imaging related tasks such as those involving diagnosis of diseases (3, 4). Such diagnoses often consist of one or a combination of several ML methods involving image segmentation or classification. Image segmentation involves identifying and marking the region of interest (ROI) while image classification involves extracting features from the ROI and uses those features as a basis for classifying patients or diseases. The UNet (5) and the ResNet (6) are examples of popular segmentation and classification algorithms, respectively. The use of ML methods such as support vector machines (SVM) (7, 8) and decision trees (DT) (9) are common in applications involving identification of latent relationships in patient phenotypes (10) and development of predictive models (11, 12). Models based on DTs methods are easy to interpret (i.e., are white box models) while those based on SVM methods are versatile and effective including in high dimensional spaces such as when the number of features is greater than the number of sample points.

In this study, our goal is to investigate and analyze pre- and post-contrast CMR images to predict post-contrast information (i.e., presence, location, and/or extent of MI scar) from pre-contrast information only (i.e., without having

Abbreviations: CMR, cardiac magnetic resonance; CVD, cardiovascular disease; DT, decision tree; DL, deep learning; DSC, dice similarity coefficient (Dice score); LGE, late gadolinium enhancement; LV, left ventricle; ML, machine learning; MI, myocardial infarction; PFI, permutation feature importance; ROI, region of interest; SAX, short axis; SI, signal intensity; SVM, support vector machines; TT, trigger time.

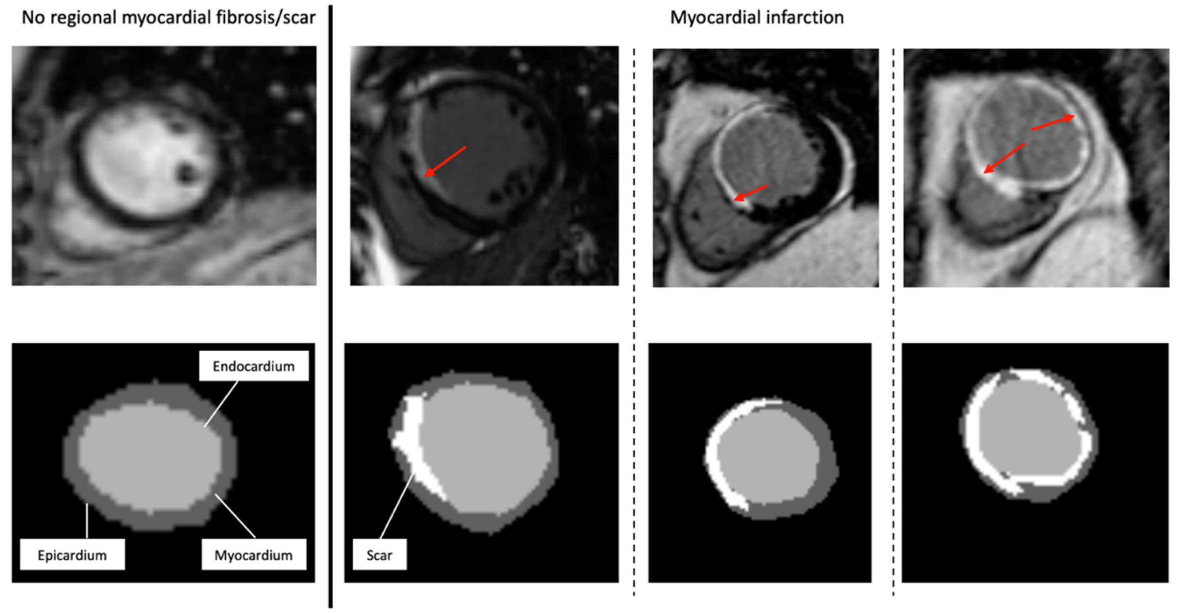


FIGURE 1
Representative short-axis late gadolinium enhancement-cardiac magnetic resonance images (LGE-CMR) of various possible locations and extents of myocardial scar. The top images are the original LGE images, with the red arrows indicating the distribution of LGE (scar) in the LV wall segments. At the bottom are the segmentation masks (expert manual contouring) with the scarred regions in white.

to administer contrast to patients) using ML. This subject has recently gained the interest of many researchers (13–21) but despite these interests, many open problems and challenges on this subject still exist. The accurate prediction of contrast information without contrast administration with ML methods is very challenging for many reasons. Qualitatively assessed interpretation by expert humans of the images are recorded in free text and are highly variable, while quantitative ground-truth of these images are not recorded. In addition, format, and quality standardization of CMR imaging data for ML does not exist.

In this article, in order to predict contrast information from non-contrast CMR images, we describe a pipeline for processing routinely acquired pre-contrast cine short-axis (SAX) CMR and post-contrast LGE images so that these images can be used for preparing ground truth for training models that can predict the location of the epicardial and endocardial walls and the location and the extent of the scar. We explore two approaches, namely, segmentation and classification approaches. For the segmentation approach, we use the popular UNet DL model in our attempts to discover the extent and location of the scar given the pre-contrast cine SAX image frames. For the classification approach, we use the ResNet50 classification model in our attempts to distinguish between the ischemic and non-ischemic cases from the non-contrast cine SAX image frames. The performance of the ML algorithms can be significantly improved by extracting features that could be useful or relevant in the model training process for solving particular problems. This feature extraction can be described as a data transformation process (and may or may not require domain knowledge of the problem). Some advantages of the feature extraction process in addition to performance and predictive accuracy improvements includes dimensionality reduction of the feature space, noise reduction, and improvement in the speed of convergence of the learning algorithms (22–24). Thus, in other to use other ML approaches for the classification task, we extract data from the cine SAX images that capture complex patterns representing subtle signal and functional changes in the cine CMR due to myocardial tissue-specific abnormalities and used for qualitative prediction. In particular, we focus on capturing three sets of data from the cine SAX images, namely, the optical flow data, the rate of change of myocardial area, and the radiomics data. Optical flow (image velocity) measurement is a fundamental method in the processing of a sequence of images (successive frames) and its goal is to compute an approximation of a 2D motion field from spatiotemporal data of image intensities (25). Radiomics features are high-dimensional handcrafted quantitative features that are based on mathematical and statistical methods extracted from images (26). They have recently been used on a wide range of problems, such as identifying the causes of myocardial hypertrophy (27) or detecting fibrosis in patients with hypertrophic cardiomyopathy (28). We use the three sets of data to explore two supervised ML methods, namely, the SVM and the DT methods to develop predictive models for classifying pre-contrast cine SAX images as being a case of MI or being free of myocardial scar.

State of the art

Recently, the attempt to predict post-contrast information without contrast administration is attracting the attention of ML researchers and clinicians alike. Efforts to tackle this challenging task either treats the problem as a pixel-wise tissue identification problem (29) where the extent and location of scar is sought or as an image synthesis problem (14) which involves the generation of images predicting what the post-contrast image would look like.

Changes in mechanical properties of myocardium caused by infarction can lead to regional wall motion abnormalities. This phenomenon inspires the pixel-wise tissue identification approach given in Xu et al. (29) where the proposed DL architecture consists of three connected function layers: the heart localization layers which automatically crop the ROI (i.e., the LV from the cine SAX image frames); the motion feature extraction layers which use long short-term memory (LSTM) recurrent neural networks and optical flow techniques to build local and global motion features through local intensity changes and global intensity changes between adjacent images; and the fully connected layers which learns to predict tissue identities (that is, infarct or not) in each pixel.

In Xu et al. (13), a so-called deep spatiotemporal generative adversarial network (DSTGAN) was used to simultaneously segment and quantify (i.e., infarct size, percentage of infarct size, percentage of segments, perimeter, centroid, major axis length, minor axis length, orientation, and transmurality) MIs directly from the cine MR image. The DSTGAN uses the conditional generative adversarial network (cGAN) DL approach and the input images are cine SAX images. After a network for heart localization process, the DSTGAN technique consists of three components: (i) a multi-level and multi-scale spatiotemporal variation encoder (which actually uses 25 temporal frames from a single slice location), (ii) the top-down and cross-task generator, and (iii) three task labels relatedness discriminators.

In Zhang et al. (14), a cGAN DL model was trained to tackle the challenge using an image synthesis approach. The generator part of the cGAN uses encoder-decoder architecture with cine SAX images, inversion recovery-weighted (IRW) images and T1 mapping images as input to produce virtual native enhancement (VNE) images. The discriminant part of the cGAN uses the VNE images as input and conditioned with LGE images during the model training process. The limitation of this approach is that it requires the acquisition of additional CMR images (IRW and T1 mapping images) that are not (yet) typically acquired in routine CMR imaging in the diagnosis of MI.

Researchers have developed a CNN-based model that identifies ischemic scar slices in computed tomography (CT) angiography of the LV without any contrast agents (15). The model's algorithm uses LGE images from CMR as ground truth (i.e., a CT-MRI paired dataset) to determine the presence or absence of scar for the binary classification problem. However, this promising approach does not give an idea of the extent or the percentage myocardium affected by scar.

In addition to wall motion abnormalities, myocardial features such as myocardial wall thinning and myocardial lipomatous metaplasia that lead to chemical shift artifacts have been shown to characterize MI (30–32) and suggest that it may be possible for radiomics analysis to identify ischemic scar from non-contrast cine CMR images (16–19). Thus, in Baessler et al. (16), for example, researchers have proposed radiomics texture analysis for the diagnosis of subacute and chronic MI on non-contrast cine CMR images. The approach analyses end-systolic cine SAX images using stepwise dimension reduction (the Boruta feature selection algorithm and the recursive feature elimination method – where a classifier, random forest classifier in this case, is recursively trained and the feature with the smallest ranking score is removed at each iteration), logistic regression ML method and correlation analysis to select features that will enable the classification of end-systolic cine images as cases with or without myocardial scar. The limitation of this approach is that it focuses on texture radiomics features only (i.e., ignoring other radiomics features such as shape) even though features such as wall thinning in myocardial regions are already reported as potential signs of myocardial scar (30). The approach also ignores wall motion abnormalities.

Another radiomics texture analysis approach was proposed in Larroza et al. (17) for differentiating acute MI from chronic MI cases using both contrast LGE images and non-contrast cine CMR images. The approach analyzed LGE images by developing three ML classifiers namely, random forest, SVM with Gaussian kernel, and SVM with polynomial (degree = 3) kernel classifiers. The three counterpart classifiers were developed for non-contrast cine CMR. The recursive feature elimination method with SVM classifier was used as the feature selection technique. Similar to Baessler et al. (16), this approach focused on texture radiomics features only. The major limitation of the texture analysis approaches in Baessler et al. (16) and Larroza et al. (17) is that they have included orientation dependent texture features (obtained from 2D ROI delineation) in their analysis which influence the results if the SAX views are not acquired in a standardized position as presented in those reports.

In Larroza et al. (18), texture analysis was used for classifying myocardial regions of patients suffering from chronic MI into three categories (segments) namely, remote segments (LGE = 0%), viable segments (0 < LGE < 50%), and non-viable segments. LGE CMR images were used to prepare the ground-truth for the non-contrast cine SAX myocardium regions using the 17-segment model. A SVM with radial basis function kernel classifier was trained. Importantly, texture features were calculated in their rotation invariance form in order to evade image rotation as a possible source of bias. Time dimension available in cine sequences are also included as part of analysis in order to take advantage of the information on temporal dimension. The recursive feature elimination method with SVM classifier was used as the feature selection technique. The proposed method also focused on texture radiomics features only.

In Di Noto et al. (19), researchers evaluated radiomics features of LGE regions of CMR images for distinguishing between MI and myocarditis. K-nearest neighbor, linear discriminant analysis (LDA), neural network (multilayer perceptron), SVM, and TreeBagger DT are the five different ML algorithms investigated in the report and the recursive feature elimination method was used as the feature selection technique. However, these analyses were carried out on LGE images and not on non-contrast cine images.

In Avard et al. (20), researchers used radiomics analysis to extract shape, first-order, and texture features for the differentiation of MI and viable tissues (normal) cases in the LV using non-contrast cine CMR images. The whole of the left ventricular myocardium (3D volume) in end-diastolic volume phase was used for the analysis. Ten ML algorithms were investigated for the classification tasks and the SVM and the logistic regression-based models show superior performance compared to other methods on evaluation dataset.

A state-of-the-art review of the methods for delineating LV scar without contrast administration can be found in Wu et al. (21). In general, as far as radiomics analysis is concerned, no specific subset of features has been found to be reliable discriminative of myocardial scar from disease-free regions of the myocardium. Research in this area is still ongoing and progress in this field and progress toward clinical application will require standardization of the discriminative features and evaluation of proposed models to ensure generalizability.

Materials and methods

Data acquisition and analysis tools

Study population

The study population consists of 272 retrospectively selected CMR studies with diagnoses of MI (n = 108) and healthy controls (n = 164) from the Barts BioResource between January 2015 to June 2018. Barts BioResource is a local biorepository of Barts Heart Centre (Barts Health NHS Trust, London, United Kingdom) that holds data from prospectively consented (written) patients for cardiovascular research (Ethics REC reference: 14/EE/0007). All images were de-identified prior to analysis.

Cardiac magnetic resonance acquisition and myocardial infarction diagnosis

Cardiac magnetic resonance examinations were obtained using 1.5T and 3T scanners (Siemens Healthineers, Germany). The steady-state free precession cine images for SAX were analyzed using CVI42® research prototype software 5.11 built-in ML tool. In order to diagnose for MI, firstly, CMR images – the cine short axis (SAX) images and images of horizontal long-axis (HLA) 4 chamber view and the vertical long-axis (VLA) 2 chamber view – of the patient are taken. The cine SAX images are spatio-temporal, meaning that, for each slice location (in space) of the left ventricle (LV), several images are taken (in time) over the cardiac cycles. Next, gadolinium contrast agent of 0.1 to 0.2 mmol/kg is administered to the patient intravenously. Then, after around 10 min wait, the second set of CMR images are taken using a conventional 2D breath-hold technique. This second set of post-contrast SAX images only have spatial component (have no time component, i.e., one slice for each slice location) and often referred to as the LGE images. It is these LGE images that are scanned for scar in the heart muscle and predominantly of the LV. Both the cine SAX and LGE images cover the whole heart. The diagnosis of MI is made in accordance with the standard definition given in (33, 34).

Machine learning software tools

All experiments were conducted on a Nvidia Tesla M40 machine using Python programming language with the following packages: Scikit-learn (Version 0.24.0) (35) was used to implement SVM and DT, Pyradiomics (Version v3.0.1) (36) was used for radiomics feature extraction, TensorFlow 2.0 Python API machine learning framework (Version 2.7.0) (37) was used for implementing UNet and ResNet50 DL architectures, and MATLAB (Version 9.7.0.1586710, R2019b, Update 8) was used for image registration as part of the ground truth data preparation pipeline.

Segmentation approach

We carry out experiments with DL-based image segmentation and classification architectures, the UNet and the ResNet50, respectively, in order to derive contrast information without contrast administration.

Ground truth data pre-processing pipeline

The three key steps of the image pre-processing pipeline for the ground truth images of the supervised ML problem are illustrated in **Figure 2** and described as follows:

1. The pre-contrast cine SAX images and their corresponding post-contrast LGE images for each slice location are extracted. Each cine SAX image is associated with a trigger time (TT), i.e., the slice acquisition time during the cardiac cycle with respect to the peak of the R wave; the peak of the R wave coincides with the early ventricular systole (in **Figure 2**, for example, the TT in ms for the 10 cine SAX images shown are given at the top of each image). As such the cine SAX image frame of a given slice location

whose TT approximately "matches" the TT of the LGE image of the same slice location is selected. The idea is that the heart muscle is at approximately the same position for the two images.

2. The LGE image is then registered with the selected cine SAX image as reference; that is, the LGE image is transformed and resampled into the coordinate system of the cine SAX image (the LGE and cine SAX images are the "moving" and the "fixed" images, respectively, in image registration terminology). The image registration process is an affine transformation consisting of translation, rotation, scale, and shear using the one-plus-one evolutionary algorithm (38) as the optimizer and the Mattes mutual information algorithm (39) as the mutual information metrics (40). MATLAB's "imregtform" function, for example, can be used to accomplish this image registration process. With image registration completed, the registered LGE image now has the same orientation, scale, and size with its matching cine SAX image. This image registration aims to correct the spatio-temporal misalignment between the pre-contrast matching cine SAX image and the post-contrast LGE image.
3. The registered LGE image is then contoured to mark regions of the epicardium, the endocardium, and the scar. Manual image segmentation was undertaken by trained observers (ER and AK). The LV structures were manually segmented to obtain three labels, namely, the LV cavity (the endocardial wall), the myocardium (the epicardial wall), and the scar. The fourth label (the background) is the non-segmented part of the image.

Given that the patients may have moved (even if slightly) between the pre- and post-contrast image acquisition, the slice location before and after contrast are not exactly the same. A further step of quality control action is taken by removing those slices with significant spatial mismatch between the cine SAX and LGE images. This quality control step was carried out by manual visual inspection. We are then left with 722 cine SAX images and segmentation mask pairs from the 272 subjects. It should be noted that the cine SAX images included images across all slice locations (i.e., all slices between and including the basal and the apical slices).

For the image segmentation task of marking the extent and location of the scar as well as the epicardial and endocardial

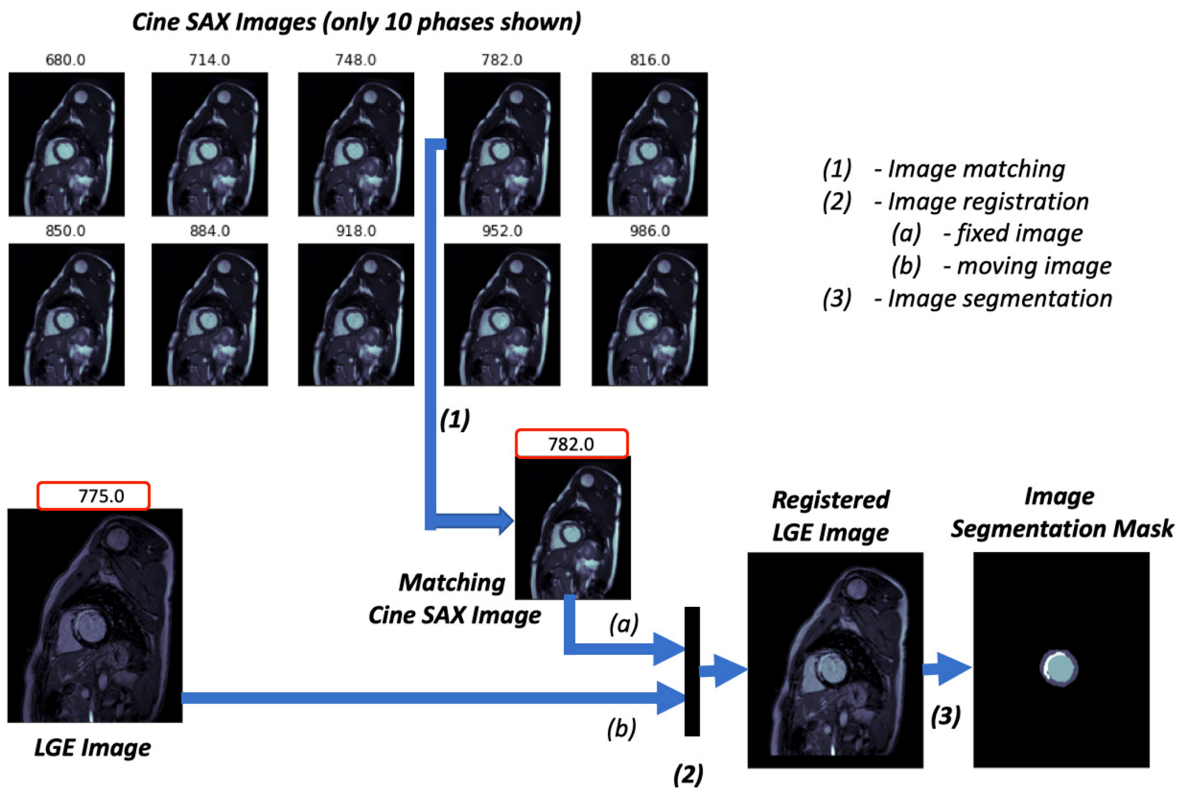


FIGURE 2
Image ground truth data pre-processing pipeline. The trigger times (TT) in ms for the 10 cine short axis (SAX) images shown are given at the top of each image. The cine SAX frame with closest TT match to the late gadolinium enhancement (LGE) image's TT is selected (1); this is followed by an image registration process (2), where the "fixed" image and the "moving" image of the registration process are denoted (a) and (b), respectively. The regions of the epicardium, the endocardium, and the scar are segmented in (3).

walls, the model training involves feeding our model with cine SAX images as inputs and their corresponding masked registered LGE images as the ground truth. The model prediction involves feeding the trained model with cine SAX stack as input so that it can predict contoured LGE masks in its output. The segmentation masks are not required for the image classification task, which involves categorizing pre-contrast cine SAX images as cases with MI or non-MI. For any set of cine SAX images, we determined from its corresponding LGE image of the same slice location whether it contained scar or not and then labeled it as such (i.e., Class 0 for non-MI cases, and Class 1 for MI cases). The cine SAX stack for each slice location contained up to 32 image frames (phases) for both segmentation and classification tasks. For the cases with fewer than 32, empty images (zero arrays) were appended with the stack to make 32 frames.

UNet segmentation model

The UNet architecture given in **Figure 3** is used as the image segmentation model, wherein the cine SAX frames are the input, and the ground truth is a segmentation mask that marks the regions of the epicardium, the endocardium and, if present, the scar. The UNet architecture includes batch normalization following the convolutional layers to enhance robustness of the model and drops out 30% hidden neurons in the first three consecutive up-sampling convolutional layers of the architecture to avoid problems associated with model overfitting. The output of the model yields an image segmentation mask with four channels: the background pixels labeled 0, the myocardium labeled l, the LV cavity labeled 2, and the scar labeled 3. The total number of parameters of the model is 75,019,204 out of which 75,011,524 parameters are trainable.

The model training settings are as follows: the 32 input images were resized to 224 × 224. The images were rotated up to ±60° and their intensities normalized as part of the on-the-fly data augmentation. In the normalization of the intensities, similar to the normalization used in Wolterink et al. (41) for the segmentation of cine CMR images, each image has been normalized between (0.0, 1.0) according to the 1 and 99% percentile of intensities in the image. The parameters of the models were randomly initialized, and training proceeded for 100 complete epochs using a batch size of 32 cine SAX images – segmentation mask pairs. The optimization method used was the Adam optimizer, with an initial learning rate of 0.001, decreasing exponentially at a rate of -0.1 after the first 5 epochs. If 30 epochs elapsed with no decrease in the loss function, training was set to cease and the weights from the best epoch is restored as the model's weights. With the total of 722 cine SAX images and segmentation mask pairs, we have used 70, 15, and 15% as the training, validation, and testing (evaluation) sets (representing, 506, 108, and 108 pairs), respectively.

Channel-weighted (class weighted) dice similarity coefficient (DSC or Dice score) function was used as the loss function. Image segmentation accuracy can be evaluated using DSC, which can be defined in terms of the per pixel classification for the i-th channel of a 2D segmentation mask as follows:

$$DSC_i = \frac{2\sum_{n=1}^{N} y_{i_n}\,\hat{y}_{i_n}}{\sum_{n=1}^{N} y_{i_n} + \sum_{n=1}^{N} \hat{y}_{i_n}} \tag{1}$$

where y_{i_n} and $\hat{y}_{i_n}$ are the ground truth mask and the predicted mask (the posterior probability obtained after the application of the "softmax" activation function on the output layer of the model), respectively, and N is the number of pixels in the mask (224 × 224). The dice loss of the i-th channel, $1 - DSC_i$, can be written as:

$$l_i = 1 - \frac{2\sum_{n=1}^{N} y_{i_n}\,\hat{y}_{i_n}}{\sum_{n=1}^{N} y_{i_n} + \sum_{n=1}^{N} \hat{y}_{i_n}} \tag{2}$$

The channel-weighted dice loss function L for the model can then be written for the 4 channels as follows:

$$L = \sum_{i=1}^{4} \beta_i l_i \tag{3}$$

where β_i is the associated with channel i. In our case, we defined $\beta_1 = 0.15$, $\beta_2 = 0.25$, $\beta_3 = 0.25$ and $\beta_4 = 0.35$; meaning that, we have assigned more weight to the scar channel than the others, and the background channel has the least weight.

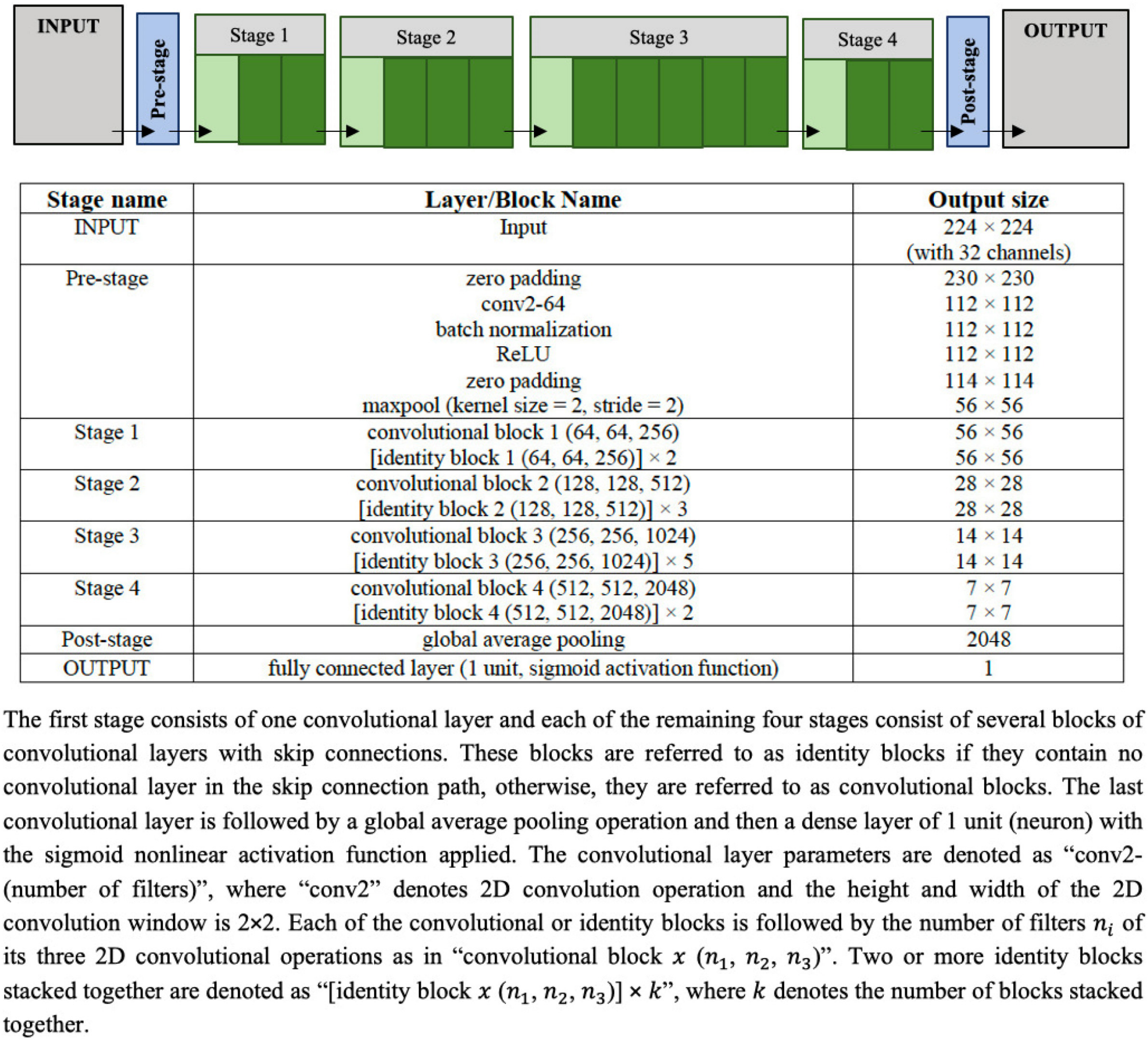
Classification approach

ResNet50 classification model

The architecture of ResNet50 given in **Figure 4** is used to train the DL classification model. Given cine SAX input frames, the ResNet model predicts whether or not the corresponding post-contrast LGE image would contain a scar as a result of MI. The main characteristic of this architecture of ResNet50 is that the number of channels of the input is 32 (i.e., 32 cine SAX images) – as against the 3 channels for a colored RGB image in a standard ResNet50 model. The output of the model is a binary prediction of whether the cine SAX images in a case without MI (class 0) or with MI (class 1) (without or with scar, respectively). The total number of parameters of the model is 23,680,705 out of which 23,627,585 parameters are trainable.

For the model training, the optimization method used was the Root Mean Squared Propagation (RMSProp) optimizer, with an initial learning rate of 0.0001. All other model training settings are the same as those of the segmentation model given in the preceding sub-section.

Of the 506 pairs that constitute the training set, the number of pairs that represent scar cases and cases with no scar are 336 and 170, respectively. In order to address this data imbalance, we used a weighted loss function. Let $\{(x_1, y_1), (x_2, y_2), (x_n, y_n), (x_N, y_N)\}$ denote a training set of N samples where x is the cine SAX input images and



Stage name	Layer/Block Name	Output size
INPUT	Input	224 × 224 (with 32 channels)
Pre-stage	zero padding conv2-64 batch normalization ReLU zero padding maxpool (kernel size = 2, stride = 2)	230 × 230 112 × 112 112 × 112 112 × 112 114 × 114 56 × 56
Stage 1	convolutional block 1 (64, 64, 256) [identity block 1 (64, 64, 256)] × 2	56 × 56 56 × 56
Stage 2	convolutional block 2 (128, 128, 512) [identity block 2 (128, 128, 512)] × 3	28 × 28 28 × 28
Stage 3	convolutional block 3 (256, 256, 1024) [identity block 3 (256, 256, 1024)] × 5	14 × 14 14 × 14
Stage 4	convolutional block 4 (512, 512, 2048) [identity block 4 (512, 512, 2048)] × 2	7 × 7 7 × 7
Post-stage	global average pooling	2048
OUTPUT	fully connected layer (1 unit, sigmoid activation function)	1

The first stage consists of one convolutional layer and each of the remaining four stages consist of several blocks of convolutional layers with skip connections. These blocks are referred to as identity blocks if they contain no convolutional layer in the skip connection path, otherwise, they are referred to as convolutional blocks. The last convolutional layer is followed by a global average pooling operation and then a dense layer of 1 unit (neuron) with the sigmoid nonlinear activation function applied. The convolutional layer parameters are denoted as "conv2-(number of filters)", where "conv2" denotes 2D convolution operation and the height and width of the 2D convolution window is 2×2. Each of the convolutional or identity blocks is followed by the number of filters n_i of its three 2D convolutional operations as in "convolutional block x (n_1, n_2, n_3)". Two or more identity blocks stacked together are denoted as "[identity block x (n_1, n_2, n_3)] × k", where k denotes the number of blocks stacked together.

FIGURE 3

Configuration of the UNet architecture to predict the segmentation mask that marks of the regions of the scar, the myocardium and left ventricle (LV) cavity. The input image is a set of 32 frames of cine SAX image frames. The ground truth (output image) consists of an image with four channels, namely, the background (label = 0), the myocardium (label = 1), the LV cavity (label = 2), and the scar (label = 3).

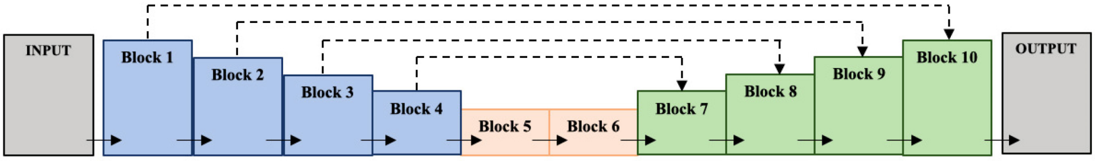
$y \in \{0, 1\}^C$ denote a binary one-hot encoded label with C 2 in our case, then the weighted loss function is defined as:

$$E_w(\theta) = -\frac{1}{N}\left[\lambda_0 \sum_{n=1}^{N} \mathbb{T}_0(x_n)\, y_n \log(\hat{y}_n(x_n,\ \theta)) + \lambda_1 \sum_{n=1}^{N} \mathbb{T}_1(x_n)\, y_n \log(\hat{y}_n(x_n,\ \theta))\right] \quad (4)$$

where θ denotes the trainable parameters of the model; $\hat{y}_n(x_n,\ \theta)$ is the posterior probability obtained after the application of sigmoid activation function on the output layer of the model; $\mathbb{T}_0(x_n)$ and $\mathbb{T}_1(x_n)$ are functions that indicate whether image x_i belongs to class 0 (cases with no scar) or class 1 (cases with scar), respectively; and λ_0 and λ_1 are weights that penalize the loss function for false negative errors and false positive errors, respectively. The weights, λ_0 and λ_1, can be computed using the following equation:

$$\lambda_i = \frac{1}{k_i} \cdot \frac{N}{c} \quad (5)$$

where k_i is the number of samples belonging to class i. In our case, $N = 506$; class 0 and class 1 are subgroups indicating the collection of samples with no scar and with scar, respectively; then, $\lambda_0 = (1/336) \times (506/2) = 0.753$ and $\lambda_1 = (1/170) \times (506/2) = 1.488$. In other words, the images



Layer/Block/Operation Name	Layer/Block/Operation Name
INPUT Input Images (32×224×224, 1)	OUTPUT Output Image (224×224, 4)
Block 1 Conv3 (16×112×112, 64) [kernel size = 3×3×3, stride = 2] Leaky ReLU	Block 10 Skip connection (concatenates the first channels of the outputs of Block 1 and Block 9) Conv3-T (2×224×224, 4) [kernel size = 3×3×3, stride = 2] softmax
Block 2 Conv3 (8×56×56, 128) [kernel size = 3×3×3, stride = 2] Batch normalization Leaky ReLU	Block 9 Skip connection (concatenates the first channels of the outputs of Block 2 and Block 8) Conv3-T (2×112×112, 128) [kernel size = 3×3×3, stride = 2] Batch normalization ReLU
Block 3 Conv3 (4×28×28, 256) [kernel size = 3×3×3, stride = 2] Batch normalization Leaky ReLU	Block 8 Skip connection (concatenates the first channels of the outputs of Block 3 and Block 7) Conv3-T (2×56×56, 256) [kernel size = 3×3×3, stride = 2] Batch normalization Dropout ReLU
Block 4 Conv3 (2×14×14, 512) [kernel size = 3×3×3, stride = 2] Batch normalization Leaky ReLU	Block 7 Skip connection (concatenates the first channels of the outputs of Block 4 and Block 6) Conv3-T (2×28×28, 512) [kernel size = 3×3×3, stride = 2] Batch normalization Dropout ReLU
Block 5 Conv3 (1×7×7, 1024) [kernel size = 3×3×3, stride = 2] Batch normalization Leaky ReLU	Block 6 Conv3-T (2×14×14, 1024) [kernel size = 3×3×3, stride = 2] Batch normalization Dropout ReLU

The convolutional layer parameters are denoted as "conv3 (dimension of output, number of filters)" where "conv3" denotes 3D convolution operation. "conv3-T (dimension of output, number of filters)" denotes 3D transpose convolution layer. The conv3 and conv2-T operations have 'same' padding (i.e. output and input of the operation have the same height and width). Each arrow points from the output of a block to the input of another block. The skip connection involves the concatenation of the first channels of the outputs of a block to the output of another block. In the case of the concatenation in Block 7 involving Block 4 (output dimension: 2×14×14×512) and Block 6 (output dimension: 2×14×14×1024) for example, the first channel of Block 4 (dimension: 1×14×14×512) is concatenated with the first channel of Block 4 (dimension: 1×14×14×1024) to form an array of dimension 1×14×14×1536.

FIGURE 4
Configuration of the ResNet50 architecture to predict the binary outcome of presence or absence of myocardial scar using the set of 32 cine short axis image frames as input to the model.

representing scar cases (class 1) are weighted as being more valuable than those representing no scar cases (class 0).

Feature extraction

In order to use ML methods for identifying MI using pre-contrast cine SAX images, we explore the data-driven approach by capturing three sets of data from the cine SAX images, namely, the optical flow data, the rate of change of myocardial area, and the radiomics data.

Optical flow data

The goal of optical flow is to compute an approximation of 2D motion field from spatiotemporal data of image intensities (25). Using Lucas–Kanade method (see Supplementary Material) for estimating optical flow velocities, we have chosen the window size (spatial neighborhood Ω) of 8 × 8 and selected a Gaussian filter w of kernel size 5 × 5 with a SD of 3 along each of x and y directions ($\sigma_x = \sigma_y = 3.0$). The magnitude of the optical flow velocities $v = |\mathbf{v}|$ can be computed as follows:

$$v = \sqrt{v_x + v_y} \tag{6}$$

v_x and v_y represent x and y component of v. The magnitude of the displacement of the optical flow field r (pixel-wise displacement) can therefore be computed as follows:

$$r = v\,\Delta t \tag{7}$$

where Δt is the time difference between acquisition of the two successive images. We further reduced the dimension of the displacement matrix r using principal component analysis (PCA) approach and vectorized (reshaped) the resulting matrix into a row vector.

For illustration, Figure 5A shows flow maps which helps to visualize the pixel-wise displacement between two cine SAX image frames images 0 and 1 in (a) and images 0 and 3 in (b). We refer to the interval between image frames as the "skip" interval, k; in (a), the $k = 1$ and in (b), the $k = 3$. The images on the right show the super-imposition of the flow map images on

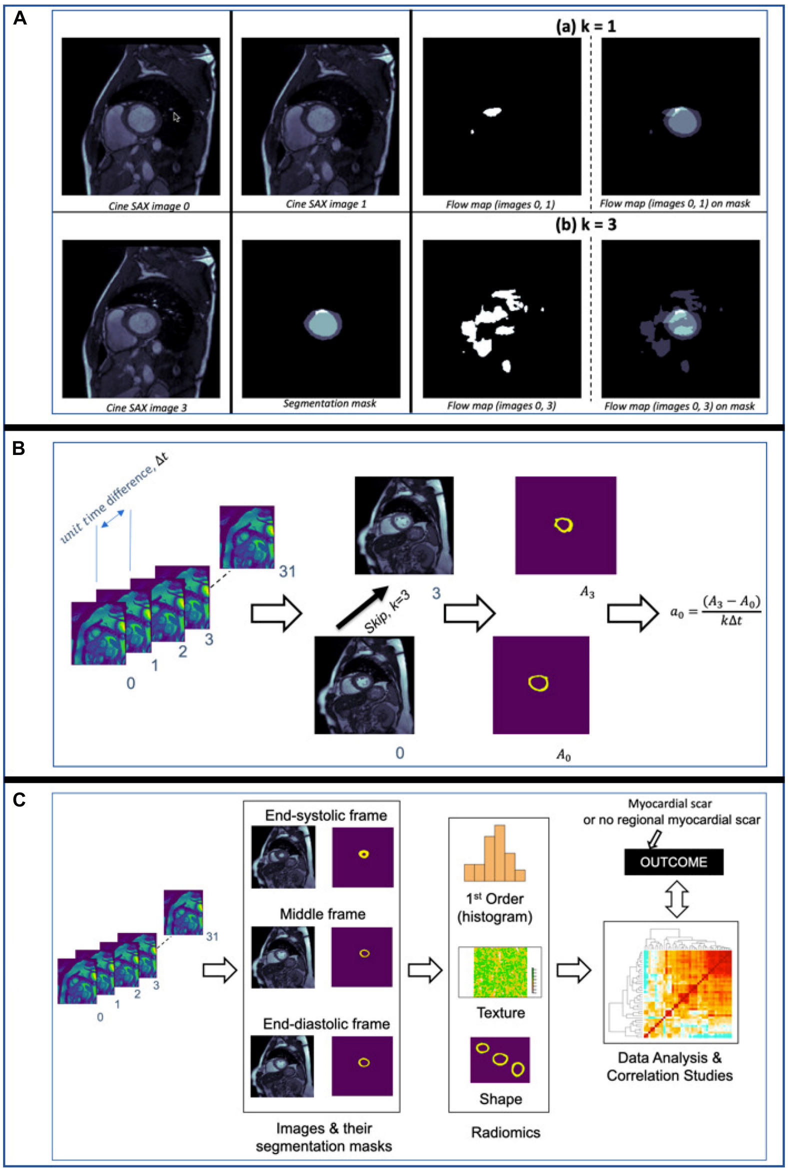


FIGURE 5
(A) Flow maps showing the pixels that have been displaced between the time the two cine SAX image frames were acquired with the skip intervals $k = 1$ in (a) and $k = 3$ in (b). On the left in (a), flow map (images 0, 1) represents the pixel-wise displacement between image 0 and image 1, and the corresponding the pixel-wise displacement between image 0 and image 3 is shown at the bottom. The images on the left show the super-imposition of the flow map images on the segmentation masks. **(B)** An illustration of the computation of rate of change of myocardial cross-section area for a given slice location. The rate of change of myocardial area between frames captures the pixel-wise area change information of slices through the cardiac cycle. **(C)** An illustration of the computation of radiomics features of three frames namely, the end-diastolic (ED) frame, end-systolic (ES) frame and the "middle" frame in between the ED and ES frames. The shape, first-order, and texture radiomics features are extracted and then a statistical test (Pearson correlation analysis) to assess the significance of these features in relation to the binary outcome (i.e., with or without MI).

the segmentation masks. Equation 7 assumes $k = 1$ (i.e., the two image frames are next to each other, e.g., the 2nd and 3rd frames, in a cine SAX set of 32 frames). The choice of $k = 1$ reduces the number of displacement matrices. In the case of 32 cine SAX frames for a given slice location, the choice $k = 3$ results in having 11 displacement matrices where the magnitude of the displacement is $r = v \times k\Delta t$ and Δt can be calculated by subtracting the TT as follows: $\Delta t = t_{i+1} - t_i$ where t_i is the TT associated with image i.

Rate of myocardial area change

The rate of change of myocardial cross-section area between successive frames of a given slice location, a_i, captures the pixel-wise area change information of slices through the cardiac cycle and can be expressed as follows:

$$a_i = \frac{\Delta A}{k\Delta t} \tag{8}$$

where $\Delta A = A_{i+k} - A_i$; A_i and A_{i+k} are the areas of myocardium for the i-th and $(i+k)$-th image frames at a given slice location, respectively, and k is the skip interval. Figure 5B illustrates the computation of a_i for a given slice location. Thus, for the up to 32 cine SAX frames for a given slice location and with $k = 3$, we compute 11 values of a_i (i.e., a_i for $i = 0, 2, 10$).

Data from radiomics

Rather than computing the radiomics features of each of the 32 cine SAX frames, we extracted radiomics from three frames: end-diastole, end-systole and the "middle" frame, which is precisely in between the end-diastolic and end-systolic frames. We extracted 306 shape, first-order, and texture radiomics features for the three frames of interest using the Pyradiomics open-source package. It should be noted that only the myocardium is segmented (i.e., an image segmentation mask with the background pixels labeled 0, and the myocardium labeled l). More details on radiomics features can be found in Freeman et al. (42) and Chu et al. (43).

The pixel spacing varies from 1.41 to 2.34 mm and to correct for differences in pixel size, each of the 2D image slices were resampled to 1.9 mm × 1.9 mm spacing through a one-dimensional (1D) area interpolation. Similar to Di Noto et al. (19), owing to strongly anisotropic CMR acquisition (i.e., out-of-plane information is intrinsically poorer), we resampled on the XY plane to preserve in-plane information. Furthermore, to account for sensitivity of radiomics features to intensity variation associated with the image acquisition process, intensity normalization of the images is carried out prior to the extraction of radiomics features. For the intensity normalization, the 1–99% intensity normalization (i.e., 1–99% percentile of intensities) with 256 intensity levels of each image has been used. In our pre-processing step, we have not performed bias correction although it may improve the inhomogeneity of images (44).

From the 306 set of radiomics features, a subset of features that are highly correlated (i.e., redundant features) are removed. In particular, features with Pearson correlation coefficient higher than 0.9 are removed while retained only one of those correlated features, resulting in 144 features (48 features for each of the three frames). We have chosen this value ($r > 0.9$), similar to Rauseo et al. (45), to ensure only highly correlated features are removed. We then carried out statistical test to assess the significance of the radiomics features in relation to the outcome – in this case, a binary outcome of whether the cine SAX images results predict MI or not. A $p < 0.001$ was considered to be statistically significant, leading to the selection of 38 radiomics features (15 systolic frame, 9 diastolic frame, and 14 middle frame features). Figure 5C illustrates the computation of radiomics features as we have described here.

Support vector machines and decision tree machine learning methods

Given that we now have information on the optical flow data, data on rate of change of myocardial cross-section area at any given slice location, and the radiomics data, we explored two supervised learning methods to qualitatively predict the presence/absence of scar, namely, the SVM and the DT methods.

Firstly, we carried out the z-score standardization. This standardization method transforms the feature space to have zero mean and unit variance, and has been shown to improve speed of convergence of SVM algorithms for classification problems and SVM model performance (46). Normalization is important for SVM method (47, 48) [in fact, SVM method may not be appropriate for some problems without normalization (49)]. While feature space transformation using standardization or normalization plays an important role in Euclidean distance minimization-based algorithms (e.g., SVM, neural networks, K-nearest neighbor, etc.), algorithms that are insensitive to feature scaling (variance scaling), such as DT, are not affected by the transformation (50).

Next, we split the dataset into random 80% train and 20% test subsets (representing, 577 and 145 sets, respectively). The training set was used for training SVM and DT models and the test set was used for unbiased evaluation of these models. Further details of the SVM with the radial basis function kernel (7, 8) and DT (9) methods used in this work are provided in the Supplementary Material.

Model evaluation methods

The DSC is a measure of similarity between the label and predicted segmentation masks and is often used to evaluate performance of ML segmentation models. Given two sets (two images in this case) A and B, DSC score can be expressed as follows:

$$DSC = \frac{2\left|A \cap B\right|}{(|A| + |B|)} \tag{9}$$

where $|A|$ and $|B|$ represent the cardinalities of set A and B (i.e., the number of elements in each set), respectively. The DSC, which has a range of [0,1], is a useful summary measure of spatial overlap that can be applied to quantify the accuracy in image segmentation tasks. Computing the DSC of several images from a segmentation model and evaluating the mean DSC (or other statistical validation metric) allows the comparison of the model with other models.

The performance of a classification models can be evaluated using the following metrics: confusion matrix, F1 score and accuracy score. The confusion matrix is a table that describes the performance of a model on a set of data for which the true labels are known by summarizing the count values for each class. For binary classification, the confusion matrix counts the number of true negative (*TN*), false negative (*FN*), true positive (*TP*) and false positive (*FP*). Precision (model's ability not to misclassify a negative sample as positive, i.e., a measure of result relevance), recall (model's ability to find all positive samples), the F1 score (the harmonic mean of the precision and recall) and the accuracy score (the fraction of the correct prediction out of the total number of samples) are other performance metrics useful in evaluating binary classification tasks.

Results

Segmentation approach

The evaluation of the predicted results of the UNet segmentation model was performed using the DSC score as the performance metric. For the testing set (n = 108), the mean DSC is 0.20 [±0.17 SD; 0.64 maximum; 0.14 median (50% percentiles)] for scar, the mean DSC is 0.51 (±0.21 SD; 0.86 maximum; 0.52 median) for epicardium (the myocardium), and the mean DSC is 0.75 (±0.20 SD; 0.94 maximum; 0.85 median) for endocardium (the LV cavity). These results are summarized in **Table 1**. The results from fivefold cross-validation are presented in **Table 2** and give the mean (±SD) DSC as 0.24 (±0.12), 0.48 (±0.23), and 0.77 (±0.18) for the scar, epicardium, and endocardium, respectively. In general, we observe that the UNet model is able to discover the regions of the endocardium and the epicardium with high degree of accuracy. The accuracy of the prediction of the extent and location of the scar is much lower. Some examples of the results of the UNet segmentation model are presented in **Figure 6**. Only the first cine SAX image in the set of 32 cine SAX images is shown in each example. In relation to the channel-weighted dice loss function, although the choices of β_i were empirical, our experiments as given in **Figure 7** shows that our choice of these values are reasonable. In **Figure 7**, we have presented the results of our simulation experiment for the first 30 epochs to show the accuracy (pixel-wise categorical accuracy) and loss (Equation 3). The arrows in (b) indicate our choice and both the loss and accuracy are satisfactory compared to other possible choices. For this experiment, we have only considered the cases of (A) $\beta_1 = 0.1$, (B) $\beta_1 = 0.15$, and (C) $\beta_{\ 1} = 0.2$.

TABLE 1 The mean, maximum and median Dice scores of the UNet segmentation model.

	Mean (SD)	Maximum	Median
Scar	0.20 (±0.17)	0.64	0.14
Epicardium	0.51 (±0.21)	0.86	0.52
Endocardium	0.75 (±0.20)	0.94	0.85

Classification approach

The performance metrics (confusion matrix, precision, recall, accuracy, and F1 scores) of the trained ResNet50 model on the evaluation dataset are given in **Figure 8A**. **Figure 8B** provides some examples of the predictions of the classification models. The precision score (0.19) is particularly poor for ResNet50 model (i.e., the number of TP of the confusion matrix is relatively small) making the model unsatisfactory in determining the presence of absence of MI.

The results of the SVM with radial basis function kernel are shown in (a) of **Table 3** with different combinations of data, namely, (i) the optical flow "plus" rate of myocardial area change data ($\frac{\triangle A}{k\triangle t}$); (ii) the optical flow "plus" $\frac{\triangle A}{k\triangle t}$ "plus" radiomics data; and (iii) $\frac{\triangle A}{k\triangle t}$ "plus" radiomics data. We observe improvement in prediction accuracy of the SVM model from the confusion matrices moving from left to right; that is, the model with the rate of myocardial area change data "plus" radiomics data has the highest accuracy and F1 scores of 0.68 and 0.69, respectively. The results of the DT model are shown in (b) of **Table 3**. Similarly, DT model where the input consists of the three combination of data features has the best performance in terms of the precision score. **Figure 9** shows the receiver operating characteristic curves (ROC) for the best performing SVM and DT classifiers calculated from 10-fold cross-validation. The area under the curve (AUC) values (0.5 < AUC < 1), that

TABLE 2 Comparing the mean, maximum, and median Dice scores of five UNet segmentation models calculated from fivefold cross-validation.

	Mean (SD)	Maximum	Median
Model 1			
Scar	0.16 (±0.14)	0.44	0.12
Epicardium	0.48 (±0.23)	0.86	0.54
Endocardium	0.74 (±0.23)	0.95	0.83
Model 2			
Scar	0.17 (±0.13)	0.42	0.13
Epicardium	0.41 (±0.27)	0.90	0.46
Endocardium	0.67 (±0.29)	0.97	0.81
Model 3			
Scar	0.24 (±0.15)	0.53	0.27
Epicardium	0.46 (±0.24)	0.87	0.48
Endocardium	0.77 (±0.18)	0.97	0.83
Model 4			
Scar	0.20 (±0.17)	0.64	0.16
Epicardium	0.44 (±0.25)	0.88	0.48
Endocardium	0.72 (±0.22)	0.96	0.80
Model 5			
Scar	0.20 (±0.12)	0.42	0.14
Epicardium	0.41 (±0.25)	0.87	0.43
Endocardium	0.72 (±0.22)	0.96	0.79

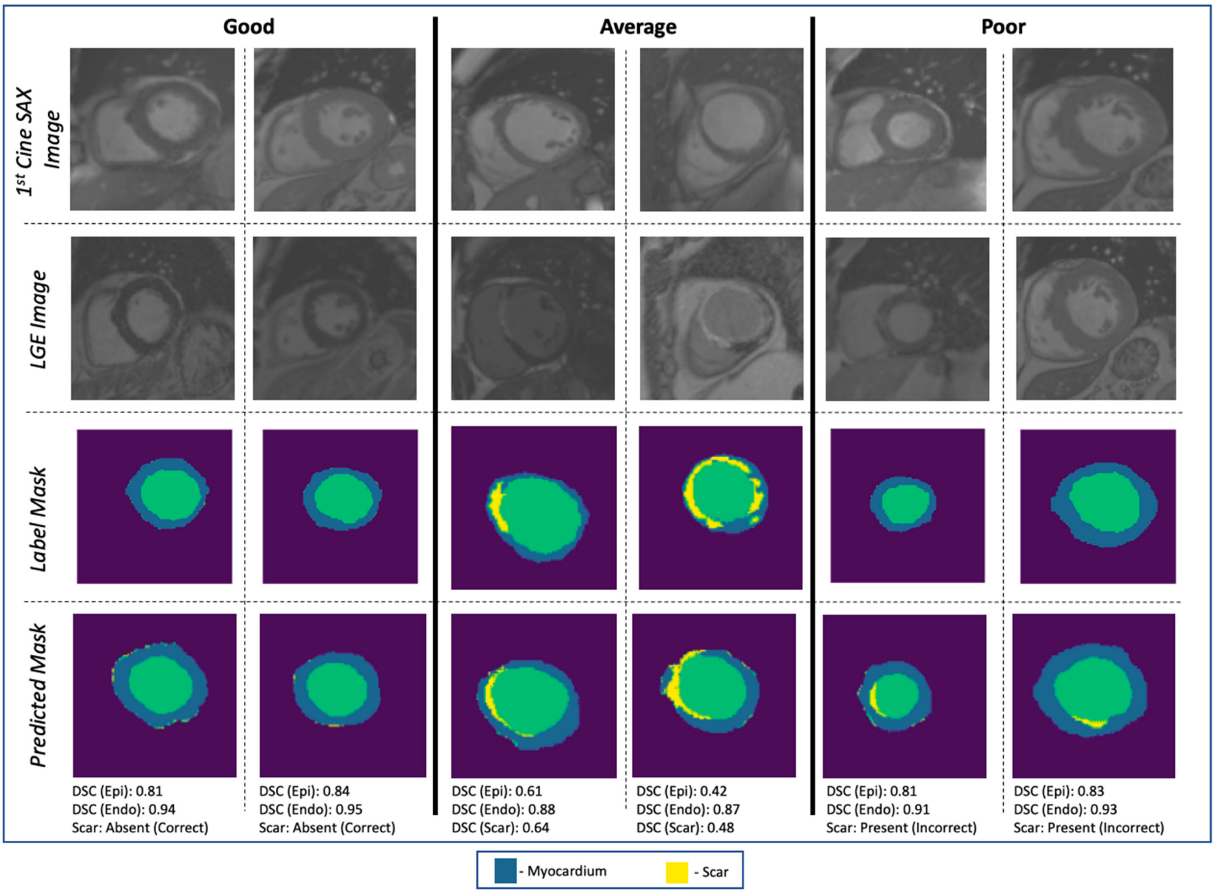


FIGURE 6
Examples of image segmentation results. The "Good" (correctly identifying the absence of myocardial scar due to MI), "Average" [correctly identifying the presence of the scar with $DSC \in (0.4,\ 0.65)$], and "Poor" (incorrect identification of the presence of the scar). Cases with $DSC\ <\ 0.4$ are not shown in the figure.

is, 0.58 ± 0.06 and 0.57 ± 0.06 for SVM and DT classifiers, respectively, show that the classifiers have some predictive power to distinguish between the positive class values from the negative class values.

The list of 38 radiomics features (15 systolic frame, 9 diastolic frame, and 14 middle frame features) that are considered statistically significant ($p\ <\ 0.001$) are given in Table 4. The definition of these feature can be found in (36). In order to estimate the importance of each radiomics feature to the SVM and DT models, permutation feature importance method (51) is used. This involves shuffling each of the features N number of times (N = 10 in our case) and estimate the importance of the feature by measuring the decrease in model predictive accuracy. Figure 10 shows the importance of each radiomics feature for the (optical flow "plus" $\frac{\Delta A}{k\Delta t}$ "plus" radiomics data) SVM and DT models computed using the test data set (i.e., using the training set data may indicate features that are important during model training only and these may not generalize).

Discussion

Summary of findings

In this study, we have presented a novel pipeline for processing routinely acquired CMR images that can be used as ground truth images for supervised and unsupervised ML methods in order to predict presence, location and extent of MI from contrast agent free cine SAX set.

For the UNet segmentation model whose input are the cine SAX frames and whose output is the segmentation mask that marks the regions of the epicardium, endocardium and the scar, the overall performance based on the average DSC score for the 108 test set is 0.75 (±0.20) for the endocardium, 0.51 (±0.21) for the epicardium and 0.20 (±0.17) for the scar; and 0.24 (±0.12), 0.48 (±0.23), and 0.77 (±0.18) for the scar, epicardium and endocardium, respectively, from fivefold cross-validation. At first glance, the prediction accuracies of the epicardium or endocardium appear low given that medical

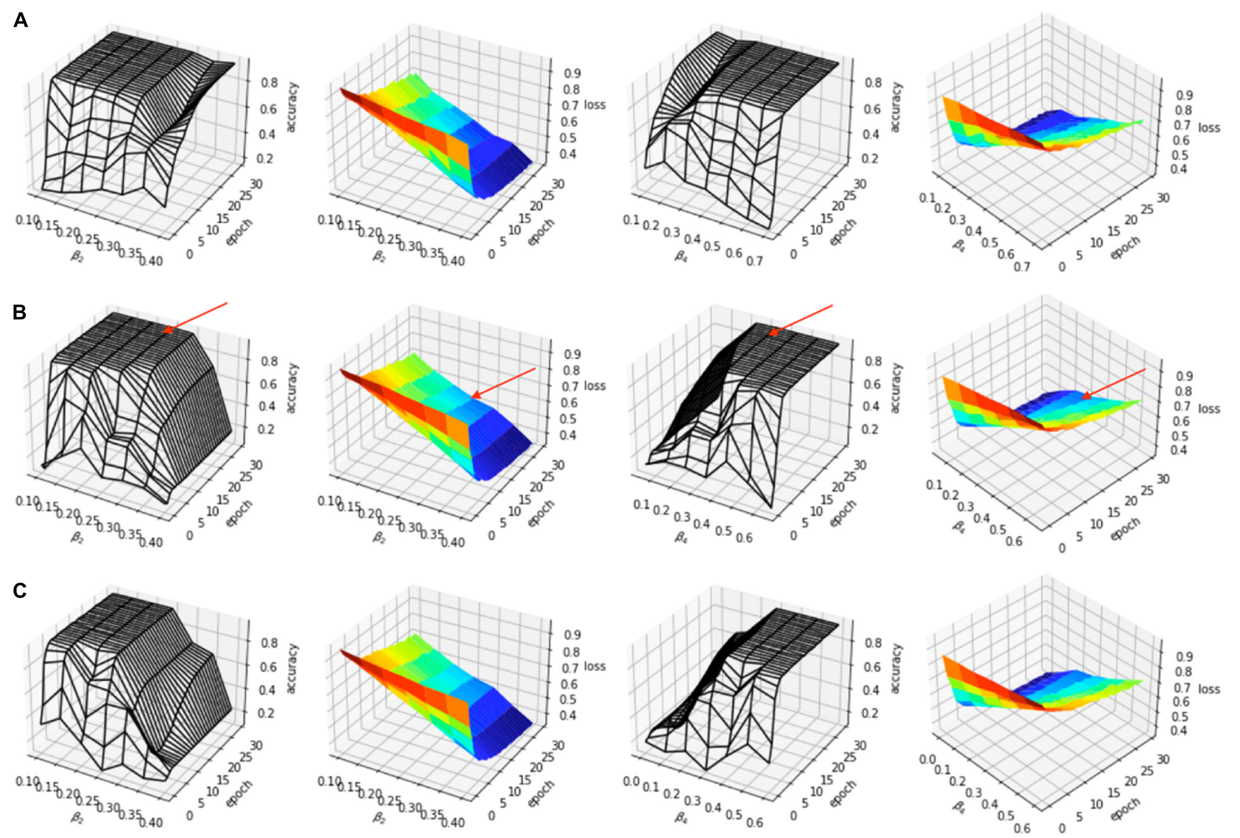
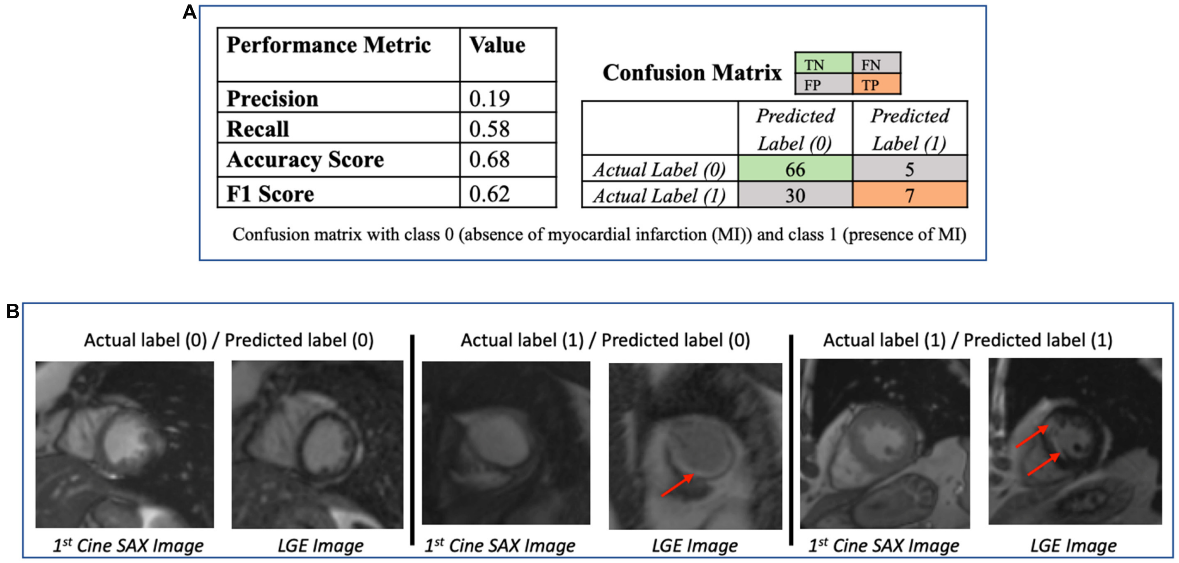


FIGURE 7
Simulation experiment for **(A)** $\beta_1 = 0.1$, **(B)** $\beta_1 = 0.15$, and **(C)** $\beta_1 = 0.2$. In each case, $\beta_2 = \beta_3$. The arrows in **(B)** indicate the choice ($\beta_1 = 0.15$, $\beta_2 = \beta_3 = 0.25$, $\beta_4 = 0.35$).

image segmentation is a widely studied subject (3) and the state-of-the-art average DSC could reach 0.94 for CMR (52) or 0.885 for cardiac CT images (53). We note that in those previous studies, the models (i.e., the one-input and one-output models) involved a single image and the prediction of the models is compared with the ground truth obtained after segmenting the input image. In our case (i.e., a 32-inputs and one-output model), the model input is a set of 32 image frames and the DSC score was obtained for only one of these images whose TT "matches" the TT of the LGE image of the same slice location; thus, comparing the predicted mask with only one mask (i.e., mask of the registered LGE image only out of the 32 possible masks) does not necessarily paint the overall picture of the level of the accuracy (and therefore, cannot be benchmarked with standard image segmentation models involving one-input, one-output models).

Also, in the LGE image registration for data pre-processing, we have used affine image transformation. Other non-rigid image registration methods, e.g., the free-form deformation (FFD) image registration methods (54, 55), may be more suitable for recovering motion and deformation since they can capture local motion of the myocardium into the registration process. Using FFD-based methods for spatio-temporal CMR image registration to correct spatial misalignment caused by patient motion and temporal misalignment caused by the motion of the heart may improve the accuracy of registration and, consequently, the manual segmentation for the ground truth and the prediction of the UNet model. Such methods may therefore be considered in a future work on this subject.

Moreover, the limited amount of dataset (the training, validation, and testing sets representing, 506, 108, and 108 cine SAX images – mask pairs, respectively) could explain this relatively low level of prediction accuracy for the scar (which is a more difficult problem even for human experts at times). Importantly, we note that we have used the channel-weighted DSC function as the loss function and have assigned 0.15, 0.25, 0.25, and 0.35 as weights of the background, epicardium, endocardium, and scar (i.e., assigning more importance to the scar channel than others). Future work on this project will involve experimenting with a much larger dataset as well as exploring different choices of weights assigned to the channels. It is worth mentioning that none of the other loss functions we have experimented so far [including the sparse categorical cross entropy loss and (unweighted) dice loss] were able to discover



A

Performance Metric	Value
Precision	0.19
Recall	0.58
Accuracy Score	0.68
F1 Score	0.62

Confusion Matrix

TN	FN
FP	TP

	Predicted Label (0)	*Predicted Label (1)*
Actual Label (0)	66	5
Actual Label (1)	30	7

Confusion matrix with class 0 (absence of myocardial infarction (MI)) and class 1 (presence of MI)

B
Actual label (0) / Predicted label (0)
Actual label (1) / Predicted label (0)
Actual label (1) / Predicted label (1)
1st Cine SAX Image
LGE Image
1st Cine SAX Image
LGE Image
1st Cine SAX Image
LGE Image

FIGURE 8
(A) Performance of the ResNet50 classification model (to determine presence/absence of myocardial scar) using the test dataset (N = 108). The low value of the precision score (0.19) makes the model unsatisfactory in determining the presence of absence of myocardial infarction.
(B) Some examples of the prediction of the ResNet50 classification model. The red arrows point to the location of the scar. The middle column shows the case that the classification models got wrong.

TABLE 3 Performance of machine learning data-driven approaches, with methods (a) support vector machines (SVM) using radial basis function (RBF) kernel and (b) decision tree, for different combinations of the optical flow, rate of change of myocardial area and radiomics data.The SVM model had the best performance in terms of accuracy and F1 scores when the input consists of the rate of change of myocardial area and the radiomics data. The DT model had the best in terms of the precision score when the input consists of the optical flow, the rate of change of myocardial area and the radiomics data.

	(a) SVM with RBF kernel			**(b) Decision Tree**		
Data Features / **Performance Metric**	Optical Flow + $\frac{\Delta A}{k\Delta t}$	Optical Flow + $\frac{\Delta A}{k\Delta t}$ + Radiomics	$\frac{\Delta A}{k\Delta t}$ + Radiomics	Optical Flow + $\frac{\Delta A}{k\Delta t}$	Optical Flow + $\frac{\Delta A}{k\Delta t}$ + Radiomics	$\frac{\Delta A}{k\Delta t}$ + Radiomics
Confusion Matrix (0) (1) (0) TN FN (1) FP TP	73 22 44 6	61 34 18 32	67 28 18 32	54 41 21 29	54 41 14 36	66 29 24 26
Precision	0.12	0.64	0.64	0.58	0.72	0.52
Recall	0.21	0.48	0.53	0.41	0.47	0.47
Accuracy Score	0.54	0.64	0.68	0.57	0.62	0.63
F1 Score	0.50	0.65	0.69	0.58	0.63	0.64

Confusion matrix with class 0 (absence of myocardial infarction (MI)) and class 1 (presence of MI)

the scar at all (despite some varying degree of successes in their discoveries of the epicardium and endocardium).

The performance metrics of ResNet classification model that predicts whether a cine SAX image frames constitute an MI case or non-MI case show that the model's performance is poor (see Figure 8). Of particular note here is the precision score ($\frac{7}{37}$, i.e., 0.19 – see also the confusion matrix). ResNet is a very successful DL architecture (3, 53) and the low performance of our ResNet model in this case (which could be for several reasons, e.g., not enough dataset or not information from the given dataset that will enable the model learn the underlying model parameters) emphasize difficulty and complexity of the problem we intend

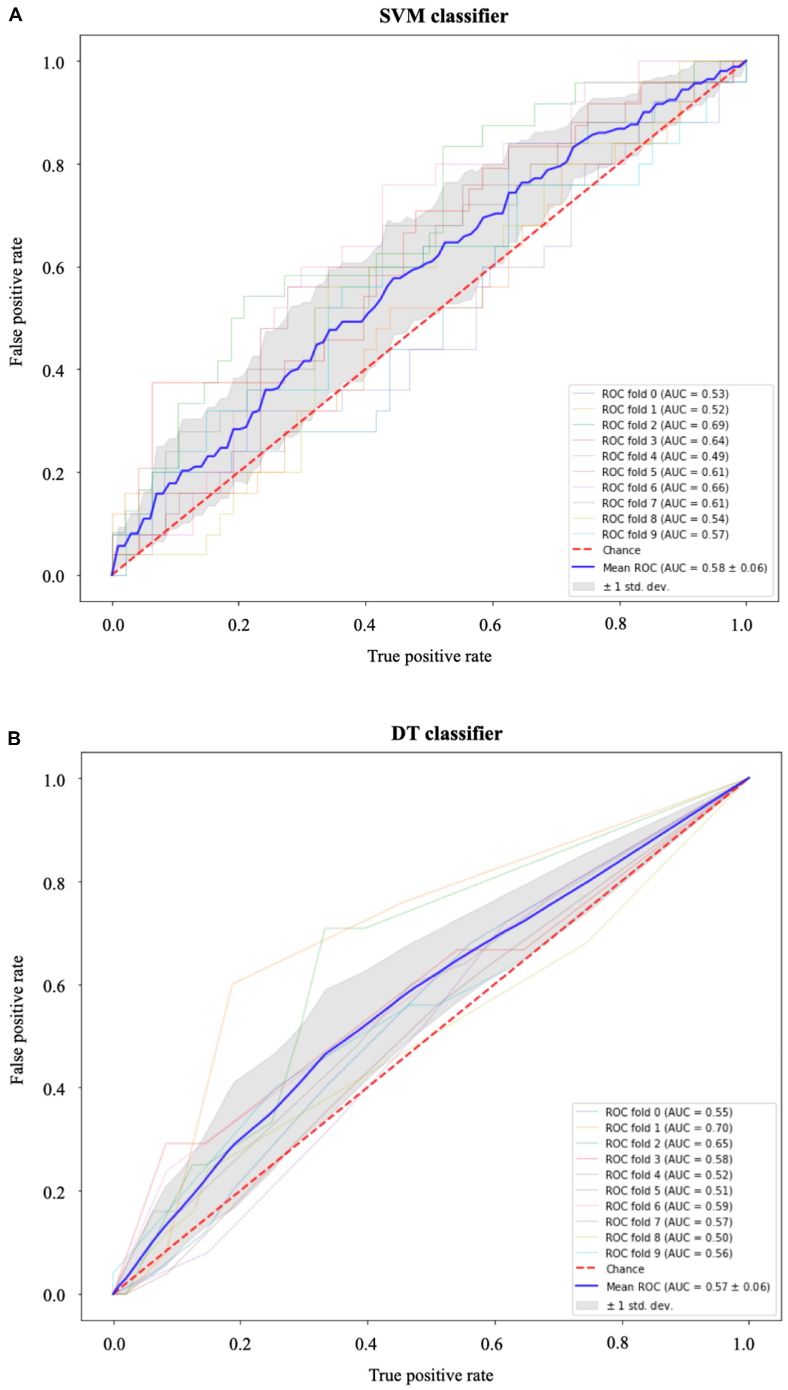


FIGURE 9
Receiver operating characteristic (ROC) curves for **(A)** the support vector machines (SVM) and **(B)** the decision tree (DT) classifiers. In both models, the area under the curve (AUC) values (0.5 < AUC < 1) show that the classifiers have some predictive power above "chance" to distinguish between the positive class values from the negative class values. The gray area indicates ±1 SD calculated from 10-fold cross-validation.

to solve. The lack of sufficient information could also be as a result of reduction of the resolution of the cine SAX images due to resizing (i.e., 224 × 224) although resizing can sometimes be necessary due to hardware limitations or to ensure all input images have common size. This motivated the data-driven ML approaches involving the use of the SVM and DT methods. In using these methods, we have experimented with the use of a combination of three sets of model input data that were captured

TABLE 4 Statistically significant radiomics features of the three cine short-axis image frames.

Systolic frame	Diastolic frame	Middle frame
2D shape-based features: ●Major axis length ●Maximum diameter ●Minor axis length ●Perimeter ●Sphericity	2D shape-based features: ●Major axis length ●Minor axis length ●Perimeter	2D shape-based features: Major axis length ●Minor axis length ●Perimeter ●Sphericity
First-order statistics features: ● 90th percentile ● Energy ● Maximum ● Mean ● Median ● Range ● Root mean squared ●Total energy	First-order statistics features: ●10th percentile ●Energy ●Total energy	First-order statistics features: ●10th percentile ●Energy ●Maximum ●Mean ●Median ●Root mean squared ●Total energy
Gray level run length matrix Texture-based features: ●Gray level non-uniformity ●Run length non-uniformity	Gray level run length matrix Texture-based features: ●Gray level non-uniformity ●Run entropy ●Run length non-uniformity normalized	Gray level run length matrix Texture-based features: ●Gray level non-uniformity ●Run entropy ●Run length non-uniformity normalized

from the cine SAX images, namely, the optical flow data, the rate of change of myocardial area, and the radiomics data. The SVM method had the best performance of accuracy score and F1 score of 0.68 and 0.69, respectively, when we included data from rate of change of myocardial area and the radiomics as input. The precision score was 0.64. The DT method had the best performance in terms of precision reaching 0.72 when the three sets of data are combined as input. In this case, the accuracy score and F1 score are 0.62 and 0.63, respectively. The best performing models of the data-driven ML methods outperforms the ResNet model on the precision score metric. Also, the SVM and DT models' AUC values ($0.5 < \text{AUC} < 1$) show that these classifiers do indeed have predictive power above "chance" to distinguish between the positive class values from the negative class values.

Related work

The pixel-wise tissue identification approach given in Xu et al. (29) is an interesting approach given that it does not require any preliminary segmentation of myocardial walls, captures the dense motion of the myocardium and integrates both local and global motion features for its prediction. The main problem with the approach however is the absence of the ground truth data preparation pipeline or any technique necessary to address the spatio-temporal misalignment between the pre-contrast and post-contrast image. Given the complex nature of this problem and the reported accuracy of 95.03% mean Dice score of the trained model for the identification and segmentation of the scar from cine SAX images, it is most likely that the model – trained on the dataset of 165 cine CMR patients (140 diagnosed with MI and 25 control cases) – suffers from overfitting issues. Moreover, even simpler DL-based image segmentation problems involving CMR images hardly achieved this level of accuracy to date [for example, in Bai et al. (52), endocardium and epicardium segmentation models achieved 0.88 (0.03) and 0.94 (0.04) mean (±SD) Dice scores, respectively; in Jacobs et al. (56), myocardial segmentation model achieved 0.86 (±0.06) Dice score on gadolinium-enhanced CMR images; and in Zhuang et al. (57), myocardial segmentation of the mid-ventricular slice achieved 0.86 (±0.07) inter-observer Dice score on LGE CMR images]. It should be noted that in Zhang et al. (58), researchers have used the framework given in Xu et al. (29) (i.e., with the following main components: LV localization; the motion feature extraction layers which use LSTM and optical flow techniques; and the fully connected layers) for a training dataset that consists of only chronic MI (n = 169) and control (n = 69) patients. The Dice score of 86.1% (±5.7) was reported in the study but this was the result of a small test set [chronic MI (n = 43) and control (n = 18) patients] from a single vendor and single center (i.e., the same vendor and center as the training set). The approach and dataset presented in this article are not limited to chronic MI cases only.

The DSTGAN approach given in Xu et al. (13) uses a total of 495 cine SAX images and segmentation mask pairs (i.e., 25 cine frames for each segmentation mask) from 165 patients (140 acute MI patients and 25 non-MI patients), the approach demonstrates impressive performance of 96.98% pixel-wise classification on a 10-fold cross-validation test (i.e., the test set consists of approximately 49 cine SAX and segmentation mask pairs).

The cGAN model proposed in Zhang et al. (14) was trained and tested on a dataset of 2,695 and 345 triplets, respectively. The performance of the model (for n = 326 datasets) measured by the correlation with LGE images are [r 0.77–0.79; intraclass correlation coefficients (ICC) = 0.77–0.87; p 0.001] in detecting and quantifying hyperintensity myocardial lesions and (r 0.70–0.76; ICC = 0.82–0.85; p 0.001) in detecting and quantifying intermediate-intensity lesions. Moreover, as the authors rightly mentioned, this approach is relevant in the image acquisition stage, meaning that, clinicians will still have to visually scrutinize each of the synthesized images for the location and extent of scar in order to diagnose MI. Importantly, the T1 mapping images contain information about the characteristics of the tissues, and MI scar can somewhat be visible in these images – for example, refer to **Figure 4** in the original article. Thus, in our view, the problem solved in Zhang et al. (14) seems to take a relatively complex approach since the task can be reduced to a simpler image synthesis or segmentation task (i.e., synthesize a postcontrast

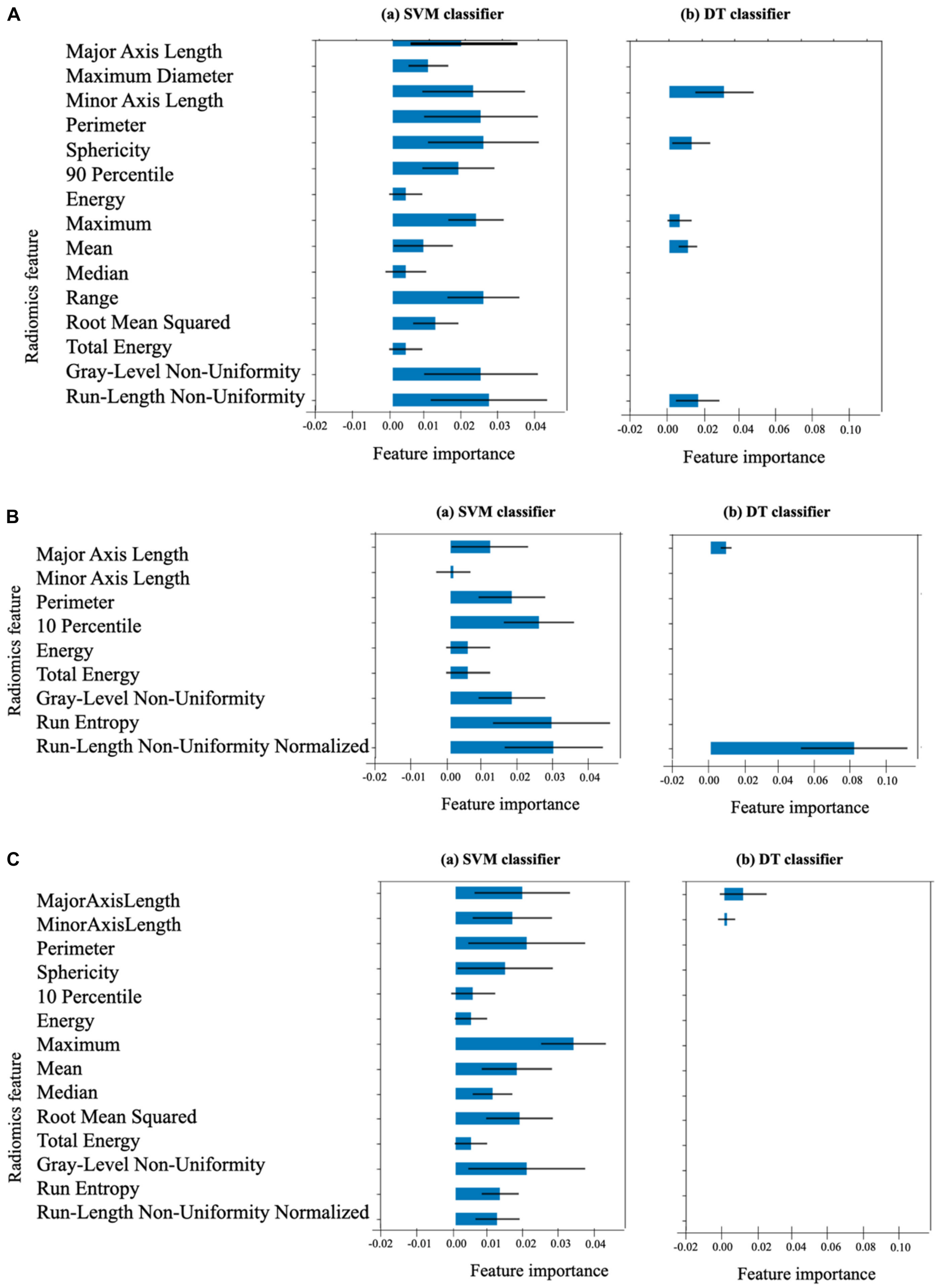


FIGURE 10
Importance of the 38 radiomics features associated with **(A)** the end-systolic frame (15 features), **(B)** the end-diastolic frame (9 features), and **(C)** the middle frame (14 features) for the SVM and DT models. The length of the blue bar indicates the importance of the feature to the generalization power of the model (the black line is the ±SD). The DT classifier considers only the 9 features indicated as the only important features for its own classification.

image or obtaining an image segmentation mask from a T1 mapping image). Moreover, as highlighted in Manisty et al. (59), T1 mapping images and LGE images are not imaging equivalent (i.e., interchangeable) myocardial disease processes, so one cannot be expected to replace the other. The approach proposed in this article differs to this model as its outcome is to diagnose with minimal amount of imaging data (cine SAX images only) as input and determine whether it is a case of infarction or not.

In the CNN-based model developed in (15) that identifies ischemic scar slices in CT angiography of the LV, with CT images as input and a training set of 200 patients of which 83 are with scar, the trained network achieved an accuracy of 88.3% on a 10-fold cross-validation metric.

The texture analysis approach proposed in (16) uses end-systolic cine SAX images from 120 MI patients [72 large transmural (>20%) patients and 48 small subacute or chronic (≤20% transmural) patients] and 60 control subjects and, using 5 textural radiomics features, reported a 10-fold cross-validation estimate of accuracy of 0.81 for patients with large myocardial scar versus control subjects, and a cross-validation estimate of accuracy of 0.75 for patients with small myocardial scar versus control subjects.

Similarly, the texture analysis approach proposed in Larroza et al. (17) uses LGE images and end-diastolic cine SAX images from 44 MI patients (22 acute MI patients and 22 chronic MI patients) and reported a fivefold cross-validation results. The SVM with polynomial kernel yielded the best classification performance with ROC providing AUC (mean ± SD) of 0.86 (±0.06) on LGE MRI using 72 textural radiomics features. For the cine CMR images, the SVM with polynomial kernel classifier's performance given by the AUC of ROC of 0.82 (±0.06) from 75 textural radiomics features.

In Larroza et al. (18) where the texture analysis was used to classify myocardial regions of chronic MI patients into remote, viable and non-viable segments, the approach uses end-diastolic cine SAX images from 50 chronic MI patients [randomly split into training (30 patients) and testing (20 patients) sets] and, using 5 textural radiomics features and a fivefold cross-validation, reported AUC under ROC of 0.849 with sensitivities of 85, 72, and 92% for remote, viable, and non-viable segments, respectively.

In Di Noto et al. (19), radiomics features where captured from LGE images in order to classify the images from 173 patients (111 with MI and 62 with myocarditis) into MI and myocarditis. The approach involved both 2D and 3D texture analysis to capture textural radiomics features; thus, the proposed method used shape and first-order features in addition to texture radiomics features. Five different ML algorithms were investigated and a stratified 10-fold cross-validation was performed. The SVM classifier achieved the best results (accuracy: 88%) for the 2D features and LDA showed the highest accuracy (85%) for 3D features. In comparison with subjective visual analyses by readers with different experience levels, the radiomics approach was superior to the less experienced reader but performed lower with the experienced reader.

Radiomics analysis was used in Avard et al. (20) to classify MI from healthy patients using a dataset of 50 MI and 20 healthy control cases, the average of univariate AUCs was 0.62 ± 0.08. For multivariate analysis, logistic regression (AUC = 0.93 ± 0.03) and SVM (AUC = 0.92 ± 0.05) yielded optimal performance. It is clear that this is a small and an imbalanced dataset (i.e., MI cases are 2.5 times the number of healthy cases) – and this can have significant influence on the predictive power of the models (classifiers are generally not robust to the change of training data size (60). In their report, the researchers have not mentioned how they have eliminated the impact of class imbalance or how its effect on their results.

Thus, while eliminating the need for contrast will save both time and cost of cardiovascular healthcare and improve the patient experience, accurate prediction of contrast information without contrast administration is a very challenging task. We have presented some promising approaches using a heterogeneous dataset for qualitative analysis using DL, SVM, and DT methods in order to predict post-contrast information accurately without requiring contrast administration. While our initial results are modest, our investigation shows that this area still poses several open challenges and opportunities for further research.

Limitations and future work

The main limitation of our study is that it included a relatively small number of patients from a single center. However, the study confirms the efficacy of ML methods and can improve our understanding of the diagnostic potentials of these emerging methods as well as data phenotypes that are yet to be standardized in CVD. Future work will focus on using novel DL architectures that combine both cine SAX image frames and the derived heterogenous variables to predict the extent and location of scar in the myocardium. To have a larger dataset for model training will involve automating the ground truth data preparation pipeline. Also, there is the need to consider risks associated with the applications when contrast is not administered when it should have been acquired and vice versa (i.e., the consequences of the false negatives and false positives).

Conclusion

Cardiac magnetic resonance imaging has potential to benefit from practical and inexpensive methods in the emerging field of ML for diagnosing MI without the use of a contrast agent. We have presented some promising approaches using a heterogeneous dataset for qualitative analysis using ML methods in an attempt to predict contrast information accurately without requiring contrast administration. Our study provided an original contribution and development in this area, presenting new parameters, such as, rate of myocardial area change, optical flow and radiomics parameters, that could be considered biomarkers of the mechanics of myocardial disease. However, further studies that would improve the proposed methods and identify other parameters are needed, just as it would be necessary to develop such models on a larger population in order to validate the results and make it possible to reach acceptable prediction level that could make it possible and safe to avoid contrast administration in clinical CMR scans.

Data availability statement

The datasets presented in this article are not readily available because restrictions apply to the availability of these raw data, which were used under license for the current study from Barts BioResource Institutional Review Board. Generated anonymised dataset can be made available from the authors upon reasonable request and with permission of Institutional Review Board of Barts BioResource, Barts Health NHS Trust, London, United Kingdom. Requests to access the datasets should be directed to SP, s.e.petersen@qmul.ac.uk.

Ethics statement

Data used for this research were Barts BioResource – a local biorepository of Barts Heart Centre (Barts Health NHS Trust, London, United Kingdom) that holds data from prospectively consented (written) patients for cardiovascular research (Ethics REC reference: 14/EE/0007). All images were de-identified prior to analysis. The patients/participants provided their written informed consent to participate in this study.

Author contributions

SP, MA, AL, KL, AY, MB, and PB conceived to the idea. MA, AA-K, ER, AS, AK, and SP developed the contouring method. MA led on the machine learning methodology and the main mathematical and statistical analysis, and drafted the first version of the manuscript. MA, AS, AK, KL, AY, MK, and SP contributed to the analysis. MA and AL advised on data governance and computing infrastructure. SP provided overall supervision. All authors contributed to the content, the writing of the final version or provided critical feedback.

Funding

This work forms part of the research areas contributed to the translational research portfolio of the Biomedical Research Centre at Barts which was supported and funded by the National Institute for Health Research. MA and SP acknowledged support from the CAP-AI program (led by Capital Enterprise in partnership with Barts Health NHS Trust and Digital Catapult and funded by the European Regional Development Fund and Barts Charity) and Health Data Research UK [HDR UK—an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities; www.hdruk.ac.uk]. SP acknowledged support from the SmartHeart EPSRC program grant (www.nihr.ac.uk; EP/P001009/1). SP had received funding from the European Union's Horizon 2020 Research and Innovation Program under grant agreement No. 825903 (euCanSHare project). SP and ER acknowledged support by the London Medical Imaging and Artificial Intelligence Centre for Value Based Healthcare (AI4VBH), which was funded from the Data to Early Diagnosis and Precision Medicine strand of the government's Industrial Strategy Challenge Fund, managed, and delivered by Innovate UK on behalf of UK Research and Innovation (UKRI). NA was supported by a Wellcome Trust Research Training Fellowship (wellcome.ac.uk; 203553/Z/Z). NA recognises the National Institute for Health Research (NIHR) Integrated Academic Training programme which supports his Academic Clinical Lectureship post.

Conflict of interest

AS, AA-K, and PB were employed by Circle Cardiovascular Imaging Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

The views expressed are those of the authors and not necessarily those of the AI4VBH Consortium members, the NHS, Innovate UK, or UKRI.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcvm.2022.894503/full#supplementary-material

References

1. WHF. *Cardiovascular Health for Everyone*. (2022). Available online at: https://world-heart-federation.org/ (accessed January 12, 2022).

2. BHF. *Heart Statistics*. (2022). Available online at: https://www.bhf.org.uk/what-we-do/our-research/heart-statistics (accessed January 12, 2022).

3. Abdulkareem M, Petersen SE. The promise of AI in detection, diagnosis and epidemiology for combating COVID-19: beyond the hype. *Front Artif Intell.* (2021) 4:652669. doi: 10.3389/frai.2021.652669

4. Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. *Annu Rev Biomed Eng.* (2017) 19:221–48. doi: 10.1146/annurev-bioeng-071516-044442

5. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham (2015). p. 234–41.

6. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV (2016). p. 770–8.

7. Smola AJ, Schölkopf B. A tutorial on support vector regression. *Stat Comput.* (2004) 14:199–222. doi: 10.1023/B:STCO.0000035301.49549.88

8. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J. LIBLINEAR: a library for large linear classification. *J Mach Learn Res.* (2008) 9:1871–4.

9. Hastie T, Tibshirani R, Friedman JH, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer (2009).

10. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Informatics Assoc.* (2014) 21:221–30. doi: 10.1136/amiajnl-2013-001935

11. Johnson KW, Shameer K, Glicksberg BS, Readhead B, Sengupta PP, Björkegren JLM, et al. Enabling precision cardiology through multiscale biology and systems medicine. *JACC Basic to Transl Sci.* (2017) 2:311–27. doi: 10.1016/j.jacbts.2016.11.010

12. Abdulkareem M, Aung N, Petersen SE. Biobanks and artificial intelligence. In: De Cecco CN, van Assen M, Leiner T editors, *Artificial Intelligence in Cardiothoracic Imaging*. Cham: Springer (2022). p. 81–93. doi: 10.1007/978-3-030-92087-6_8

13. Xu C, Howey J, Ohorodnyk P, Roth M, Zhang H, Li S. Segmentation and quantification of infarction without contrast agents via spatiotemporal generative adversarial learning. *Med Image Anal.* (2020) 59:101568. doi: 10.1016/j.media.2019.101568

14. Zhang Q, Burrage MK, Lukaschuk E, Shanmuganathan M, Popescu IA, Nikolaidou C, et al. Toward replacing late gadolinium enhancement with artificial intelligence virtual native enhancement for gadolinium-free cardiovascular magnetic resonance tissue characterization in hypertrophic cardiomyopathy. *Circulation.* (2021) 144:589–99. doi: 10.1161/CIRCULATIONAHA.121.054432

15. O'Brien H, Whitaker J, Sidhu BS, Gould J, Kurzendorfer T, O'Neill MD, et al. Automated left ventricle ischemic scar detection in CT using deep neural networks. *Front Cardiovasc Med.* (2021) 8:655252. doi: 10.3389/fcvm.2021.655252

16. Baessler B, Mannil M, Oebel S, Maintz D, Alkadhi H, Manka R. Subacute and chronic left ventricular myocardial scar: accuracy of texture analysis on nonenhanced cine MR images. *Radiology.* (2018) 286:103–12. doi: 10.1148/radiol.2017170213

17. Larroza A, Materka A, López-Lereu MP, Monmeneu JV, Bodi V, Moratal D. 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.* (2017) 92:78–83. doi: 10.1016/j.ejrad.2017.04.024

18. Larroza A, López-Lereu MP, Monmeneu JV, Gavara J, Chorro FJ, Bodi V, et al. Texture analysis of cardiac cine magnetic resonance imaging to detect nonviable segments in patients with chronic myocardial infarction. *Med Phys.* (2018) 45:1471–80. doi: 10.1002/mp.12783

19. Di Noto T, von Spiczak J, Mannil M, Gantert E, Soda P, Manka R, et al. Radiomics for distinguishing myocardial infarction from myocarditis at late gadolinium enhancement at MRI: comparison with subjective visual analysis. *Radiol Cardiothorac Imaging.* (2019) 1:e180026. doi: 10.1148/ryct.2019180026

20. Avard E, Shiri I, Hajianfar G, Abdollahi H, Kalantari KR, Houshmand G, et al. Non-contrast cine cardiac magnetic resonance image radiomics features and machine learning algorithms for myocardial infarction detection. *Comput Biol Med.* (2022) 141:105145. doi: 10.1016/j.compbiomed.2021.105145

21. Wu Y, Tang Z, Li B, Firmin D, Yang G. Recent advances in fibrosis and scar segmentation from cardiac MRI: a state-of-the-art review and future perspectives. *Front Physiol.* (2021) 12:709230. doi: 10.3389/fphys.2021.709230

22. Ladha L, Deepa T. Feature selection methods and algorithms. *Int J Comput Sci Eng.* (2011) 3:1787–97.

23. Khalid S, Khalil T, Nasreen S. A survey of feature selection and feature extraction techniques in machine learning. In: *Proceedings of the 2014 Science and Information Conference*, London. (2014). p. 372–8. doi: 10.1109/SAI.2014.6918213

24. Liu H, Motoda H. *Feature Extraction, Construction and Selection: A Data Mining Perspective*. New York, NY: Springer Science & Business Media (1998). doi: 10.1007/978-1-4615-5725-8

25. Barron JL, Fleet DJ, Beauchemin SS. Performance of optical flow techniques. *Int J Comput Vis.* (1994) 12:43–77. doi: 10.1007/BF01420984

26. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology.* (2016) 278:563–77. doi: 10.1148/radiol.2015151169

27. Schofield R, Ganeshan B, Kozor R, Nasis A, Endozo R, Groves A, et al. CMR myocardial texture analysis tracks different etiologies of left ventricular hypertrophy. *J Cardiovasc Magn Reson.* (2016) 18:1–2. doi: 10.1186/1532-429X-18-S1-O82

28. Baeßler B, Mannil M, Maintz D, Alkadhi H, Manka R. Texture analysis and machine learning of non-contrast T1-weighted MR images in patients with hypertrophic cardiomyopathy—preliminary results. *Eur J Radiol.* (2018) 102:61–7. doi: 10.1016/j.ejrad.2018.03.013

29. Xu C, Xu L, Gao Z, Zhao S, Zhang H, Zhang Y, et al. Direct delineation of myocardial infarction without contrast agents using a joint motion feature learning architecture. *Med Image Anal.* (2018) 50:82–94. doi: 10.1016/j.media.2018.09.001

30. Rajiah P, Desai MY, Kwon D, Flamm SD. MR imaging of myocardial infarction. *Radiographics.* (2013) 33:1383–412. doi: 10.1148/rg.335125722

31. Lücke C, Schindler K, Lehmkuhl L, Grothoff M, Eitel I, Schuler G, et al. Prevalence and functional impact of lipomatous metaplasia in scar tissue following myocardial infarction evaluated by MRI. *Eur Radiol.* (2010) 20:2074–83. doi: 10.1007/s00330-010-1791-x

32. Shriki JE, Surti KS, Farvid AF, Lee CC, Samadi S, Hirschbeinv J, et al. . *Can J Cardiol.* (2011) 27:664.e17–23. doi: 10.1016/j.cjca.2010.12.074

33. Thygesen K, Alpert JS, Jaffe AS, Chaitman BR, Bax JJ, Morrow DA, et al. Fourth universal definition of myocardial infarction (2018). *J Am Coll Cardiol.* (2018) 72:2231–64. doi: 10.1016/j.jacc.2018.08.1038

34. Chapman AR, Adamson PD, Mills NL. Assessment and classification of patients with myocardial injury and infarction in clinical practice. *Heart.* (2017) 103:10–8. doi: 10.1136/heartjnl-2016-309530

35. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res.* (2011) 12:2825–30.

36. Van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* (2017) 77:e104–7. doi: 10.1158/0008-5472.CAN-17-0339

37. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. (2015). Available online at: https://www.tensorflow.org/ (accessed March 11, 200).

38. Styner M, Brechbuhler C, Szckely G, Gerig G. Parametric estimate of intensity inhomogeneities applied to MRI. *IEEE Trans Med Imaging.* (2000) 19:153–65. doi: 10.1109/42.845174

39. Mattes D, Haynor DR, Vesselle H, Lewellyn TK, Eubank W. "Nonrigid multimodality image registration," in *Proceedings of the Medical Imaging 2001: Image Processing*, San Diego, CA (2001). p. 1609–20.

40. Rahunathan S, Stredney D, Schmalbrock P, Clymer BD. "Image registration using rigid registration and maximization of mutual information," in *Proceedings of the 13th Annual Medicine Meets Virtual Reality Conference*, Long Beach, CA (2005).

41. Wolterink JM, Leiner T, Viergever MA, Išgum I. Automatic segmentation and disease classification using cardiac cine MR images. In: Bernard O, Jodoin P-M, Zhuang X, Yang G, Young A, Sermesant M, et al. editors, *International Workshop on Statistical Atlases and Computational Models of the Heart*. Cham: Springer (2018). p. 101–10. doi: 10.1007/978-3-319-75541-0_11

42. Freeman J. The modelling of spatial relations. *Comput Graph Image Process.* (1975) 4:156–71. doi: 10.1016/S0146-664X(75)80007-4

43. Chu A, Sehgal CM, Greenleaf JF. Use of gray value distribution of run lengths for texture analysis. *Pattern Recognit Lett.* (1990) 11:415–9. doi: 10.1016/0167-8655(90)90112-F

44. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging.* (2010) 29:1310–20. doi: 10.1109/TMI.2010.2046908

45. Rauseo E, Izquierdo Morcillo C, Raisi-Estabragh Z, Gkontra P, Aung N, Lekadir K, et al. New imaging signatures of cardiac alterations in ischaemic heart disease and cerebrovascular disease using CMR radiomics. *Front Cardiovasc Med.* (2021) 8:716577. doi: 10.3389/fcvm.2021.716577

46. Luor D-C. A comparative assessment of data standardization on support vector machine for classification problems. *Intell Data Anal.* (2015) 19:529–46. doi: 10.3233/IDA-150730

47. Graf A, Borer S. Normalization in support vector machines. In: *Proceedings of the Joint Pattern Recognition Symposium*. Berlin (2001). p. 277–82. doi: 10.1007/3-540-45404-7_37

48. Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* (1995) 20:273–97. doi: 10.1007/BF00994018

49. Graf ABA, Smola AJ, Borer S. Classification in a normalized feature space using support vector machines. *IEEE Trans Neural Netw.* (2003) 14:597–605. doi: 10.1109/TNN.2003.811708

50. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. New York, NY: Springer (2013). doi: 10.1007/978-1-4614-7138-7

51. Breiman L. Random forests. *Mach Learn.* (2001) 45:5–32. doi: 10.1023/A:1010933404324

52. Bai W, Sinclair M, Tarroni G, Oktay O, Rajchl M, Vaillant G, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *J Cardiovasc Magn Reson.* (2018) 20:1–12. doi: 10.1186/s12968-018-0471-x

53. Abdulkareem M, Brahier MS, Zou F, Taylor A, Thomaides A, Bergquist PJ, et al. Generalizable framework for atrial volume estimation for cardiac CT images using deep learning with quality control assessment. *Front Cardiovasc Med.* (2022) 9:822269. doi: 10.3389/fcvm.2022.822269

54. Rueckert D, Sonoda LI, Hayes C, Hill DLG, Leach MO, Hawkes DJ. Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Trans Med Imaging.* (1999) 18:712–21. doi: 10.1109/42.796284

55. Rueckert D, Sonoda LI, Denton ERE, Rankin S, Hayes C, Leach MO, et al. Comparison and evaluation of rigid and nonrigid registration of breast MR images. In: *Proceedings of the Medical Imaging 1999: Image Processing*. San Diego, CA (1999). p. 78–88. doi: 10.1117/12.348637

56. Jacobs M, Benovoy M, Chang L-C, Corcoran D, Berry C, Arai AE, et al. Automated segmental analysis of fully quantitative myocardial blood flow maps by first-pass perfusion cardiovascular magnetic resonance. *IEEE Access.* (2021) 9:52796–811. doi: 10.1109/ACCESS.2021.3070320

57. Zhuang X, Xu J, Luo X, Chen C, Ouyang C, Rueckert D, et al. Cardiac segmentation on late gadolinium enhancement MRI: a benchmark study from multi-sequence cardiac MR segmentation challenge. *arXiv*[Preprint]. (2020). arXiv:200612434

58. Zhang N, Yang G, Gao Z, Xu C, Zhang Y, Shi R, et al. Deep learning for diagnosis of chronic myocardial infarction on nonenhanced cardiac cine MRI. *Radiology.* (2019) 291:606–17. doi: 10.1148/radiol.2019182304

59. Manisty CH, Jordan JH, Hundley WG. Automated noncontrast myocardial tissue characterization for hypertrophic cardiomyopathy: holy grail or false prophet? *Circulation.* (2021) 144:600–3.

60. Zheng W, Jin M. The effects of class imbalance and training data size on classifier learning: an empirical study. *SN Comput Sci.* (2020) 1:1–13. doi: 10.1007/s42979-020-0074-0

frontiers | Frontiers in Cardiovascular Medicine

TYPE Original Research
PUBLISHED 28 July 2022
DOI 10.3389/fcvm.2022.919751

OPEN ACCESS

EDITED BY
Alistair A. Young,
King's College London,
United Kingdom

REVIEWED BY
Gustav Strijkers,
Amsterdam University Medical
Center, Netherlands
Steven Alexander Niederer,
King's College London,
United Kingdom
Bharath Ambale Venkatesh,
Johns Hopkins University,
United States

*CORRESPONDENCE
Elliot McVeigh
emcveigh@ucsd.edu

SPECIALTY SECTION
This article was submitted to
Cardiovascular Imaging,
a section of the journal
Frontiers in Cardiovascular Medicine

RECEIVED 13 April 2022
ACCEPTED 27 May 2022
PUBLISHED 28 July 2022

CITATION
Chen Z, Contijoch F, Colvert GM, Manohar A, Kahn AM, Narayan HK and McVeigh E (2022) Detection of left ventricular wall motion abnormalities from volume rendering of 4DCT cardiac angiograms using deep learning.
Front. Cardiovasc. Med. 9:919751.
doi: 10.3389/fcvm.2022.919751

Detection of left ventricular wall motion abnormalities from volume rendering of 4DCT cardiac angiograms using deep learning

Zhennong Chen[1], Francisco Contijoch[1,2], Gabrielle M. Colvert[1], Ashish Manohar[3], Andrew M. Kahn[4], Hari K. Narayan[5] and Elliot McVeigh[1,2,4]*

[1]Department of Bioengineering, UC San Diego School of Engineering, La Jolla, CA, United States, [2]Department of Radiology, UC San Diego School of Medicine, La Jolla, CA, United States, [3]Department of Mechanical and Aerospace Engineering, UC San Diego School of Engineering, La Jolla, CA, United States, [4]Department of Cardiology, UC San Diego School of Medicine, La Jolla, CA, United States, [5]Department of Pediatrics, UC San Diego School of Medicine, La Jolla, CA, United States

Background: The presence of left ventricular (LV) wall motion abnormalities (WMA) is an independent indicator of adverse cardiovascular events in patients with cardiovascular diseases. We develop and evaluate the ability to detect cardiac wall motion abnormalities (WMA) from dynamic volume renderings (VR) of clinical 4D computed tomography (CT) angiograms using a deep learning (DL) framework.

Methods: Three hundred forty-three ECG-gated cardiac 4DCT studies (age: 61 ± 15, 60.1% male) were retrospectively evaluated. Volume-rendering videos of the LV blood pool were generated from 6 different perspectives (i.e., six views corresponding to every 60-degree rotation around the LV long axis); resulting in 2058 unique videos. Ground-truth WMA classification for each video was performed by evaluating the extent of impaired regional shortening visible (measured in the original 4DCT data). DL classification of each video for the presence of WMA was performed by first extracting image features frame-by-frame using a pre-trained Inception network and then evaluating the set of features using a long short-term memory network. Data were split into 60% for 5-fold cross-validation and 40% for testing.

Results: Volume rendering videos represent ~800-fold data compression of the 4DCT volumes. Per-video DL classification performance was high for both cross-validation (accuracy = 93.1%, sensitivity = 90.0% and specificity = 95.1%, κ: 0.86) and testing (90.9, 90.2, and 91.4% respectively, κ: 0.81). Per-study performance was also high (cross-validation: 93.7, 93.5, 93.8%, κ: 0.87; testing: 93.5, 91.9, 94.7%, κ: 0.87). By re-binning per-video results into the 6 regional views of the LV we showed DL was accurate (mean accuracy = 93.1 and 90.9% for cross-validation and testing cohort, respectively) for every region. DL classification strongly agreed (accuracy = 91.0%, κ: 0.81) with expert visual assessment.

Conclusions: Dynamic volume rendering of the LV blood pool combined with DL classification can accurately detect regional WMA from cardiac CT.

KEYWORDS

computed tomography, left ventricle (LV), wall motion abnormality detection, volume rendering (VR), deep learning

Introduction

Left Ventricular (LV) wall motion abnormalities (WMA) are an independent indicator of adverse cardiovascular events and death in patients with cardiovascular diseases such as myocardial infarction (MI), dyssynchrony and congenital heart disease (1, 2). Further, regional WMA have greater prognostic values after acute MI than LV ejection fraction (EF) (3, 4). Multidetector computed tomography (CT) is routinely used to evaluate coronary arteries (5, 6). Recently, ECG-gated acquisition of cardiac 4DCT enables the combined assessment of coronary anatomy and LV function (7, 8). Recent publications show that regional WMA detection with CT agrees with echocardiography (9, 10) as well as with cardiac magnetic resonance (11, 12).

Dynamic information of the 3D cardiac motion and regional WMA is encoded in 4DCT data. Visualization of regional WMA with CT usually requires reformatting the acquired 3D data along standard 2D short- and long-axis imaging planes. However, it requires experience in practice to resolve the precise region of 3D wall motion abnormalities from these 2D planes. Further, these 2D plane views may be confounded by through-plane motion and foreshortening artifacts (13). We propose to directly view 3D regions of wall motion abnormalities through the use of volumetric visualization techniques such as volume rendering (VR) (14), which can preserve high resolution anatomical information and visualize 3D (15, 16) and 4D (17) data simultaneously over large regions of the LV in cardiovascular CT. In VR, the 3D CT volume is projected onto a 2D viewing plane and different colors and opacities are assigned to each voxel based on intensity. It has been shown that VR provides a highly representative and memory efficient way to depict 3D tissue structures and anatomic abnormalities (18, 19). In this paper, we perform dynamic 4D volume rendering by sequentially combining the VR of each CT time frame into a video of LV function (we call this video a "Volume Rendering video"). We propose to use volume rendering videos of 4DCT data to depict 3D motion dynamics and visualize highly local wall motion dynamics to detect regional WMA.

Analytical approaches to quantify 3D motion from 4DCT using image registration and deformable LV models have been developed (9, 20, 21). However, these approaches usually require complex and time-consuming steps such as user-guided image segmentation and point-to-point registration or feature tracking. Further, analysis of multiple frames at the native image resolution/size of 4DCT can lead to significant memory limitations (22), especially when running deep learning experiments using current graphical processing units (GPU). Volume rendering (VR) videos provide a high-resolution representation of 4DCT data which clearly depicts cardiac motion at a significantly reduced memory footprint (~1 Gigabyte when using original 4DCT for motion analysis and only 100 kilobytes when using volume rendering video). Given the lack of methods currently available to analyze motion observed in VR videos, we sought to create an objective observer that could automate VR video interpretation. Doing so would facilitate clinical adoption as it would avoid the need for training individuals on VR video interpretation and the approach could be readily shared. Deep learning approaches have been successfully used to perform classification of patients using medical images (23, 24). Further, DL methods, once trained, are very inexpensive and can be easily deployed.

Therefore, in this paper, we propose a novel framework which combines volume rendering videos of clinical cardiac CT cases with a DL classification to detect WMA. We outline a straightforward process to generate VR videos from 4DCT data and then utilize a combination of a convolutional neural network (CNN) and recurrent neural network (RNN) to assess regional WMA observable in the videos.

Methods and materials

CT data collection

Under institutional review board approval, 343 ECG-gated contrast enhanced cardiac CT patient studies between Jan 2018 and Dec 2020 were retrospectively collected with waiver of informed consent. Inclusion criteria were: each study (a) had images reconstructed across the entire cardiac cycle, (b) had a field-of-view which captured the entire LV, (c) was free from significant pacing lead artifact in the LV and (d) had a radiology report including assessment of cardiac function. Images were collected by a single, wide detector CT scanner with 256 detector rows (Revolution scanner, GE Healthcare, Chicago IL) allowing for a single heartbeat axial 16cm acquisition across the cardiac cycle. The CT studies were performed for range of clinical cardiac indications including suspected coronary artery disease (CAD, $n = 153$), pre-procedure assessment of pulmonary vein isolation (PVI, $n = 126$), preoperative assessment of transcatheter aortic valve replacement (TAVR, $n = 42$), preoperative assessment of left ventricular assist device placement (LVAD, $n = 22$).

Production of volume rendering video of LV blood-pool

Figure 1 step 1-4 shows the pipeline of VR video production. The CT images were first rotated using visual landmarks such as the RV insertion and LV apex, so that every study had the same orientation (with the LV long axis along the z-axis of the images and the LV anterior wall at 12 o'clock in cross-sectional planes). Structures other than LV blood-pool (such as LV myocardium, ribs, the right ventricle, and great vessels) were automatically removed by a pre-trained DL segmentation U-Net (25) which has previously shown high accuracy in localizing the LV in CT images (25, 26). If present, pacing leads were removed manually.

The resultant grayscale images of the LV blood-pool (as shown in Fig. 1 step 2) were then used to produce Volume renderings (VR) *via* MATLAB (version: 2019b, MathWorks, Natick MA). Note the rendering was performed using the native CT scan resolution. The LV endocardial surface shown in VR was defined by automatically setting the intensity window level (WL) equal to the mean voxel intensity in a small ROI placed at the centroid of the LV blood pool and setting the window width (WW) equal to 150 HU (thus WL is study-specific, and WW is uniform for every study). Additional rendering parameters are listed in Supplementary Materials 1A. VR of all frames spanning one cardiac cycle was then saved as a video ("VR video," Figure 1).

Each VR video projects the 3D LV volume from one specific projection view angle θ, thus it shows only part of the LV blood-pool and misses parts that are on the backside. Therefore, to see and evaluate all AHA segments, 6 VR videos were generated per study, with six different projection views $\theta_{60\times n},\ \ n\in[0,1,2,3,4,5]$ corresponding to 60-degree rotations around the LV long axis (Supplementary Materials 1B for details). With our design, each projection view had a particular mid-cavity AHA segment shown on the foreground (meaning this segment was the nearest to and in front of the ray source-point of rendering) as well as its corresponding basal and apical segments. Two adjacent mid-cavity AHA segments and their corresponding basal and apical segments were shown on the left and right boundary of the rendering in that view. In standard regional terminology, the six projection views ($n = 0, 1, 2, 3, 4, 5$ in $\theta_{60\times n}$) looked at the LV from the view with mid-cavity Anterolateral, Inferolateral, Inferior, Inferoseptal, Anteroseptal and Anterior segments on the foreground, respectively. In this paper, to simplify the text we call them six "regional LV views" from anterolateral to anterior. In total, a *large* dataset of 2058 VR videos (343 patients × 6 views) with unique projections were generated.

Classification of wall motion

Figure 1 steps a-d shows how the ground truth presence or absence of WMA at each location on the endocardium was determined. It is worth clarifying first that the ground truth is made on the original CT data not the volume rendered data. First, voxel-wise LV segmentations obtained using the U-Net were manually refined in ITK-SNAP (Philadelphia, PA, USA) (27). Then, regional shortening (RS_{CT}) (8, 28, 29) of the endocardium was measured using a previously-validated surface feature tracking (21) technique. The accuracy of RS_{CT} in detecting WMA has been validated previously with strain measured by tagged MRI (12) [a validated non-invasive approach for detecting wall motion abnormalities in myocardial ischemia (30, 31)]. Regional shortening was calculated at each face on the endocardial mesh as:

$$RS_{CT} = \sqrt{\frac{Area_{ES}}{Area_{ED}}} - 1$$

where $Area_{ES}$ is the area of a local surface mesh at end-systole (ES) and $Area_{ED}$ is the area of the same mesh at end-diastole (ED). ED and ES were determined based on the largest and smallest segmented LV blood-pool volumes, respectively. RS_{CT} for an endocardial surface voxel was calculated as the average RS_{CT} value of a patch of mesh faces directly connected with this voxel. RS_{CT} values were projected onto pixels in each VR video view (see Supplementary Material 2 for details about projection) to generate a ground truth map of endocardial function for each region from the perspective of each VR video. Then, each angular position was classified as abnormal (WMA present) if >35% of the endocardial surface in that view had impaired RS_{CT} (RS_{CT} ≥-0.20). Supplementary Material 2A explains how these thresholds were selected.

To do per-study classification in this project, we defined that a CT study is abnormal if it has more than one VR videos labeled as abnormal ($N_{ab\_videos} \geq 2$). Other thresholds (e.g., $N_{ab\_videos} \geq 1$ or 3) were also chosen and the corresponding results were shown in the Supplementary Material 3.

DL framework design

The DL framework (see Figure 2) consists of three components, (*a*) a pre-trained 2D convolutional neural network (CNN) used to extract spatial features from each input frame of a VR video, (*b*) a recurrent neural network (RNN) designed

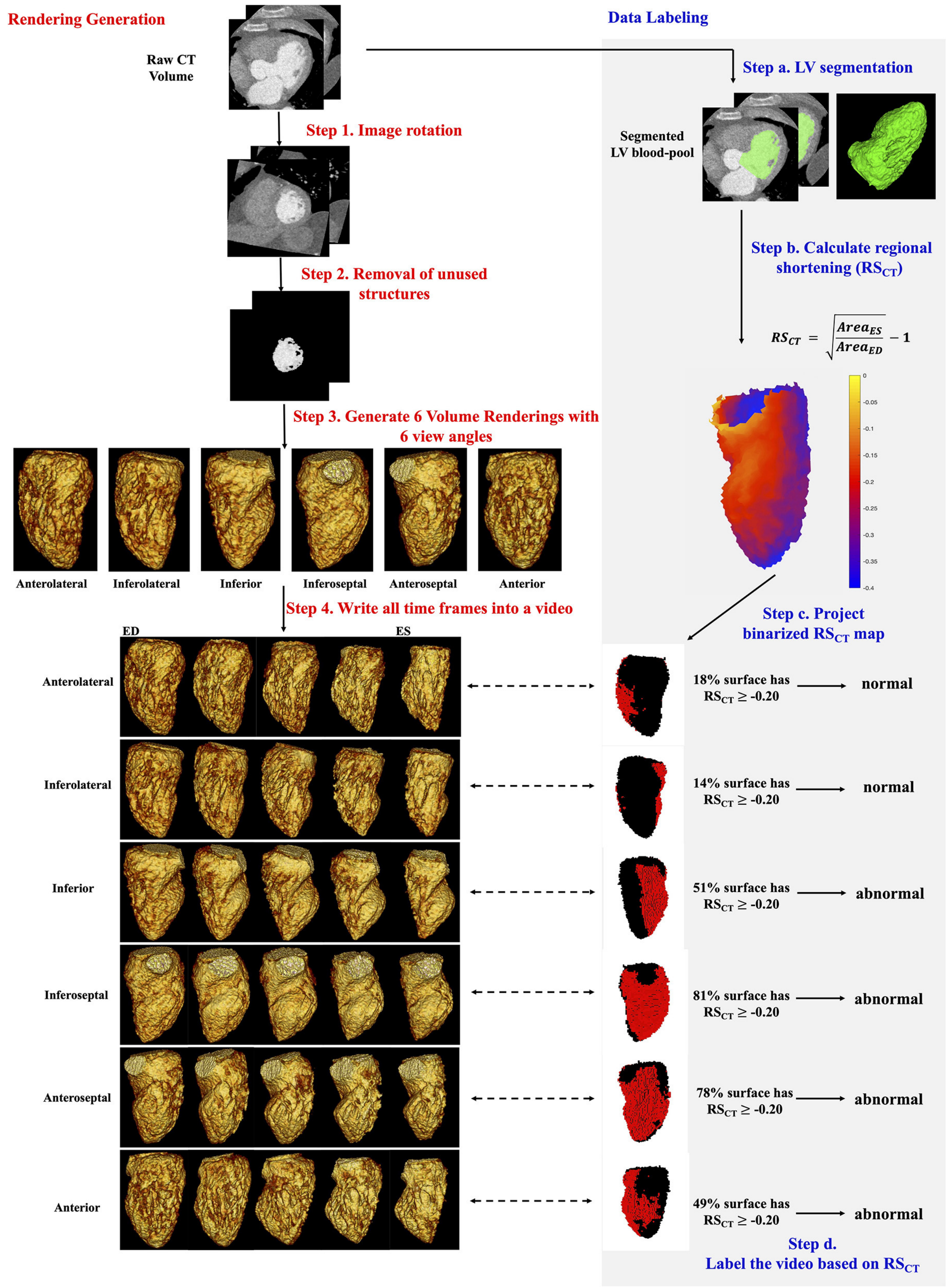

Rendering Generation
Raw CT Volume
Step 1. Image rotation
Step 2. Removal of unused structures
Step 3. Generate 6 Volume Renderings with 6 view angles
Anterolateral
Inferolateral
Inferior
Inferoseptal
Anteroseptal
Anterior
Step 4. Write all time frames into a video
ED
ES
Anterolateral
Inferolateral
Inferior
Inferoseptal
Anteroseptal
Anterior
Data Labeling
Step a. LV segmentation
Segmented LV blood-pool
Step b. Calculate regional shortening (RS_{CT})
$RS_{CT} = \sqrt{\frac{Area_{ES}}{Area_{ED}}} - 1$
0
-0.05
-0.1
-0.15
-0.2
-0.25
-0.3
-0.35
-0.4
Step c. Project binarized RS_{CT} map
18% surface has $RS_{CT} \geq -0.20$
normal
14% surface has $RS_{CT} \geq -0.20$
normal
51% surface has $RS_{CT} \geq -0.20$
abnormal
81% surface has $RS_{CT} \geq -0.20$
abnormal
78% surface has $RS_{CT} \geq -0.20$
abnormal
49% surface has $RS_{CT} \geq -0.20$
abnormal
Step d. Label the video based on RS_{CT}

FIGURE 1
Automatic generation and quantitative labeling of volume rendering video. This figure contains two parts: Rendering Generation: automatic generation of VR video (left column, white background, step 1-4 in red) and Data Labeling: quantitative labeling of the video (right column, light gray background, step a-d in blue). Rendering Generation: *Step 1 and 2*: Prepare the greyscale image of LV blood-pool with all other structures

(Continued)

FIGURE 1
removed. *Step 3*: For each study, 6 volume renderings with 6 view angles rotated every 60 degrees around the long axis were generated. The mid-cavity AHA segment in the foreground was noted under each view. *Step 4*: For each view angle, a volume rendering video was created to show the wall motion across one heartbeat. Five systolic frames in VR video were presented. ED, end-diastole; ES, end-systole. Data Labeling: *Step* a: LV segmentation. LV, green. *Step b*: Quantitative RS_{CT} was calculated for each voxel. *Step c*: The voxel-wise RS_{CT} map was binarized and projected onto the pixels in the VR video. See Supplementary Material 2 for more details. In rendered RS_{CT} map, the pixels with $RS_{CT} \geq -0.20$ (abnormal wall motion) were labeled as red and those with $RS_{CT} < -0.20$ (normal) were labeled as black. *Step d*: a video was labeled as abnormal if >35% endocardial surface has $RS_{CT} \geq -0.20$ (red pixels).

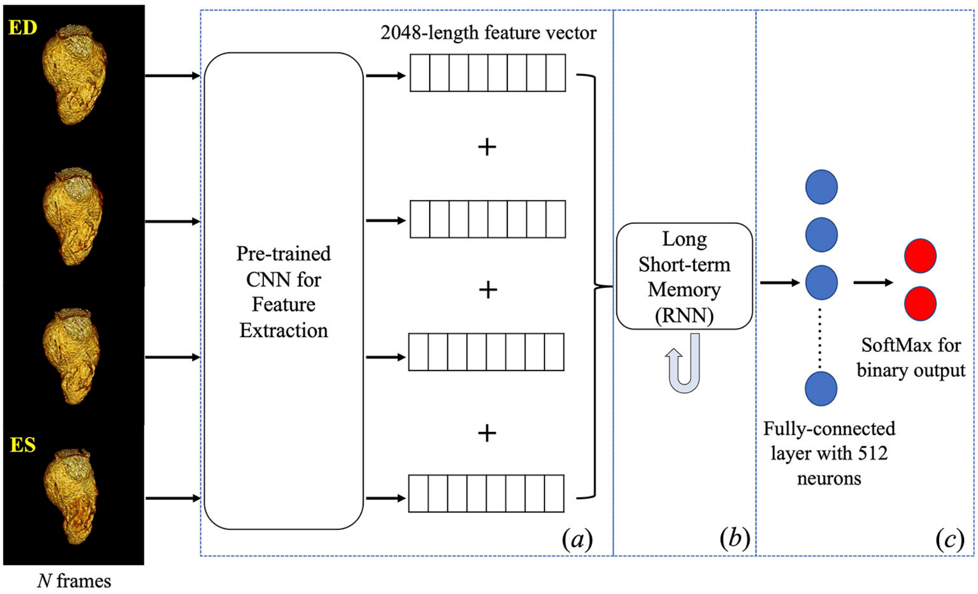

ED
ES
N frames
Pre-trained CNN for Feature Extraction
2048-length feature vector
+
+
+
Long Short-term Memory (RNN)
Fully-connected layer with 512 neurons
SoftMax for binary output
(a)
(b)
(c)

FIGURE 2
Deep learning framework. Four frames were input into a pre-trained inception-v3 individually to obtain a 2048-length feature vector for each frame. Four vectors were concatenated into a feature matrix which was then input to the next components in the framework. A Long Short-term Memory followed by fully connected layers was trained to predict a binary classification of the presence of WMA in the video. CNN, convolutional neural network; RNN, recurrent neural network.

to incorporate the temporal relationship between frames, and (*c*) a fully connected neural network designed to output the classification.

Given our focus on systolic function, four frames (ED, two systolic frames, and ES) were input to the DL architecture. This sampling was empirically found to maximize DL performance (32). Given the CT gantry rotation time, this also minimizes view sharing present in each image frame while providing a fuller picture of endocardial deformation. Each frame was resampled to 299×299 pixels to accommodate the input size of the pre-trained CNN.

Component (*a*) is a pre-trained CNN with the Inception architecture (Inception-v3) (33) and the weights obtained after training on the ImageNet (34) database. The reason to pick Inception-v3 architecture can be found in this reference (32). This component was used to extract features and create a 2048-length feature vector for each input image. Feature vectors from the four frames were then concatenated into a 2D feature matrix with size = (4, 2048).

Component (*b*) is a long short-term memory (35) RNN with 2048 nodes, tanh activation and sigmoid recurrent activation. This RNN analyzed the (4, 2048) feature matrix from component (*a*) to synthesize temporal information (RNN does this by passing the knowledge learned from the previous instance in a sequence to the learning process of the current instance in that sequence then to the next instance). The final component (*c*), the fully connected layer, logistically regressed the binary prediction of the presence of WMA in the video.

Cross-validation and testing

In our DL framework, component (*a*) was pre-trained and directly used for feature extraction whereas components (*b*) and (*c*) were trained end-to-end as one network for WMA classification. Parameters were initialized randomly. The loss function was categorical cross-entropy.

The dataset was split randomly into 60% and 40% subsets. 60% (205 studies, 1230 videos) were used for 5-fold cross-validation, meaning in each fold of validation we had 164 studies (984 videos) to train the model and the rest 41 studies (246 videos) to validate the model. We report model performance across all folds. 40% (138 studies, 828 videos) were used only for testing.

Experiment settings

We performed all DL experiments using TensorFlow on an 8-core Ubuntu workstation with 32 GB RAM and with a GeForce GTX 1080 Ti (NVIDIA Corporation, Santa Clara, CA, USA). The file size of each 4DCT study and VR video were recorded. Further, the time needed to run each step in the entire framework (including the image processing, VR video generation and DL prediction) on the new cases was recorded.

Model performance and LVEF

The impact of systolic function, measured *via* LVEF on DL classification accuracy was evaluated in studies with LVEF <40%, LVEF between 40-60%, LVEF >60%. We hypothesized that the accuracy of the model would be different for different LVEF intervals since because the "obviously abnormal" LV with low EF, and the "obviously normal" LV with high EF would be easier to classify. The consequence of a local WMA in hearts with LVEF between 40-60% might be a more subtle pattern and harder to detect. These subtle cases are also difficult for human observers.

Comparison with expert visual assessment

While not the primary goal of the study we investigated the consistency of the DL classifications with the results from two human observers using traditional views. 100 CT studies were randomly selected from the testing cohort for independent analysis of WMA by two cardiovascular imaging experts with different levels of experiences: expert 1 with >20 years of experience (author A.K.) and expert 2 with >5 years of experience (author H.K.N.) The experts classified the wall motion in each AHA segment into 4 classes (normal, hypokinetic, akinetic and dyskinetic) by visualizing wall motion from standard 2D short- and long-axis imaging planes, in a blinded fashion. Because of the high variability in the inter-observer classifications of abnormal categories we: (1) combined the last three classes into a single "abnormal" class indicating WMA detection, and (2) we performed the comparison on a per-study basis. A CT study was classified as abnormal by the experts if it had more than one abnormal segment. The interobserver variability is reported in the result Section Model performance-comparison with expert assessment. It should be noted that our model was only trained on ground truth based on quantitative RS_{CT} values; the expert readings were performed as a measure of consistency with clinical performance.

Statistical evaluation

Two-tailed categorical z-test was used to compare data proportions (e.g., proportions of abnormal videos) in two independent cohorts: a cross-validation cohort and a testing cohort. Statistical significance was set at $P \leq 0.05$.

DL Model performance against the ground truth label was reported *via* confusion matrix and Cohen's kappa value. Both regional (per-video) and per-study comparison were performed. A CT study is defined as abnormal if it has more than one VR videos labeled as abnormal ($N_{ab\_videos} \geq 2$). As stated in Section Production of volume rendering video of LV blood-pool, every projection view of the VR video corresponded to a specific regional LV view. Therefore, we re-binned the per-video results into 6 LV views to test the accuracy of the DL model when looking at each region of the LV. We also calculated the DL per-study accuracy for patients with each clinical cardiac indication in the testing cohort and use pair-wise Chi-squared test to compare the accuracies between indications.

Results

Of the 1230 views (from 205 CT studies) used for 5-fold cross-validation, 732 (from 122 studies, 59.5%) were male (age: 63 ± 15) and 498 (from 83 studies, 40.5%) were female (age: 62 ± 15). The LV blood pool had a median intensity of 516 HU (IQR: 433 to 604). 40.0% (492/1230) of the videos were labeled as abnormal based on RS_{CT} analysis, and 45.4% (93/205) of studies had WMA in ≥2 videos. 104 studies had LVEF > 60%, 54 studies had LVEF < 40% and the rest 47 (47/205 = 22.9%) studies had LVEF between 40-60%. For clinical cardiac indications, 85 studies have suspect CAD, 77 studies have the pre-PVI assessment, 31 studies have the pre-TAVR assessment, and 12 studies have the pre-VAD assessment.

Of the 828 views (from 138 CT studies) used for testing, 504 (from 84 studies, 60.9%) were male (age: 57 ± 16) and 324 (from 54 studies, 39.1%) were female (age: 63 ± 13). The LV blood pool had a median intensity of 520 HU (IQR: 442 to 629). 37.0% (306/828) of the videos were labeled as abnormal, and 45.0% (62/138) of studies had WMA in ≥2 videos. 72 studies had LVEF > 60%, 25 studies had LVEF < 40% and the rest 41 (41/138 =

TABLE 1 DL classification performance in cross-validation and testing.

		Cross-validation				Testing			
		Per-video		Per-study ($N_{ab\_videos} \geq 2$)		Per-video		Per-study ($N_{ab\_videos} \geq 2$)	
		Ground truth		Ground truth		Ground truth		Ground truth	
		Abnormal	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal	Normal
DL	Abnormal	443	36	87	7	276	45	57	4
	Normal	49	702	6	105	30	477	5	72
		Sens	0.900	Sens	0.935	Sens	0.902	Sens	0.919
		Spec	0.951	Spec	0.938	Spec	0.914	Spec	0.947
		Acc	0.931	Acc	0.937	Acc	0.909	Acc	0.935
		κ	0.855	κ	0.872	κ	0.808	κ	0.868

Two hundred five CT studies and 1230 Volume Rendered (VR) videos were used for 5-fold cross-validation. One hundred thirty-eight CT studies and 828 VR videos were in the testing. The four confusion matrices correspond to per-video classification (light gray) and per-study classification (dark gray) for cross-validation (left) and testing (right). $N_{ab\_videos} \geq 2$ (number of views classified as abnormal) was used to classify a study as abnormal. Sens, sensitivity; Spec, specificity; Acc, accuracy. Cohen's kappa κ is also reported.

28.7%) studies had LVEF between 40-60%. For clinical cardiac indications, 68 studies have suspect CAD, 49 studies have the pre-PVI assessment, 11 studies have the pre-TAVR assessment, and 10 studies have the pre-VAD assessment.

There were no significant differences (all P-values > 0.05) in data proportions between the cross-validation and testing cohorts in terms of the percentages of sex, abnormal videos, abnormal CT studies.

Model performance—per-video and per-study classification

Per-video and per-study DL classification performance for WMA were excellent in both cross-fold validation and testing. Table 1 shows that the per-video classification for the *cross-validation* had high accuracy = 93.1%, sensitivity = 90.0% and specificity = 95.1%, Cohen's kappa κ = 0.86 with 95% CI as [0.83, 0.89]. Per-study classification also had excellent performance with accuracy = 93.7%, sensitivity = 93.5% and specificity = 93.8%, κ = 0.87[0.81, 0.94]. Table 1 also shows that the per-video classification for the *testing cohort* had high accuracy = 90.9%, sensitivity = 90.2% and specificity = 91.4%, κ = 0.81[0.77, 0.85]. We obtained per-study classification accuracy = 93.5%, sensitivity = 91.9% and specificity = 94.7%, κ = 0.87[0.78, 0.95] in the testing cohort.

Figure 3 shows the relationship between DL classification accuracy and LVEF in the cross-validation. Table 2 shows that CT studies with LVEF between 40 and 60% in the cross-validation cohort were classified with per-video accuracy = 78.7%, sensitivity = 78.0% and specificity = 79.8%. In the testing cohort, per-video classification accuracy = 80.1%, sensitivity = 82.9% and specificity = 75.5% accuracy for this LVEF group remained relatively high but was lower (P < 0.05) than the accuracy for patients with LVEF < 40% and LVEF > 60% due to the more difficult nature of the classification task in this group with more "subtle" wall motion abnormalities.

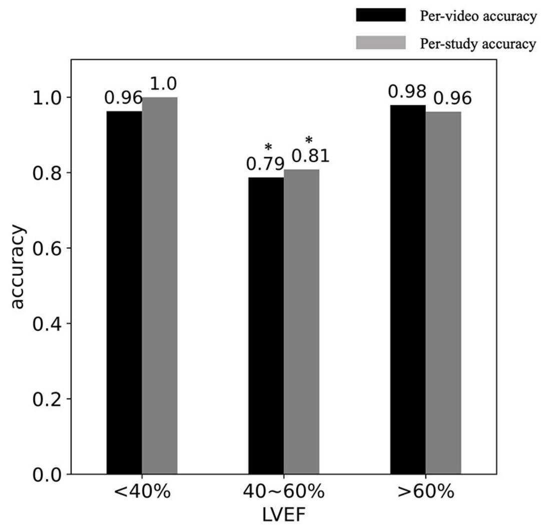


FIGURE 3
DL classification accuracy vs. LVEF. The per-video (black) and per-study (gray) accuracy are shown in studies with (LVEF < 40%), (40 < LVEF < 60%) and (LVEF > 60%). *Indicates the significant difference.

Model performance—regional LV views

Table 3 shows that our DL model was accurate for detection of WMA in all 6 regional LV views both in cross-validation cohort (mean accuracy = 93.1% ± 0.03) and testing cohort (mean accuracy = 90.9% ± 0.06).

TABLE 2 DL classification performance in CT studies with 40 < LVEF < 60%.

		Cross-validation				Testing			
		Per-video		Per-study ($N_{ab\_videos} \geq 2$)		Per-video		Per-study ($N_{ab\_videos} \geq 2$)	
		Ground truth		Ground truth		Ground truth		Ground truth	
		Abnormal	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal	Normal
DL	Abnormal	131	23	33	5	126	23	32	3
	Normal	37	91	4	5	26	71	1	5
		Sens	0.780	Sens	0.892	Sens	0.829	Sens	0.970
		Spec	0.798	Spec	0.500	Spec	0.755	Spec	0.625
		Acc	0.787	Acc	0.809	Acc	0.801	Acc	0.902
		κ	0.567	κ	0.407	κ	0.581	κ	0.657

Forty-seven CT studies with 40% < LVEF < 60% were in the cross-validation and 41 CT studies were in the testing. The light gray indicates per-video evaluation, dark gray indicates per-study evaluation.

TABLE 3 Results re-binned into six regional LV views.

		Per-video classification							
		Cross-validation				Testing			
Projection view	LV wall on the foreground	Sens	Spec	Acc	κ	Sens	Spec	Acc	κ
0	Anterolateral	0.845	0.964	0.922	0.824	0.886	0.936	0.920	0.818
60	Inferolateral	0.938	0.952	0.946	0.888	0.909	0.915	0.913	0.805
120	Inferior	0.879	0.974	0.932	0.860	0.917	0.910	0.913	0.824
180	Inferoseptal	0.882	0.946	0.917	0.832	0.847	0.861	0.855	0.705
240	Anteroseptal	0.963	0.944	0.951	0.899	0.927	0.952	0.942	0.879
300	Anterior	0.893	0.931	0.917	0.822	0.932	0.904	0.913	0.807

This table shows the per-video classification of our DL model when detecting WMA from each regional view of LV. See the definition of regional LV views in Section Production of volume rendering video of LV blood-pool. Sens, sensitivity; Spec, specificity; Acc, accuracy.

Model performance—different clinical cardiac indications

We calculated the DL per-study classification accuracy equal to 91.2% for CT studies with suspect CAD ($n = 68$ in the testing cohort), 93.9% for studies with pre-PVI assessment ($n = 49$), 100% for patients with pre-TAVR assessment ($n = 11$), 100% for studies with pre-LVAD assessment ($n = 10$). Using Chi-squared test pairwise, there was no significant difference of DL performance between indications (all P-values > 0.5).

TABLE 4 Comparison between DL and expert visual assessment.

		Expert visual assessment			
		Expert 1		Expert 2	
		Abnormal	Normal	Abnormal	Normal
DL	Abnormal	37	5	33	9
	Normal	4	54	4	54
		κ	0.815	κ	0.729

Per-study comparison were run on 100 CT studies randomly selected from the testing cohort. The light gray indicates per-video evaluation, dark gray indicates per-study evaluation.

Model performance—comparison with expert assessment

First, we report the interobserver variability of two experts. The Cohen's kappa for the agreement between observers on per-AHA-segment basis was 0.81[0.79, 0.83] and on the per-CT-study basis was 0.88[0.83, 0.93]. For those segments labeled as abnormal by both experts, the Kappa for the two experts to further classify an abnormal segment into hypokinetic, akinetic and dyskinetic dramatically dropped to 0.34.

Second, we show in the Table 4 that per-study comparison between DL prediction and expert visual assessment on 100 CT studies in the testing cohort led to Cohen's Kappa $\kappa = 0.81[0.70,0.93]$ for expert 1 and $\kappa = 0.73[0.59,0.87]$ for expert 2.

Data-size reduction

The average size of the CT study across one cardiac cycle was 1.52 ± 0.67 Gigabytes. One VR video was 341 ± 70 Kilobytes, resulting in 2.00 ± 0.40 Megabytes for 6 videos per study. VR videos led to a data size that is ∼800 times smaller than the conventional 4DCT study.

Run time

Regarding image processing, the image rotation took 14.1 ± 1.2 seconds to manually identify the landmarks and then took 38.0 ± 16.2 seconds to automatically rotate the image using the direction vectors derived from landmarks. The DL automatic removal of unnecessary structures took 141.0 ± 20.3 seconds per 4DCT study. If needed, manual pacing lead artifacts removal took around 5–10 mins per 4DCT study depending on the severity of artifacts. Regarding automatic VR video generation, it took 32.1 ± 7.0 seconds (to create 6 VR videos from the processed CT images). Regarding DL prediction of WMA presence in one CT study, it took 0.7 ± 0.1 seconds to extract image features from frames of the video and took ∼0.1 seconds to predict binary classification for all 6 VR videos in the study. To summarize, the entire framework requires approximately 4 minutes to evaluate a new study if no manual artifacts removal is needed.

Discussions

In this study, we developed and evaluated a DL framework that detects the presence of WMA in dynamic 4D volume rendering (VR videos) depicting the motion of the LV endocardial boundary. VR videos enabled a highly compressed (in terms of memory usage) representation of large regional fields of view with preserved high spatial-resolution features in clinical 4DCT data. Our framework analyzed four frames spanning systole extracted from the VR video and achieved high per-video (regional LV view) and per-study accuracy, sensitivity and specificity (≥0.90) and concordance ($\kappa \geq 0.8$) both in cross-validation and testing.

Benefits of the volume visualization approach

Assessment of regional WMA with CT is usually performed on 2D imaging planes reformatted from the 3D volume. However, 2D approaches often confuse the longitudinal bulk displacement of tissue into and out of the short-axis plane with true myocardial contraction. Various 3D analytical approaches (9, 20, 28) to quantify 3D motion using image registration and deformable LV models have been developed; our novel use of regional VR videos as input to DL networks has several benefits when compared to these traditional methods. First, VR videos contain 3D endocardial surface motion features which are visually apparent. This enables simultaneous observation of the complex 3D motion of a large region of the LV in a single VR video instead of requiring synthesis of multiple 2D slices. Second, our framework is extremely memory efficient with reduced data size while preserving key anatomical and motion information; a set of 6 VR videos is ∼800 times smaller in data size than the original 4DCT data. The use of VR videos also allows our DL experiments to run on the current graphic processing unit (GPU), whereas the original 4DCT data is too large to be imported into the GPU. Third, our framework is simple as it does not require complex and time-consuming computations such as point registration or motion field estimation included in analytical approaches. The efficiency of our technique will enable retrospective analysis of large numbers of functional cardiac CT studies; this cannot be said for traditional 3D tracking methods which require significant resources and time for segmentation and analysis.

Model performance for each LV view

We re-binned the per-video results into 6 projection views corresponding to 6 regional LV views and showed that our DL model is accurate to detect WMA from specific regions of the LV. The results shown in Table 3 indicate that all results for classification can be labeled with a particular LV region. For example, to evaluate the wall motion on the inferior wall of a CT study, the classification from the VR video with the corresponding projection view θ (=120) would be used.

Comparison with experts and its limitations

To evaluate the consistency of our model with standard clinical evaluation, we compared DL results with two cardiovascular imaging experts and showed high per-study classification correspondence. This comparison study has its limitations. First, we did not perform a per-AHA-segment comparison. Expert visual assessment was subjective (by definition) and had greater inter-observer variability on per-AHA-segment basis than the per-study basis the variability (Kappa increased from 0.81 for per-segment to 0.88 for per-study). Second, the interobserver agreement for experts to further classify an abnormal motion as hypokinetic, akinetic or dyskinetic was also too poor (Kappa = 0.34) to use expert visual labels for three severities as the ground truth; therefore, we used one "abnormal" class instead of three levels of severity of WMA. Third, experts could only visualize the wall motion from 2D imaging planes while our DL model evaluated the 3D wall motion from VR videos. A future study using a larger number of observers, and a larger number of cases could be performed in which trends could be observed; however, it is clear that variability in subjective calls for degree of WMA will likely persist in the expert readers.

Using RS_{CT} for ground truth labeling

Direct visualization of wall motion abnormalities in volume rendered movies from 4DCT is a truly original application; hence, as can be expected there are no current clinical standards/guidelines for visual detection of WMA from volume rendered movies. In fact, we believe our paper is the first to introduce this method of evaluating myocardial function in a formal pipeline. In our recent experience, visual detection of patches of endocardial "stasis" in these 3D movies highly correlates with traditional markers of WMA such as wall thickening, circumferential shortening and longitudinal shortening. However, specific guidance on how to clinically interpret VR movies is not yet available. We expect human interpretation to depend on both experience and training. Thus, we used quantitative regional myocardial shortening (RS_{CT}) derived from segmentation and 3D tracking to delineate regions of endocardial WMA. RS_{CT} has been previously shown to be a robust method for quantifying regional LV function (8, 12, 28, 29).

Limitations and future directions

First, our current DL pipeline has several manual image processing such as manual rotation of the image and manual removal of lead artifacts. These steps lengthen the time required to run the entire pipeline (see Section Run time) and limit the clinical utility. One important future direction of our technique is to integrate the DL-driven automatic image processing to get a fully automatic pipeline. Chen et al. (26) have proposed a DL technique to define the short-axis planes from CT images so that the LV axis can be subsequently derived for correct image orientation. Zhang and Yu (36) and Ghani and Karl (37) have proposed DL techniques to remove the lead artifacts.

Second, our work only focuses on the systolic function and only takes 4 systolic frames from the VR video as the model input. The future direction is to input diastolic frames into the model to enable the evaluation of diastolic function and to use a 4D spatial-temporal convolutional neural network (38) to directly process the video without requiring explicit selection of temporal frames.

Third, we currently perform binary classification of the presence of WMA in the video. The DL model integrates all information from all the AHA segments that can be seen in the video and only evaluates the extent of pixels with WMA (i.e., whether it's larger than 35% of the total pixels). The DL evaluation is independent of the position of WMA; thus, we do not identify which of the AHA segments contribute to the WMA just based on the DL binary classification. Future research is needed to "focus" the DL model's evaluation on specific AHA segments using such as local attention (39) and evaluate whether the approach can delineate the location and extent of WMA in terms of AHA segments. Further, by using a larger dataset with a balanced distribution of all four severities of WMA, we aim to train the model to estimate the severity of the WMA in the future.

Fourth, tuning the inceptionV3 (the CNN) weights to extract features most relevant to detection of WMA is expected to further increase performance as it would further optimize how the images are analyzed. However, given our limited training data, we chose not to train weights of the inception network and the high performance we observed seems to have supported this choice.

In conclusion, we developed a framework that combines the video of the volume rendered LV endocardial blood pool with deep learning classification to detect WMA and observed high per-region (per-video) and per-study accuracy. This approach has promising clinical utility to screen for cases with WMA simply and accurately from highly compressed data.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Ethics statement

The studies involving human participants were reviewed and approved by University of California, San Diego, Human Research Protections Program, IRB_approval_#191797. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

ZC, FC, and EM designed the overall study and performed the final analysis. ZC developed, trained, and validated the deep learning network, collected all the retrospective cardiac 4DCT studies, performed data curation, and drafted the whole manuscript. GC, AM, and ZC designed the pipeline to measure

RS_{CT}. AK and HN provided the expert visual assessment on 100 CT studies. All authors participated in the analysis, interpretation of data, revising the manuscript critically, and final approval of the submitted manuscript.

Funding

FC was supported by NHLBI HL145817, HL143113, and HL144678. EM and ZC were supported by HL144678 and HL153250. AM was supported by AHA 20PRE35210261.

Conflict of interest

Author EM has founder shares in Clearpoint Neuro Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcvm.2022.919751/full#supplementary-material

References

1. Carluccio E, Tommasi S, Bentivoglio M, Buccolieri M, Prosciutti L, Corea L. Usefulness of the severity and extent of wall motion abnormalities as prognostic markers of an adverse outcome after a first myocardial infarction treated with thrombolytic therapy. *Am J Cardiol.* (2000) 85:411–5. doi: 10.1016/S0002-9149(99)00764-X

2. Cicala S, de Simone G, Roman MJ, Best LG, Lee ET, Wang W, et al. Prevalence and prognostic significance of wall-motion abnormalities in adults without clinically recognized cardiovascular disease: the strong heart study. *Circulation.* (2007) 116:143–50. doi: 10.1161/CIRCULATIONAHA.106.652149

3. Møller JE, Hillis GS, Oh JK, Reeder GS, Gersh BJ, Pellikka PA. Wall motion score index and ejection fraction for risk stratification after acute myocardial infarction. *Am Heart J.* (2006) 151:419–25. doi: 10.1016/j.ahj.2005.03.042

4. Jurado-Román A, Agudo-Quílez P, Rubio-Alonso B, Molina J, Díaz B, García-Tejada J, et al. Superiority of wall motion score index over left ventricle ejection fraction in predicting cardiovascular events after an acute myocardial infarction. *Eur Heart J Acute Cardiovasc Care.* (2019) 8:78–85. doi: 10.1177/2048872616674464

5. Chen Z, Contijoch F, Schluchter A, Grady L, Schaap M, Stayman W, et al. Precise measurement of coronary stenosis diameter with CCTA using CT number calibration. *Med Phys.* (2019) 46:5514–27. doi: 10.1002/mp.13862

6. Douglas PS, Hoffmann U, Patel MR, Mark DB, Al-Khalidi HR, Cavanaugh B, et al. Outcomes of Anatomical versus Functional Testing for Coronary Artery Disease. *N Engl J Med.* (2015) 372:1291–300. doi: 10.1056/NEJMoa1415516

7. Cardiac Computed Tomography Writing Group, Taylor Allen J, Cerqueira Manuel, Hodgson John McB, Mark Daniel, Min James, O'Gara Patrick, et al. ACCF/SCCT/ACR/AHA/ASE/ASNC/NASCI/SCAI/SCMR 2010 appropriate use criteria for cardiac computed tomography. *Circulation.* (2010) 122:e525–555. doi: 10.1161/CIR.0b013e3181fcae66

8. McVeigh ER, Pourmorteza A, Guttman M, Sandfort V, Contijoch F, Budhiraja S, et al. Regional myocardial strain measurements from 4DCT in patients with normal LV function. *J Cardiovasc Comput Tomogr.* (2018) 12:372–8. doi: 10.1016/j.jcct.2018.05.002

9. Tavakoli V, Sahba N. Cardiac motion and strain detection using 4D CT images: comparison with tagged MRI, and echocardiography. *Int J Cardiovasc Imaging.* (2014) 30:175–84. doi: 10.1007/s10554-013-0305-8

10. Buss SJ,Schulz F, Mereles D, Hosch W, Galuschky C, Schummers G, et al. Quantitative analysis of left ventricular strain using cardiac computed tomography. *Eur J Radiol.* (2014) 83:e123–130. doi: 10.1016/j.ejrad.2013.11.026

11. Kaniewska M, Schuetz GM, Willun S, Schlattmann P, Dewey M. Noninvasive evaluation of global and regional left ventricular function using computed tomography and magnetic resonance imaging: a meta-analysis. *Eur Radiol.* (2017) 27:1640–59. doi: 10.1007/s00330-016-4513-1

12. Pourmorteza A, Chen MY, van der Pals J, Arai AE, McVeigh ER. Correlation of CT-based regional cardiac function (SQUEEZ) with myocardial strain calculated from tagged MRI: an experimental study. *Int J Cardiovasc Imaging.* (2016) 32:817–23. doi: 10.1007/s10554-015-0831-7

13. Ünlü S, Duchenne J, Mirea O, Pagourelias ED, Bézy S, Cvijic M, et al. EACVI-ASE Industry Standardization Task Force. Impact of apical foreshortening on deformation measurements: a report from the EACVI-ASE Strain Standardization Task Force. *Eur Heart J Cardiovasc Imaging.* (2020) 21:337–43. doi: 10.1093/ehjci/jez189

14. Levoy M. Display of surfaces from volume data. *IEEE Comput Graph Appl.* (1988) 8:29–37. doi: 10.1109/38.511

15. Zhang Q, Eagleson R, Peters TM. Volume visualization: a technical overview with a focus on medical applications. *J Digit Imaging.* (2011) 24:640–64. doi: 10.1007/s10278-010-9321-6

16. Cutroneo G, Bruschetta D, Trimarchi F, Cacciola A, Cinquegrani M, Duca A, et al. In Vivo CT direct volume rendering: a three-dimensional anatomical description of the heart. *Pol J Radiol.* (2016) 81:21–8. doi: 10.12659/PJR.895476

17. Zhang Q, Eagleson R, Peters TM. Dynamic real-time 4D cardiac MDCT image display using GPU-accelerated volume rendering. *Comput Med Imaging Graph.* (2009) 33:461–76. doi: 10.1016/j.compmedimag.2009.04.002

18. Mor-Avi V, Sugeng L, Lang RM. Real-time 3-dimensional echocardiography: an integral component of the routine echocardiographic examination in adult patients? *Circulation.* (2009) 119:314–29. doi: 10.1161/CIRCULATIONAHA.107.751354

19. Mori S, Takaya T, Kinugasa M, Ito T, Takamine S, Fujiwara S, et al. Three-dimensional quantification and visualization of aortic calcification by multidetector-row computed tomography: a simple approach using a volume-rendering method. *Atherosclerosis.* (2015) 239:622–8. doi: 10.1016/j.atherosclerosis.2014.12.041

20. Lamash Y, Fischer A, Carasso S, Lessick J. Strain Analysis From 4-D Cardiac CT Image Data. *IEEE Trans Biomed Eng.* (2015) 62:511–21. doi: 10.1109/TBME.2014.2359244

21. Pourmorteza Amir, Schuleri Karl H, Herzka Daniel A, Lardo Albert C, McVeigh Elliot R. A new method for cardiac computed tomography regional function assessment. *Circ Cardiovasc Imaging.* (2012) 5:243–50. doi: 10.1161/CIRCIMAGING.111.970061

22. Gupta K, Sekhar N, Vigneault DM, Scott AR, Colvert B, Craine A, et al. Octree representation improves data fidelity of cardiac CT images and convolutional neural network semantic segmentation of left atrial and ventricular chambers. *Radiol Artif Intell.* (2021) 3:e210036. doi: 10.1148/ryai.2021210036

23. Zhang N, Yang G, Gao Z, Xu C, Zhang Y, Shi R, et al. Deep learning for diagnosis of chronic myocardial infarction on nonenhanced caridac cine MRI. *Radiology.* (2019) 291:606–17. doi: 10.1148/radiol.2019182304

24. Zreik M, Lessmann N, van Hamersvelt RW, Wolterink JM, Voskuil M, Viergever MA, et al. Deep learning analysis of the myocardium in coronary CT angiography for identification of patients with functionally significant coronary artery stenosis. *Med Image Anal.* (2018) 44:72–85. doi: 10.1016/j.media.2017.11.008

25. Baskaran L, Maliakal G. Al'Aref SJ, Singh G, Xu Z, Michalak K, et al. Identification and quantification of cardiovascular structures from CCTA: an end-to-end, rapid, pixel-wise, deep-learning method. *JACC Cardiovasc Imaging.* (2020) 13:1163–71. doi: 10.1016/j.jcmg.2019.08.025

26. Chen Z, Rigolli M, Vigneault DM, Kligerman S, Hahn L, Narezkina A, et al. Automated cardiac volume assessment and cardiac long- and short-axis imaging plane prediction from electrocardiogram-gated computed tomography volumes enabled by deep learning. *Eur Heart J - Digit Health.* (2021) 2:311–22. doi: 10.1093/ehjdh/ztab033

27. Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage.* (2006) 31:1116–28. doi: 10.1016/j.neuroimage.2006.01.015

28. Manohar A, Colvert GM, Schluchter A, Contijoch F, McVeigh ER. Anthropomorphic left ventricular mesh phantom: a framework to investigate the accuracy of SQUEEZ using coherent point drift for the detection of regional wall motion abnormalities. *J Med Imaging.* (2019) 6:045001. doi: 10.1117/1.JMI.6.4.045001

29. Colvert GM, Manohar A, Contijoch FJ, Yang J, Glynn J, Blanke P, et al. Novel 4DCT method to measure regional left ventricular endocardial shortening before and after transcatheter mitral valve implantation. *Struct Heart.* (2021) 5:410–9. doi: 10.1080/24748706.2021.1934617

30. Götte MJ, van Rossum AC, Twisk JWR. null, Kuijer JPA null, Marcus JT, Visser CA. Quantification of regional contractile function after infarction: strain analysis superior to wall thickening analysis in discriminating infarct from remote myocardium. *J Am Coll Cardiol.* (2001) 37:808–17. doi: 10.1016/S0735-1097(00)01186-4

31. Moore CC, McVeigh ER, Zerhouni EA. Noninvasive measurement of three-dimensional myocardial deformation with tagged magnetic resonance imaging during graded local ischemia. *J Cardiovasc Magn Reson Off J Soc Cardiovasc Magn Reson.* (1999) 1:207–22. doi: 10.3109/10976649909088333

32. Chen Z, Contijoch F, McVeigh E. Development of deep learning pipeline for direct observation of wall motion abnormality from 4DCT. In: *Medical Imaging 2022: Biomedical Applications in Molecular, Structural, and Functional Imaging.* Vol 12036. San Diego, CA: SPIE (2022). p. 429–39. doi: 10.1117/12.2607387

33. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. *ArXiv151200567 Cs.* (2015). Available online at: http://arxiv.org/abs/1512.00567 (accessed November 16, 2020).

34. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition.* (2009). p. 248–55. doi: 10.1109/CVPR.2009.5206848

35. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput.* (1997) 9:1735–80. doi: 10.1162/neco.1997.9.8.1735

36. Zhang Y, Yu H. Convolutional neural network based metal artifact reduction in X-ray computed tomography. *IEEE Trans Med Imaging.* (2018) 37:1370–81. doi: 10.1109/TMI.2018.2823083

37. Ghani MU, Karl WC. Fast enhanced CT metal artifact reduction using data domain deep learning. In: *IEEE Transactions on Computational Imaging*, Vol. 6. (2020). p. 181–193. doi: 10.1109/TCI.2019.2937221

38. Choy C, Gwak J, Savarese S. 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. *arXiv.* (2019). doi: 10.1109/CVPR.2019.00319

39. Li Z, Zhuang X, Wang H, Nie Y, Tang J. Local attention sequence model for video object detection. *Appl Sci.* (2021) 11:4561. doi: 10.3390/app11104561

frontiers | Frontiers in Cardiovascular Medicine

TYPE Original Research
PUBLISHED 11 August 2022
DOI 10.3389/fcvm.2022.952089

Deep learning assessment of left ventricular hypertrophy based on electrocardiogram

Xiaoli Zhao[1†], Guifang Huang[2†], Lin Wu[1†], Min Wang[1], Xuemin He[3], Jyun-Rong Wang[4,5], Bin Zhou[1], Yong Liu[1], Yesheng Lin[1], Dinghui Liu[1], Xianguan Yu[1], Suzhen Liang[1], Borui Tian[1], Linxiao Liu[1], Yanming Chen[3], Shuhong Qiu[2], Xujing Xie[1]*, Lanqing Han[6]* and Xiaoxian Qian[1]*

[1]Department of Cardiology, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China, [2]China Unicom (Guangdong) Industrial Internet Ltd., Guangzhou, China, [3]Department of Endocrine and Metabolic Diseases, The Third Affiliated Hospital, Sun Yat-sen University, Guangzhou, China, [4]LCFC (Hefei) Electronics Technology Co., Ltd., Hefei, China, [5]Hefei LCFC Information Technology Co., Ltd., Hefei, China, [6]Center for Artificial Intelligence, Research Institute of Tsinghua, Pearl River Delta, Guangzhou, China

Background: Current electrocardiogram (ECG) criteria of left ventricular hypertrophy (LVH) have low sensitivity. Deep learning (DL) techniques have been widely used to detect cardiac diseases due to its ability of automatic feature extraction of ECG. However, DL was rarely applied in LVH diagnosis. Our study aimed to construct a DL model for rapid and effective detection of LVH using 12-lead ECG.

Methods: We built a DL model based on convolutional neural network-long short-term memory (CNN-LSTM) to detect LVH using 12-lead ECG. The echocardiogram and ECG of 1,863 patients obtained within 1 week after hospital admission were analyzed. Patients were evenly allocated into 3 sets at 3:1:1 ratio: the training set (n = 1,120), the validation set (n = 371) and the test set 1 (n = 372). In addition, we recruited 453 hospitalized patients into the internal test set 2. Different DL model of each subgroup was developed according to gender and relative wall thickness (RWT).

Results: The LVH was predicted by the CNN-LSTM model with an area under the curve (AUC) of 0.62 (sensitivity 68%, specificity 57%) in the test set 1, which outperformed Cornell voltage criteria (AUC: 0.57, sensitivity 48%, specificity 72%) and Sokolow-Lyon voltage (AUC: 0.51, sensitivity 14%, specificity 96%). In the internal test set 2, the CNN-LSTM model had a stable performance in predicting LVH with an AUC of 0.59 (sensitivity 65%, specificity 57%). In the subgroup analysis, the CNN-LSTM model predicted LVH by 12-lead ECG with an AUC of 0.66 (sensitivity 72%, specificity 60%) for male patients, which performed better than that for female patients (AUC: 0.59, sensitivity 50%, specificity 71%).

Conclusion: Our study established a CNN-LSTM model to diagnose LVH by 12-lead ECG with higher sensitivity than current ECG diagnostic criteria. This CNN-LSTM model may be a simple and effective screening tool of LVH.

KEYWORDS

left ventricular hypertrophy, electrocardiogram, echocardiography, deep learning model, convolutional neural network-long short-term memory

OPEN ACCESS

EDITED BY
Alistair A. Young,
King's College London,
United Kingdom

REVIEWED BY
Emanuela Teresina Locati,
IRCCS San Donato Polyclinic, Italy
Stefan van Duijvenboden,
University College London,
United Kingdom

*CORRESPONDENCE
Xiaoxian Qian
qianxx@mail.sysu.edu.cn
Lanqing Han
hanlance@tsinghua-gd.org
Xujing Xie
xiexj@mail.sysu.edu.cn

[†]These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Cardiovascular Imaging, a section of the journal Frontiers in Cardiovascular Medicine

RECEIVED 24 May 2022
ACCEPTED 21 July 2022
PUBLISHED 11 August 2022

CITATION
Zhao X, Huang G, Wu L, Wang M, He X, Wang J-R, Zhou B, Liu Y, Lin Y, Liu D, Yu X, Liang S, Tian B, Liu L, Chen Y, Qiu S, Xie X, Han L and Qian X (2022) Deep learning assessment of left ventricular hypertrophy based on electrocardiogram. *Front. Cardiovasc. Med.* 9:952089. doi: 10.3389/fcvm.2022.952089

Introduction

Left ventricular hypertrophy (LVH) is an early structural and functional cardiac change of hypertension, with an estimated echocardiographic prevalence of 36–41% (1). Other causes of LVH include aortic stenosis, hypertrophic cardiomyopathy, valvular heart disease, infiltrative heart muscle disease, storage and metabolic disorders (2, 3). The incidence of LVH is further affected by age and obesity (4, 5). Previous studies showed that LVH is an independent risk of arrhythmias (3), heart failure (6) and mortality (7).

Echocardiography is the current standard diagnostic method (8), whereas 12-lead ECG is the most commonly used diagnostic tool in clinical cardiology as it allows a rapid screening of LVH. However, current ECG criteria of LVH including the Cornell voltage and the Sokolow-Lyon voltage criteria have low sensitivity (7, 9). These criteria mainly focus on increased QRS complex amplitude, but overlook a leftward shift of electrical axis in the frontal plane, ST segment deviation and T wave changes, which are also principal ECG diagnostic characteristics for LVH (10). Besides, the interpretation of these ECG criteria are tedious for doctors, affecting the efficiency and accuracy of diagnosis. To improve these limitations of the current ECG criteria for LVH, new methods for analysis of ECG are urgently needed.

Since the digitalization of ECG, artificial intelligence methods have been employed in computerized interpretation of ECGs (11). Recently, few studies were presented by machine learning for the ECG characteristics to detect presence of LVH (12, 13). Among these methods, deep learning (DL) techniques are superior to conventional machine learning techniques due to its ability of automatic feature extraction. The Convolutional Neural Network (CNN), combined with the Long Short-Term Memory (LSTM) model, appear to be the most useful architectures for classification (14). A 16-layer CNN-LSTM model was efficaciously used to classify coronary atherosclerotic disease (CAD), myocardial infarction, and chronic heart failure signals, with a precision rate of 98.5% (15). Our previous study also showed that the CNN-LSTM performed better than the CNN, LSTM, and doctors in detecting acute ST-segment elevation myocardial infarction (STEMI) based on 12-lead ECG, with an area under the curve (AUC) of 0.99 (16). Accordingly, our study aimed to establish a DL model based on the CNN-LSTM for reliable and rapid detection of LVH using 12-lead ECG.

Methods

Study population

A total of 3,120 patients hospitalized at the Third Affiliated Hospital of Sun Yat-sen University in China from January 2017 to December 2019 were recorded. Only the first admission for each patient was included; repeated hospitalizations were not evaluated in this study. Finally, 1,863 patients with ECG obtained within 1 week after hospitalization were included for analysis. Exclusion criteria were as follows: complete left or right bundle branch block, ventricular paced rhythm, ventricular arrhythmia at the time of ECG acquisition. Another independent cohort consisted of 453 patients was used as the internal test set 2 using the same inclusion and exclusion criteria. All personal details were erased to protect the confidentiality of patients' data. Data collection was approved by the ethics committee at the Third Affiliated Hospital of Sun Yat-sen University.

Baseline data collection

Data was extracted from the standard clinical electronic medical record (EMR) database of the Third Affiliated Hospital of Sun Yat-sen University, including demographic characteristics, comorbidities, laboratory tests, and medicines. The comorbidities were retrieved according to ICD-10 diagnostic codes.

Acquisition and procession of echocardiography data

Comprehensive 2-dimensional Doppler echocardiography, the gold standard to assess LVH, was routinely performed using commercially available ultrasound equipment. Acquisitions and measurements were performed by two experienced cardiac ultrasound doctors. LVH is defined as a left ventricular mass index (LVMI) >115 g/m^2 in male subjects and >95 g/m^2 in female subjects (17). Calculation of relative wall thickness (RWT) with the formula (2× posterior wall thickness)/(LV internal diameter at end-diastole), permits categorization of an increase in LV mass as either concentric (RWT $>$ 0.42) or eccentric (RWT $\leq$ 0.42) hypertrophy (17).

Acquisition and procession of ECG data

ECG was performed at a sampling rate of 1,000 Hz, and acquired in the supine position using the ECGNET Vision 3.0 (SanRui Electronic Technology, Guangdong, China). The ECG signal had to be clear, stable baseline with no interference. All ECG data were labeled with the study ID, and stored as XML file format following the H7L standard on a secure server. The quality of ECG data and ECG interpretations were independently reviewed by 2 cardiologists. The comparison of our model was referred to the Cornell voltage criteria and the Sokolow-Lyon voltage, given their relative higher sensitivity and specificity (9, 18). The sex-specific Cornell voltage criteria was computed as the amplitude of R in aVL plus the amplitude of S or QS complex in V3 (RaVL + SV3) with a cutoff of >2.8 mV in men and >2.2 mV in women. The Sokolow-Lyon voltage was obtained by adding the amplitude of S in V1 and the amplitude of R in V5 or V6 ≥3.5 mV (SV1 + RV5 or RV6) (19).

Deep-learning modeling

ECG data extraction

ECG data was extracted from XML files, consisted of 12 channels. The duration of ECG generally lasted from 10 to 90 s, and were cut into 5-s segment. The specification of each ECG segment was finally intercepted (5,000, 12), which was then utilized in the input model.

Data balance

There was imbalance in the quantity of cases and ECG segments between the control and LVH groups, as the latter group had less cases and ECG segments. To solve this problem, we drew sample cases and ECG segments of the control group referring to these of the LVH group, at last the cases and ECG segments were balanced in two groups.

The model was evaluated through 5-fold cross-validation technique. In each repetition of the cross-validation process, one part was selected as the validation set, another part was selected as the test set, while the remaining parts were served as the training set. Thus, the datasets of cases and ECG segments were needed to be equally split into 5 parts following below steps: (1) ECG segments of each case were ranked by number; (2) counted the frequency of the number of ECG segments; (3) if the ECG data of cases had the same quantity of segments and the number of those cases was more than 5, the ECG data of those five cases were selected and evenly divided into five parts, and then the remaining ECG data of cases were partitioned into five proximately equal parts, making the total number of cases and ECG segments among 5-fold subsets approximate. In order to split five equal parts rapidly, we developed an algorithm replacing manual processing with automation. The final dataset included 36,350 ECG segments ($n = 931$) and 36,348 ECG segments ($n = 932$) in the control and LVH groups, respectively. Previous studies have showed that the sensitivity and specificity of ECG criteria could be influenced by gender and left ventricular geometry, therefore we performed subgroup analyses. And we also balanced data for all subgroup analyses using the same method.

Model architecture and training

The architecture of CNN-LSTM model has been described in our previous study (16). In the training process, the model input was 12-lead ECG segment which had the specification of (5,000, 12). The first part of the CNN-LSTM model was CNN layers. The (5,000, 12) ECG segments were split into m smaller segments (length of smaller segments = 5,000/m) to train m CNN time Distributed layers simultaneously. The time Distributed layer is fully connected in the time dimension. In CNN time Distributed layer, weight parameters or convolution kernels were shared, instead of each have its own weight. We made the number of smaller segments (m) as a parameter with a value scope in (1, 2, 5, 10, 20, 25, 50, 100, 200, 250, and 500). The number of smaller segment (m) was settled according to the best validation output during training process. The number of CNN layers ranged from 1 to 5, and that of LSTM layers was 2. The CNN layer kernels would be selected from the scope of (16, 24, 32, 48, 56, and 64). All hyper-parameters would be Grid Search by keras tuner tool, which could automatically record and compare the accuracy of different models. Finally, the model with the best performance and corresponding hyper-parameters were selected, and then the parameters of the best model were utilized for LVH prediction based on ECG. Among all of the models explored, the CNN-LSTM model which had 200 smaller segments to input data, and contained 3 CNN layers (16 kernels in each layer), followed by 2 LSTM layers (200 LSTM units and 2 LSTM units in each LSTM layer, respectively) performed the best. The last 2 LSTM units output was predictive probability of the control and LVH groups. The DL models of each subgroup, including gender and RWT, were developed in the same process. More details of all models were showed in Supplementary Table 1.

Statistical analysis

The baseline characteristics were described as mean (standard deviation) (SD) or median with interquartile range (IQR) for continuous variables, and categorical variables were described as proportions. Kolmogorov-Smirnov test was used for continuous variables whether conforming to normal distribution. Differences in baseline characteristics were compared using *t*-test between two groups or analysis of variance (ANOVA) for continuous variables, while Mann-Whitney *U*-test between two groups and Kruskal-Wallis test among three groups was for abnormal distribution, and Chi-square test was applied for categorical variables. The statistical analyses were performed using the SPSS 22.0. A 2-tailed *P*-value < 0.05 was considered statistically significant.

Receiver operating characteristic (ROC) curve analysis and area under the curve (AUC) were used to evaluate the diagnostic efficacy of CNN-LSTM models and conventional ECG indexes of LVH. Delong's test was used to compare the performance of two ROC curves. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and F1-score were calculated with python (version 3.6.9).

Results

Study population and baseline characteristics

The study flow chart was shown in Figure 1. In the first cohort, 1,863 patients were divided into 2 groups according to the LVMI criteria: LVH group ($n = 932$) and control group ($n = 931$). Compared with the patients of control, patients in the LVH group were older, composed of higher proportion of female, more combined with hypertension, chronic heart failure (CHF), chronic kidney disease (CKD), as well as more likely to receive angiotensin-converting enzyme inhibitor (ACEI) and diuretics, but had lower level of hemoglobin (HGB). More details were shown in Table 1.

Further, patients in the first cohort were evenly split into 3 sets: the training set ($n = 1{,}120$), the validation set ($n = 371$) and the test set 1 ($n = 372$). Patients in the training set had higher prevalence of CAD. There were no significant difference in other clinical characteristics among the three sets. More baseline characteristics were summarized in Supplementary Table 2. In the internal test set 2, patients in the LVH group were older, composed of higher proportion of female and had higher prevalence of CHF, but had lower level of HGB (Supplementary Table 3).

The predictive value of DL models in LVH diagnosis

LVH was predicted by the CNN-LSTM model with an AUC of 0.62 (sensitivity 68%, specificity 57%) in the test set

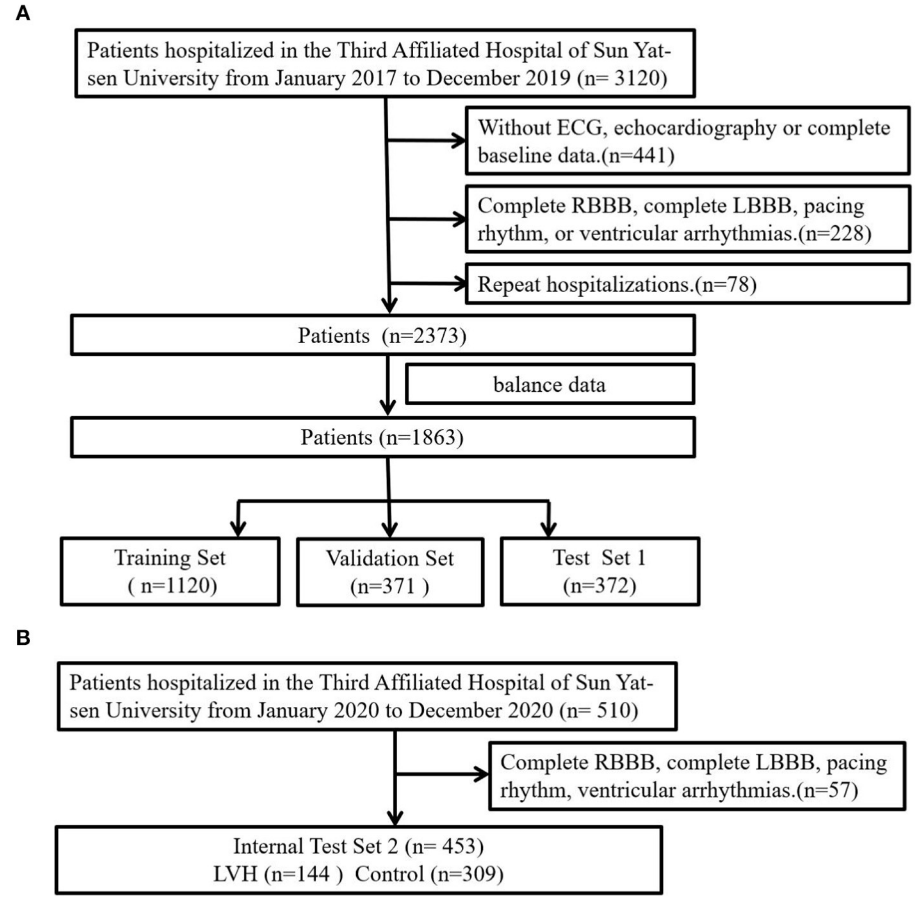

A
Patients hospitalized in the Third Affiliated Hospital of Sun Yat-sen University from January 2017 to December 2019 (n= 3120)
Without ECG, echocardiography or complete baseline data.(n=441)
Complete RBBB, complete LBBB, pacing rhythm, or ventricular arrhythmias.(n=228)
Repeat hospitalizations.(n=78)
Patients (n=2373)
balance data
Patients (n=1863)
Training Set (n=1120)
Validation Set (n=371)
Test Set 1 (n=372)
B
Patients hospitalized in the Third Affiliated Hospital of Sun Yat-sen University from January 2020 to December 2020 (n= 510)
Complete RBBB, complete LBBB, pacing rhythm, ventricular arrhythmias.(n=57)
Internal Test Set 2 (n= 453)
LVH (n=144) Control (n=309)

FIGURE 1
Study flow diagram. **(A)** The first cohort was further divided into training, validation and test sets. **(B)** The second cohort was used as internal test 2 to evaluate developed DL model.

TABLE 1 Patient characteristics between LVH and control groups.

Characteristics	LVH ($n = 932$)	Control ($n = 931$)	*P*-value
Demographic			
Female, *n* (%)	511 (54.8)	273 (29.3)	<0.001
Age, years	67.3 (10.5)	63.9 (11.3)	<0.001
Medical history			
CAD, *n* (%)	601 (64.5)	564 (60.6)	0.082
HT, *n* (%)	586 (62.9)	486 (52.2)	<0.001
CHF, *n* (%)	330 (35.4)	223 (24.0)	<0.001
DM, *n* (%)	306 (32.8)	333 (35.8)	0.182
Stroke, *n* (%)	129 (13.8)	122 (13.1)	0.641
CKD, *n* (%)	85 (9.1)	47 (5.0)	0.001
STEMI, *n* (%)	23 (2.5)	16 (1.7)	0.259
Laboratory examination			
HDL-C (mmol/L)	1.07 (0.29)	1.04 (0.27)	0.270
LDL-C (mmol/L)	2.77 (1.06)	2.82 (1.03)	1.030
HGB (g/L)	125.71 (19.06)	133.02 (18.74)	<0.001
PLT (10∧9/L)	230.52 (85.34)	228.64 (68.03)	0.607
BUN (mmol/L)	6.82 (4.58)	6.16 (2.70)	<0.001
Cr (umol/L)	103.92 (122.34)	87.81 (58.00)	<0.001
UA (umol/L)	394.94 (124.77)	394.92 (112.51)	0.997
potassium (mmol/L)	3.99 (0.45)	4.00 (0.40)	0.620
sodium (mmol/L)	141.58 (3.24)	141.39 (5.17)	0.767
ECG			
RV5 (mV)	1.49 (1.12, 1.99)	1.40 (1.09, 1.75)	<0.001
RV6 (mV)	1.20 (0.87, 1.58)	1.10 (0.86, 1.42)	<0.001
RaVL (mV)	0.42 (0.24, 0.63)	0.33 (0.17, 0.54)	<0.001
SV1 (mV)	−0.81 (−1.12 to −0.53)	−0.72 (−0.97 to −0.50)	<0.001
SV3 (mV)	−0.93 (−1.32 to −0.58)	−0.86 (−1.17 to −0.55)	0.002
Cornell voltage LVH, *n* (%)	414 (45.2)	240 (26.3)	<0.001
Sokolow-Lyon LVH, *n* (%)	109 (11.9)	18 (2.0)	<0.001
Echocardiography			
LVEF (%)	64.08 (9.35)	67.43 (5.17)	<0.001
LVEDD (mm)	49.23 (5.34)	44.6 (3.97)	<0.001
LVPW (mm)	10.44 (1.10)	9.58 (1.01)	<0.001
IVS (mm)	11.77 (1.68)	10.50 (1.34)	<0.001
LVMI (g/m^2)	129.28 (28.93)	89.97 (14.47)	<0.001
Concentric LVH, *n* (%)	515 (55.3)	532 (57.1)	0.412
Treatment			
ACEI, *n* (%)	192 (20.6)	113 (12.1)	<0.001
ARB, *n* (%)	264 (28.3)	234 (25.1)	0.120
Spirolactone, *n* (%)	134 (14.4)	107 (11.5)	0.064
CCB, *n* (%)	366 (39.3)	326 (35.0)	0.057
BB, *n* (%)	586 (62.9)	563 (60.5)	0.286
Diuretics, *n* (%)	240 (25.8)	186 (20.0)	0.003

CAD, coronary artery disease; HT, hypertension; DM, diabetes mellitus; CHF, chronic heart failure; CKD, chronic kidney disease; STEMI, ST-segment elevation myocardial infarction; HDL-C, high density lipoprotein cholesterol; LDL-C, low density lipoprotein cholesterol; HGB, hemoglobin; PLT, platelet; BUN, blood urea nitrogen; Cr, creatinine; UA, uric acid; LVEF, left ventricular ejection fraction; LVEDD, left ventricular end-diastolic dimension; LVPW, left ventricle posterior wall; IVS, ventricular septum; LVMI, left ventricular mass index; ACEI, angiotensin-converting enzyme inhibitor; ARB, angiotensin receptor blocker; CCB, calcium channel blocker; BB, beta-block.

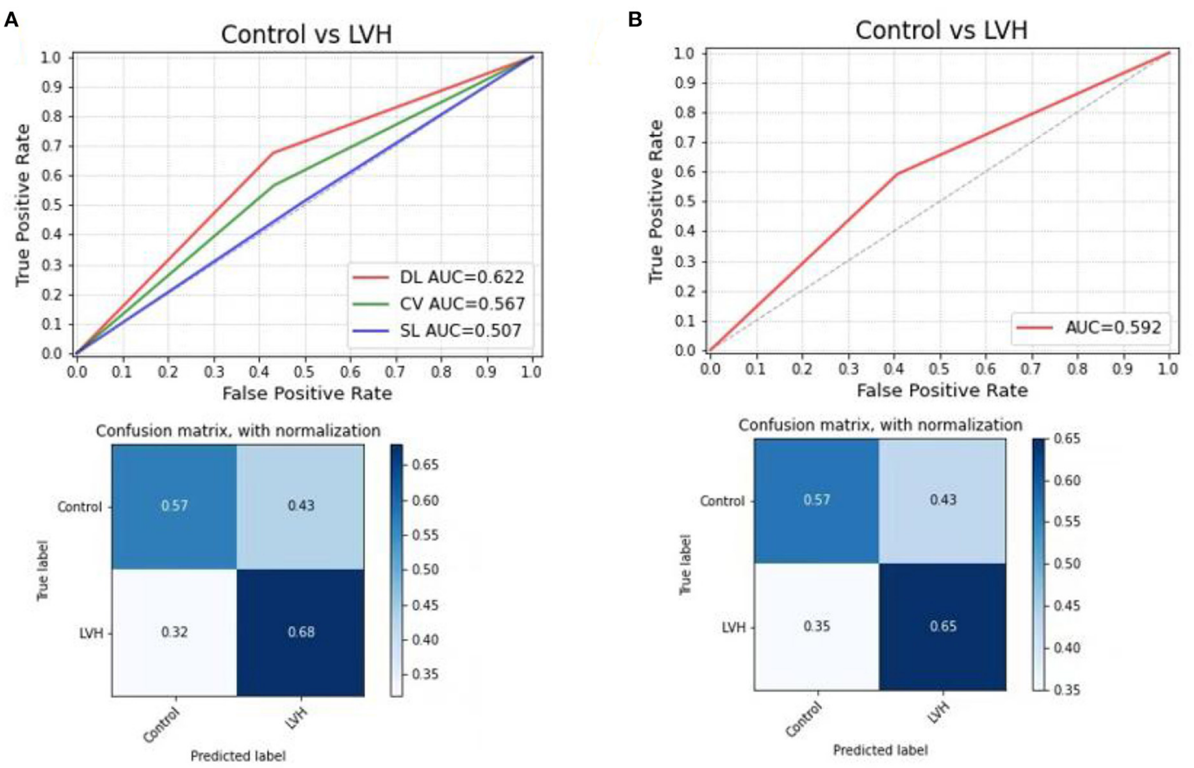


FIGURE 2
Receiver operating characteristic curve analysis, **(A)** compared the DL model with Cornell voltage and Sokolow-Lyon voltage in test set 1, the confusion matrix for predicting control and LVH using the DL model in the test set 1; **(B)** to test the DL model in internal test set 2. DL, deep learning model; CV, Cornell voltage, SL, Sokolow-Lyon voltage.

1, which had a better performance than the Cornell voltage criteria (AUC: 0.57, sensitivity 48%, specificity 72%) and the Sokolow-Lyon voltage (AUC: 0.51, sensitivity 14%, specificity 96%). Differences in ROC curves were statistically compared *via* Delong's test (CNN-LSTM model vs. Cornell voltage criteria, p-value = 0.075; CNN-LSTM model vs. Sokolow-Lyon, p-value = 0.037). Although no significant difference was found between CNN-LSTM model and Cornell voltage criteria, the sensitivity of CNN-LSTM model was higher than that of Cornell voltage criteria. In the internal test set 2, the CNN-LSTM model had a stable performance in predicting LVH with an AUC of 0.59 (sensitivity 65%, specificity 57%) (Figure 2), which was comparable to that of the internal test set 1.

In the subgroup analysis, the first step was to train different DL models according to gender. In the test sets, the CNN-LSTM model predicted LVH with an AUC of 0.66 (sensitivity 72%, specificity 60%) for male patients, which was better than that for female patients (AUC: 0.59, sensitivity 50%, specificity 71%) (Figure 3). The second step was to evaluate the effect of left ventricular geometry on the diagnosis of ventricular hypertrophy based on ECG. The DL models were trained for concentric and eccentric hypertrophy according to RWT. In the test sets, the CNN-LSTM model predicted concentric hypertrophy with an AUC of 0.66 (sensitivity 62%, specificity 70%) and eccentric hypertrophy with an AUC of 0.68 (sensitivity 65%, specificity 71%) in male patients, and an AUC of 0.58 (sensitivity 48%, specificity 68%) for concentric hypertrophy and an AUC of 0.58 (sensitivity 47%, specificity 69%) for eccentric hypertrophy in female patients (Figure 4) (Supplementary Table 4).

Discussion

This is a study to develop DL models of LVH diagnosis based on a large real-world ECG database. Our main achievement was that we built a DL model based on CNN-LSTM with higher sensitivity than current ECG diagnostic criteria. Moreover, we constructed different CNN-LSTM models to predict LVH for male and female patients separately, and the predictive value was better in male patients.

Our DL model predicted LVH with higher sensitivity than the Cornell voltage criteria and Sokolow-Lyon voltage (68, 48, and 14%, respectively), whereas its specificity was inferior to

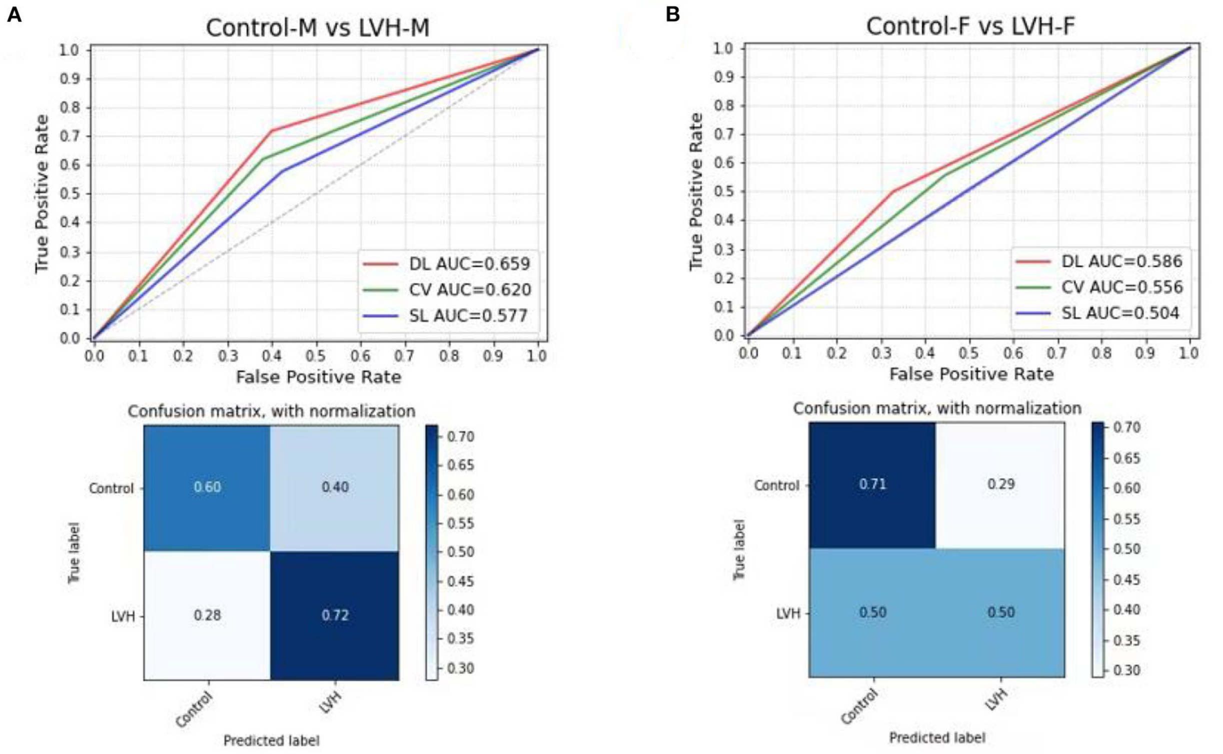


FIGURE 3
Comparing the DL model with Cornell voltage and Sokolow-Lyon voltage to predict LVH, the confusion matrix for predicting control and LVH using the DL model in the test set; **(A)** for male patients; **(B)** for female patients. DL, deep learning model; CV, Cornell voltage; SL, Sokolow-Lyon voltage.

these two criteria (57, 72, and 96%, respectively). The accuracy of our model still needed to be improved. In the study of Bressman et al., found that the sensitivity and specificity of ECG for left ventricular hypertrophy were 30.7 and 84.4% in a cohort of 13,960 subjects using a computer-generated algorithm, which is similar to the combination of the Sokolow-Lyon and Framingham criteria (20). Peguero et al. proposed a new ECG criteria involved measuring the amplitude of the deepest S wave (SD) in any single lead and adding it to the S wave amplitude of lead V4 (SV4), which outperformed Cornell voltage with a significantly higher sensitivity (62 vs. 35%) in a relatively small sample size (21). However, another study found that the Cornell voltage carried the best AUC of 0.678 (sensitivity 33.1%, specificity 88.8%), while Peguero Lo Presti criterion had an AUC of 0.64 (sensitivity 42.3%, specificity 75.8%) in a cohort of 2,134 patients (19). Current ECG criteria of LVH have low sensitivity, limit the application of ECG in screening for LVH. Recently, a few studies utilized machine learning techniques for ECG and clinical characteristics to diagnose LVH. Lin et al. used a support vector machine classifier as the machine learning method for 31 clinical characteristics and 28 ECG parameters to detect LVH, successfully achieving a specificity of 73.3%, and a much better sensitivity of 86.7%, compared to 3.3 and 52.7% of the Cornell and Sokolow-Lyon voltage criteria in a large sample of 2,196 males (12). Although this research developed a method with high accuracy in a large sample size, the patients included were only of younger males, and this model needed lots of clinical characteristics. Additionally, a machine-learning technique called Bayesian Additive Regression Trees was developed to predict LVH based on ECG and participant characteristics, and the result showed a specificity more than 93% but a poor sensitivity of only 29.0% in a cohort of 4,714 participants from the Multi-Ethnic Study of Atherosclerosis study (13). Khurshid et al. trained a CNN to predict cardiac magnetic resonance (CMR)-derived LV mass using 12-lead ECGs (LVM-AI) in the UK Biobank prospective cohort of 32,239 individuals. The results showed that the LVH discrimination of LVM-AI was 0.653 (sensitivity 34%, specificity 96%) and 0.621 (sensitivity 41%, specificity 83%) in the independent UK Biobank test set and Mass General Brigham, respectively. However, low sensitivity was still limiting the application of these models. On the other hand, the CNN-LSTM was able to detect CAD ECG signals with a diagnostic accuracy of 99.85% with blind-fold strategy (22). Our previous study showed the

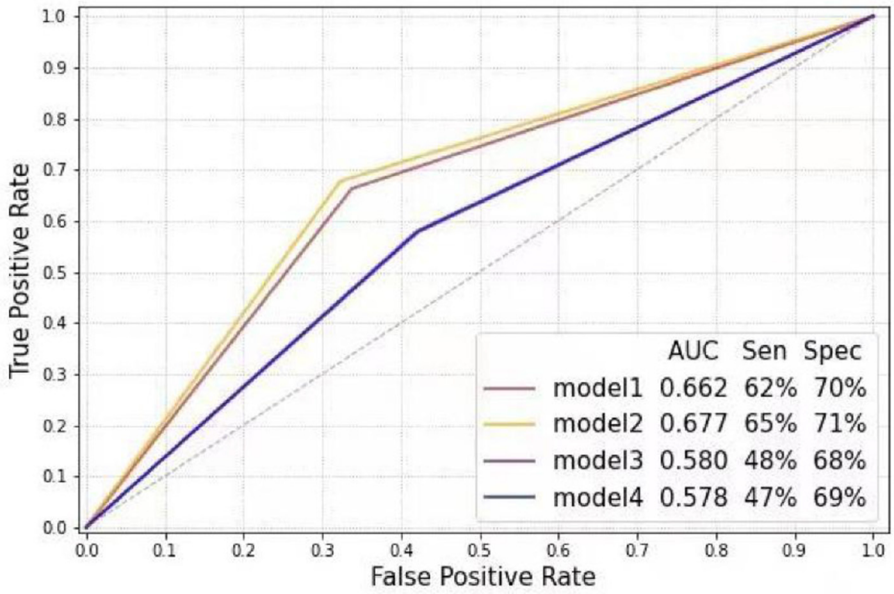


FIGURE 4
Receiver operating characteristic curve analysis of different models according to gender and relative wall thickness (Model 1: Control-M vs. concentric LVH-M; Model 2: Control-M vs. eccentric LVH-M; Model 3: Control-F vs. concentric LVH-F; Model 4: Control-F vs. eccentric LVH-F). LVH-F, female patients with left ventricular hypertrophy; LVH-M, male patients with left ventricular hypertrophy; Control-F, female patients in control group; Control-M, male patients in control group.

ECG DL diagnosis systems based on the CNN-LSTM have a good performance to detect STEMI and predict culprit vessel occlusion (16). On this basis, we developed a DL model of LVH diagnosis that showed higher sensitivity than current ECG criteria.

Moreover, previous work showed that female gender was associated with lower sensitivity but higher specificity (20, 23). Consistently, in our study, the LVH diagnosed by the DL model was lower in female patients (50% sensitivity, 71% specificity), compared to 72% of sensitivity and 60% of specificity for male patients. Additionally, left ventricular geometry is associated with ECG-defined left ventricular hypertrophy (24). An RWT > 0.42 demonstrated an increased sensitivity and decreased specificity for LVH (20). However, our models showed similar sensitivity to predict eccentric and concentric hypertrophy in female patients, and even higher sensitivity for eccentric hypertrophy in male patients.

There are some advantages in our DL models based on CNN-LSTM. First, the most common method of model training is to manually set a parameter, and the optimal value is selected after repeated experiments, which is inconvenient for clinical application. In our model training stage, the grid search method was used to search all possible parameter combinations of each model. For each parameter, grid search algorithm can extensively search the whole possible parameters space, and these parameters searching can be done in parallel, regardless of computing resource constraints, to reduce the training time. Moreover, our DL model did not need additional preprocessing for ECG data like removing noise, which may also perform well with different sources of ECG. However, the accuracy of our model still needed to be improved. Previous study showed multiple patient characteristics were associated with differences in sensitivity and specificity of LVH prediction by ECG. Therefore, adding the baseline characteristics like age, gender, body mass index, comorbidities to our model training may improve its performance. On the other hand, the attention module integrates channel information, obtains the importance of features and allocates attention weight to make the network pay attention to important features, so channel-wise attention could be added to different convolution layers in order to optimize the CNN-LSTM model. In terms of clinical application, our DL model was established in a real-world ECG database, in which all patients were included regardless of the admitting diagnosis. In addition, LVH is a modifiable risk factor related to systolic BP and regression of LVH may reduce subsequent CV events (25). Therefore, it might be helpful in the better management of hypertension. Besides, our CNN-LSTM model is an end-to-end approach, it only utilized raw ECG data input and built binary classification and multiclassification without experts or experienced cardiologists. It could be able to give primary diagnosis timely and reduce the workload of doctors.

Limitations

Some limitations of our study should be considered. Our study was a single-center study, the models may have the risk of generalizing poorly to other hospital systems and other datasets. Besides, our ECG diagnostic models based on CNN-LSTM have higher sensitivity at the expenses of relatively lower specificity compared to currently commonly used ECG diagnostic criteria. But ECG used as a screening tool, the interpretation method with higher sensitivity is more likely to identify more individuals with LVH who need confirmation of the diagnosis with echocardiography or MRI. Moreover, this study population mainly included south China population. Therefore, more researches from different regions and ethnic groups are necessary to confirm these findings.

Conclusion

Our ECG diagnostic model based on the CNN-LSTM has higher sensitivity than currently used ECG diagnostic criteria. The performance of the model trained for male patients was better than that for female patients. Therefore, this CNN-LSTM model may be a simple and effective screening tool of LVH in hypertensive patients and general population.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author/s.

Ethics statement

All traceable personal identifiers were removed from the analytic dataset to protect patients' privacy. The study protocol was approved by the Third Affiliated Hospital, Sun Yat-sen University ethics committee and the study was performed according to the declaration of Helsinki.

Author contributions

XZ, GH, and LW: research idea and study design. MW, J-RW, BZ, YeL, XY, SL, BT, and LL: data acquisition. DL and YoL: data analysis/interpretation. XZ and GH: statistical analysis. XQ, LH, and XX: supervision and mentorship. XQ, LH, XX, SQ, YC, and XH: writing guidance. Each author contributed important intellectual content during manuscript drafting or revision and accepts accountability for the overall work by ensuring that questions on the accuracy or integrity of any portion of the work are appropriately investigated and resolved. All authors read and approved the final version.

Funding

This study received funding from Guangdong Medical Research Foundation (A2019079), the National Natural Science Foundation of China (81770826 and 81370447), Science and Technology Planning Project of Guangdong Province (2016A050502014), National Key R&D Program (2018yfc1705105 and 2017YFA0105803), the 5010 Clinical Research Projects of Sun Yat-sen University (2015015), the Key Area R&D Program of Guangdong Province (2019B020227003), and the Science and Technology Plan Project of Guangzhou City (202007040003). The funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

Conflict of interest

Authors XZ, LW, XX, and XQ work in the Third Affiliated Hospital of Sun Yat-sen University, GH was employed by China Unicom (Guangdong) Industrial Internet Ltd., SQ works for China Unicom (Guangdong) Industrial Internet Ltd., LH works in Research Institute of Tsinghua, Pearl River Delta. Author J-RW was employed by LCFC (Hefei) Electronics Technology Co., Ltd., Hefei LCFC Information Technology Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcvm.2022.952089/full#supplementary-material

References

1. Cuspidi C, Sala C, Negri F, Mancia G, Morganti A. Prevalence of left-ventricular hypertrophy in hypertension: an updated review of echocardiographic studies. *J Human Hypert.* (2012) 26:343–9. doi: 10.1038/jhh.2011.104

2. Grajewski KG, Stojanovska J, Ibrahim EH, Sayyouh M, Attili A. Left ventricular hypertrophy: evaluation with cardiac MRI. *Curr Prob Diagn Radiol.* (2020) 49:460–75. doi: 10.1067/j.cpradiol.2019.09.005

3. Shenasa M, Shenasa H, El-Sherif N. Left ventricular hypertrophy and arrhythmogenesis. *Card Electrophysiol Clin.* (2015) 7:207–20. doi: 10.1016/j.ccep.2015.03.017

4. Cuspidi C, Vaccarella A, Negri F, Sala C. Resistant hypertension and left ventricular hypertrophy: an overview. *J Am Soc Hypert.* (2010) 4:319–24. doi: 10.1016/j.jash.2010.10.003

5. Cuspidi C, Rescaldani M, Sala C, Grassi G. Left-ventricular hypertrophy and obesity: a systematic review and meta-analysis of echocardiographic studies. *J Hypert.* (2014) 32:16–25. doi: 10.1097/HJH.0b013e328364fb58

6. Lewis AA, Ayers CR, Selvin E, Neeland I, Ballantyne CM, Nambi V, et al. Racial differences in malignant left ventricular hypertrophy and incidence of heart failure: a multicohort study. *Circulation.* (2020) 141:957–67. doi: 10.1161/CIRCULATIONAHA.119.043628

7. Cao X, Broughton ST, Waits GS, Nguyen T, Li Y, Soliman EZ. Interrelations between hypertension and electrocardiographic left ventricular hypertrophy and their associations with cardiovascular mortality. *Am J Cardiol.* (2019) 123:274–83. doi: 10.1016/j.amjcard.2018.10.006

8. Ruilope LM, Schmieder RE. Left ventricular hypertrophy and clinical outcomes in hypertensive patients. *Am J Hypert.* (2008) 21:500–8. doi: 10.1038/ajh.2008.16

9. Alfakih K, Walters K, Jones T, Ridgway J, Hall AS, Sivananthan M. New gender-specific partition values for ECG criteria of left ventricular hypertrophy: recalibration against cardiac MRI. *Hypertension.* (2004) 44:175–9. doi: 10.1161/01.HYP.0000135249.66192.30

10. Bacharova L, Ugander M. Left ventricular hypertrophy: the relationship between the electrocardiogram and cardiovascular magnetic resonance imaging. *Ann Noninv Electrocardiol.* (2014) 19:524–33. doi: 10.1111/anec.12223

11. Nagarajan VD, Lee SL, Robertus JL, Nienaber CA, Trayanova NA, Ernst S. Artificial intelligence in the diagnosis and management of arrhythmias. *Eur Heart J.* (2021) 42:3904–16. doi: 10.1093/eurheartj/ehab544

12. Lin GM, Liu K. An electrocardiographic system with anthropometrics via machine learning to screen left ventricular hypertrophy among young adults. *IEEE J Transl Eng Health Med.* (2020) 8:1800111. doi: 10.1109/JTEHM.2020.2990073

13. Sparapani R, Dabbouseh NM, Gutterman D, Zhang J, Chen H, Bluemke DA, et al. Detection of left ventricular hypertrophy using bayesian additive regression trees: the MESA. *J Am Heart Assoc.* (2019) 8:e009959. doi: 10.1161/JAHA.118.009959

14. Somani S, Russak AJ, Richter F, Zhao S, Vaid A, Chaudhry F, et al. Deep learning and the electrocardiogram: review of the current state-of-the-art. *Europace.* (2021) 23:1179–91. doi: 10.1093/europace/euaa377

15. Lih OS, Jahmunah V, San TR, Ciaccio EJ, Yamakawa T, Tanabe M, et al. Comprehensive electrocardiographic diagnosis based on deep learning. *Artific Int Med.* (2020) 103:101789. doi: 10.1016/j.artmed.2019.101789

16. Wu L, Huang G, Yu X, Ye M, Liu L, Ling Y, et al. Deep learning networks accurately detect st-segment elevation myocardial infarction and culprit vessel. *Front Cardiovasc Med.* (2022) 9:797207. doi: 10.3389/fcvm.2022.797207

17. Lang RM, Badano LP, Mor-Avi V, Afilalo J, Armstrong A, Ernande L, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American society of echocardiography and the European association of, cardiovascular imaging. *Eur Heart J Cardiovasc Imag.* (2016) 17:412. doi: 10.1093/ehjci/jew041

18. Hancock EW, Deal BJ, Mirvis DM, Okin P, Kligfield P, Gettes LS. AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: part V: electrocardiogram changes associated with cardiac chamber hypertrophy: a scientific statement from the American heart association electrocardiography and arrhythmias committee, council on clinical cardiology; the American college of cardiology foundation; and the heart rhythm society: endorsed by the international society for computerized electrocardiology. *Circulation.* (2009) 119:e251–61. doi: 10.1161/CIRCULATIONAHA.108.191097

19. Ricciardi D, Vetta G, Nenna A, Picarelli F, Creta A, Segreti A, et al. Current diagnostic ECG criteria for left ventricular hypertrophy: is it time to change paradigm in the analysis of data? *J Cardiovasc Med.* (2020) 21:128–33. doi: 10.2459/JCM.0000000000000907

20. Bressman M, Mazori AY, Shulman E, Chudow JJ, Goldberg Y, Fisher JD, et al. Determination of sensitivity and specificity of electrocardiography for left ventricular hypertrophy in a large, diverse patient population. *Am J Med.* (2020) 133:e495–500. doi: 10.1016/j.amjmed.2020.01.042

21. Peguero JG, Lo Presti S, Perez J, Issa O, Brenes JC, Tolentino A. Electrocardiographic criteria for the diagnosis of left ventricular hypertrophy. *J Am Coll Cardiol.* (2017) 69:1694–703. doi: 10.1016/j.jacc.2017.01.037

22. Tan JH, Hagiwara Y, Pang W, Lim I, Oh SL, Adam M, et al. Application of stacked convolutional and long short-term memory network for accurate identification of CAD ECG signals. *Comp Biol Med.* (2018) 94:19–26. doi: 10.1016/j.compbiomed.2017.12.023

23. Colossimo AP, Costa Fde A, Riera AR, Bombig MT, Lima VC, Fonseca FA, et al. Electrocardiogram sensitivity in left ventricular hypertrophy according to gender and cardiac mass. *Arquiv Brasil Cardiol.* (2011) 97:225–31. doi: 10.1590/S0066-782X20110050000850

24. Tomita S, Ueno H, Takata M, Yasumoto K, Tomoda F, Inoue H. Relationship between electrocardiographic voltage and geometric patterns of left ventricular hypertrophy in patients with essential hypertension. *Hypert Res.* (1998) 21:259–66. doi: 10.1291/hypres.21.259

25. Brooks JE, Soliman EZ, Upadhya B. Is left ventricular hypertrophy a valid therapeutic target? *Curr Hypert Rep.* (2019) 21:47. doi: 10.1007/s11906-019-0952-9

frontiers | Frontiers in Cardiovascular Medicine

TYPE Original Research
PUBLISHED 22 August 2022
DOI 10.3389/fcvm.2022.903660

OPEN ACCESS

EDITED BY
Tim Leiner,
Mayo Clinic, United States

REVIEWED BY
Kenya Kusunose,
Tokushima University Hospital, Japan
Evelyne Yehudit Bischof,
Shanghai University of Medicine
and Health Sciences, China
Bas Kietselaer,
Zuyderland Medical Centre,
Netherlands

*CORRESPONDENCE
Kunlun He
kunlunhe@plagh.org

†These authors have contributed
equally to this work

SPECIALTY SECTION
This article was submitted to
Cardiovascular Imaging,
a section of the journal
Frontiers in Cardiovascular Medicine

RECEIVED 24 March 2022
ACCEPTED 04 August 2022
PUBLISHED 22 August 2022

CITATION
Lin X, Yang F, Chen Y, Chen X, Wang W, Chen X, Wang Q, Zhang L, Guo H, Liu B, Yu L, Pu H, Zhang P, Wu Z, Li X, Burkhoff D and He K (2022) Echocardiography-based AI detection of regional wall motion abnormalities and quantification of cardiac function in myocardial infarction. *Front. Cardiovasc. Med.* 9:903660. doi: 10.3389/fcvm.2022.903660

Echocardiography-based AI detection of regional wall motion abnormalities and quantification of cardiac function in myocardial infarction

Xixiang Lin[1,2†], Feifei Yang[1†], Yixin Chen[3], Xiaotian Chen[3], Wenjun Wang[1], Xu Chen[1,2], Qiushuang Wang[4], Liwei Zhang[4], Huayuan Guo[1], Bohan Liu[1], Liheng Yu[1], Haitao Pu[3], Peifang Zhang[3], Zhenzhou Wu[3], Xin Li[5], Daniel Burkhoff[6] and Kunlun He[1*]

[1]Medical Big Data Center, Chinese PLA General Hospital, Beijing, China, [2]Medical School of Chinese PLA, Beijing, China, [3]BioMind Technology, Beijing, China, [4]Fourth Medical Center of PLA General Hospital, Beijing, China, [5]Sixth Medical Center of PLA General Hospital, Beijing, China, [6]Cardiovascular Research Foundation, New York, NY, United States

Objective: To compare the performance of a newly developed deep learning (DL) framework for automatic detection of regional wall motion abnormalities (RWMAs) for patients presenting with the suspicion of myocardial infarction from echocardiograms obtained with portable bedside equipment versus standard equipment.

Background: Bedside echocardiography is increasingly used by emergency department setting for rapid triage of patients presenting with chest pain. However, compared to images obtained with standard equipment, lower image quality from bedside equipment can lead to improper diagnosis. To overcome these limitations, we developed an automatic workflow to process echocardiograms, including view selection, segmentation, detection of RWMAs and quantification of cardiac function that was trained and validated on image obtained from bedside and standard equipment.

Methods: We collected 4,142 examinations from one hospital as training and internal testing dataset and 2,811 examinations from other hospital as the external test dataset. For data pre-processing, we adopted DL model to automatically recognize three apical views and segment the left ventricle. Detection of RWMAs was achieved with 3D convolutional neural networks (CNN). Finally, DL model automatically measured the size of cardiac chambers and left ventricular ejection fraction.

Results: The view selection model identified the three apical views with an average accuracy of 96%. The segmentation model provided good agreement with manual segmentation, achieving an average Dice of 0.89. In the internal test dataset, the model detected RWMAs with AUC of 0.91 and 0.88 respectively for standard and bedside ultrasound. In the external test dataset, the AUC were 0.90 and 0.85. The automatic cardiac function measurements agreed with echocardiographic report values (e. g., mean bias is 4% for left ventricular ejection fraction).

Conclusion: We present a fully automated echocardiography pipeline applicable to both standard and bedside ultrasound with various functions, including view selection, quality control, segmentation, detection of the region of wall motion abnormalities and quantification of cardiac function.

KEYWORDS

artificial intelligence - AI, myocardial infarction, echocardiography, deep learning, bedside ultrasound

Introduction

Myocardial infarction (MI) is the most severe manifestation of coronary heart disease, resulting in disability or sudden cardiac death. According to Report on Cardiovascular Health and Diseases in China 2021, AMI mortality increased by a factor 3.5 in rural areas and by a factor of 2.66 in urban areas from 2002 to 2019. In 2019, AMI mortality was 0.08% in rural areas and 0.06% in urban areas (1). Recent studies show that there is significant variability in the care and outcomes of MI patients in hospitals with different levels of care (2). Rapid diagnosis and prompt reperfusion treatment are of primary importance to reduce mortality from MI.

With the advantages of easy availability, low cost, fast performance and safety, transthoracic echocardiography (especially bedside ultrasound) is the most commonly used non-invasive imaging tool for detecting regional wall motion abnormalities (RWMAs) and providing information on short- and long-term outcomes after acute myocardial infarction (AMI) (3–5). The American College of Cardiology/American Heart Association and the European Heart Association guidelines give a Class I recommendation for using transthoracic echocardiography to detect RWMAs in chest pain patients presenting to the emergency ward without delaying angiography (6, 7).

However, accurate recognition of RWMAs by echocardiography requires highly trained and experienced physicians which are in short supply and typically not available around the clock in many hospitals. Furthermore, visual diagnosis of RWMAs often varies amongst doctors with various level of expertise (8). Therefore, an effective solution for efficient, accurate and objective diagnosis of RWMAs is needed.

Deep learning (DL) models have strong data processing capabilities and have been used for automated interpretation of images obtained from various modalities. Related to echocardiography, DL models can perform a variety of analyses, such as image quality assessment, view classification, boundary segmentation, and disease diagnosis (9–14). With the help of DL, tedious and time-consuming tasks like segmentation and quantification of different parameters can be performed quickly and precisely, saving increasingly scarce human resources (11, 13, 14).

Recently, tremendous advances have been made in DL models for the detection of RWMAs (15, 16). However, these studies applied relatively strict, up front image quality criteria such that ~40% of studies were excluded from analysis, indicating that those models may not be practical for widespread use. Furthermore, in those studies, standard echocardiographic equipment was used which, in general, produce higher quality images than newer, portable bedside ultrasound equipment. With the advantages of portability and availability, bedside ultrasound is becoming increasingly applied in emergency rooms and intensive care units for specific applications, such as real-time assessment of cardiac function and RWMA in patients presenting with chest pain syndromes. Use of DL models to analyze images from these machines has not been specifically explored in prior studies (9, 10).

We developed a novel DL model to analyze echocardiographic videos to detect RWMAs and standard

Abbreviations: A4C, apical four chambers; A2C, apical two chambers; ALX, apical long axis; CNN, convolutional neural networks; DL, deep learning; LVEF, left ventricular ejection fraction; LVEDV, left ventricular end-diastolic volume; LVESV, left ventricular end-systolic volume; LV EDTD, left ventricular end-diastolic transversal dimension; LA ESTD, left atrial end-systolic transversal dimension; LOA, limits of agreement; MI, myocardial infarction; RV EDTD, right ventricular end-diastolic transversal dimension; RA ESTD, right atrial end-systolic transversal dimension; RWMAs, regional wall motion abnormalities; ROC, receiver operating characteristic.

indexes of cardiac size and function from three standardized apical views. In contrast to prior studies, the structure of our model and the training dataset were geared toward analysis of images from bedside echocardiograms while fully retaining the ability to analyze images from standard equipment. Accordingly, the main purpose of this study was to compare the accuracy of this model for analyzing videos obtained from bedside ultrasound to those of standard equipment.

Materials and methods

Study population

The methods used in the design, implementation, and reporting of this study were consistent with the recently published PRIME (Proposed Requirements for Cardiovascular Imaging Related Machine Learning Evaluation) checklist (17), which was provided in the **Supplementary Appendix**. We retrospectively accessed a total of 2,274 transthoracic echocardiographic examinations obtained between May 2015 and September 2019 from the Fourth Medical Center of Chinese PLA General Hospital as our training and validation dataset (ratio 8:2). MI and control cases were matched for age and sex. We then prospectively collected 1,868 examinations between May 2020 and May 2021 from the same hospital as an internal test dataset. For the external test dataset, we collected 3,026 examinations between Jan 2021 and Dec 2021 from the Second Affiliated Hospital of Shandong University of Traditional Chinese Medicine. Training and testing datasets each included echocardiographic studies from standard and bedside echocardiographic equipment as detailed in the **Figure 1**. The diagnosis of acute or prior MI were based on information from the electronic medical records (including echocardiographic report, blood tests, ECGs and angiograms). The presence and extent of RWMAs were extracted from echocardiographic reports generated by experienced sonographers and the echocardiograms were reviewed a second time by an experienced cardiologists who authorized the final diagnoses.

Echocardiography

Each echocardiographic examination was acquired through standard methods. Videos from three standard apical views were include in this study: apical 4-chamber (A4C), apical 2-chamber (A2C), and apical long axis (ALX). Images were acquired from a diverse array of standard echocardiography machine manufacturers including Phillips EPIQ 7C and iE-elite with S5-1 and X5-1 transducers (Phillips, Andover, MA, United States), and Vivid E95 (General Electric, Fairfield, CT, United States) and portable bedside machines including Philips CX50 and Mindray M9cv with transducer SP5-1s (Mindray, Shenzhen, Guangdong, China). All images were stored with a standard Digital Imaging and Communication in Medicine (DICOM) format according to the instructions from each manufacturer.

View selection and quality control

We labeled 33,404 images to develop a method to classify 29 standard views and then selected the three apical views required for the subsequent analysis. View selection was performed using a Xception Net neural network model according to methods that were previously described (18).

An automated algorithm was developed to assess image quality and exclude images whose quality were insufficient for analysis. Expert echocardiographers manually labeled 2,837 A4C, 1,880 A2C, and 1,910 ALX images as qualified or unqualified. These labeled images were then used to build the AI model. Examples of qualified and unqualified images as assigned by the model are shown in **Supplementary Figure 1**. As seen in these examples, the contour of left ventricle was ambiguous so that the endocardial or epicardial border was rarely identified in

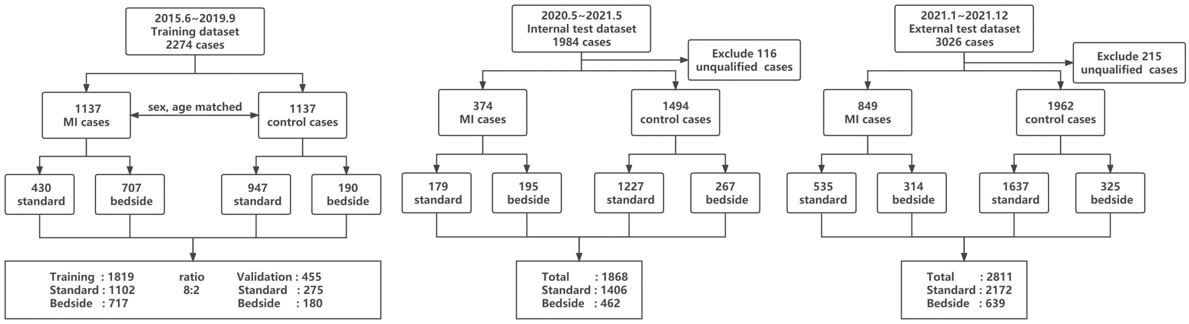


FIGURE 1
Summary of number of echocardiograms used in this study.

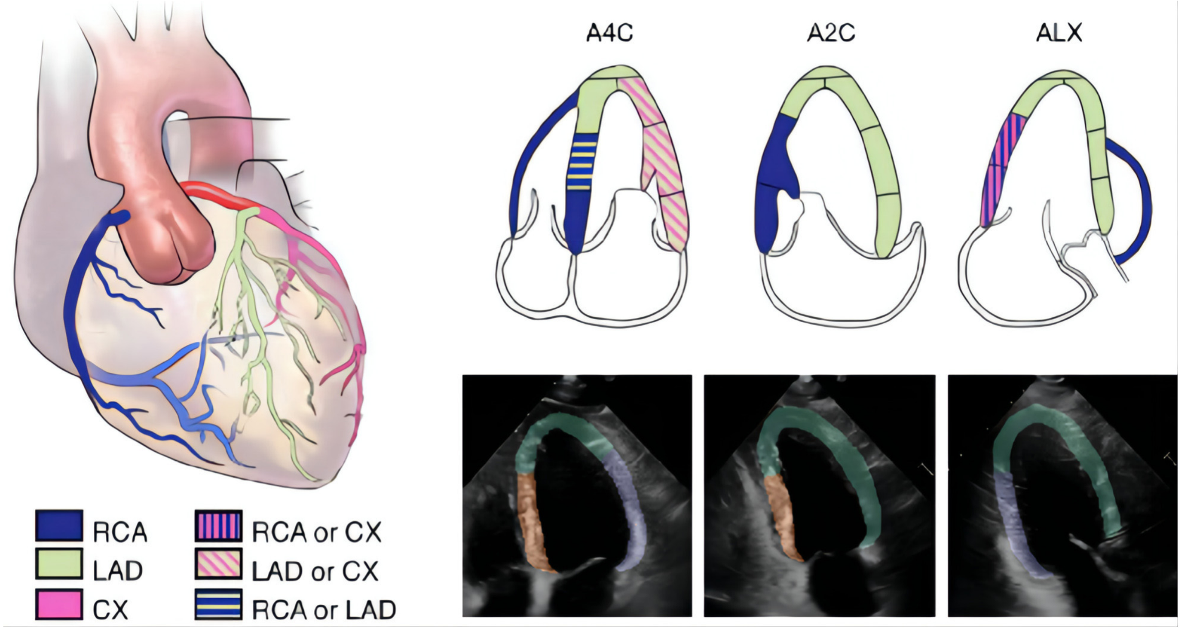


FIGURE 2
The segmentation of different wall regions. The 2015 ASE guideline recommend typical distributions of the coronary artery in apical four-chamber (A4C), apical two-chamber (A2C), and apical long-axis (ALX) views. In the echocardiographic images, we labeled A for apical, anterior and anteroseptal walls (green area), F for inferior and inferoseptal walls (orange area), and L for anterolateral and inferolateral walls (purple area).

unqualified views. Subsequently, images automatically classified as unqualified were excluded from analysis.

Segmentation

The segmentation model was developed to outline the endocardial and epicardial borders of the left ventricle and the endocardial borders of the left atrium, right atrium and right ventricle. Left ventricle was grouped into 3 different regions, designated A (apical, anterior, and anteroseptal walls), F (inferior and inferoseptal walls) and L (anterolateral and inferolateral walls) according to 2015 American Society of Echocardiography guidelines (19) (**Figure 2**). We annotated 493 apical 4-chamber videos (8,555 frames), 332 apical 2-chamber videos (5,768 frames) and 366 apical long-axis videos (6,389 frames) which served as ground truth for developing and testing this algorithm to segmented the heart into regions A, F and L as detailed above (19, 20). Myocardial segmentation masks were generated for every frame of each video with the pretrained segmentation LSTM-Unet (21–23). Three separate segmentation models with the same structure were developed to analyze the A4C, A2C, and ALX views. For the A4C video, the model segmented the left ventricle, the left atrium, the right ventricle and the right atrium which were used to quantify the size of each chamber. For detection of RWMAs, each video frame was cropped into a 128 × 128 pixels square with the left ventricle at the center and pixel values are normalized to the range from 0 to 1 (**Figure 2**).

Detection of regional wall motion abnormalities

The overall process for detecting RWMAs was summarized in **Figure 3**. Each original DICOM video was concatenated with the mask of the myocardium obtained by the segmentation model. The mask and video were then input into A, F, and L classification models. Detection of the presence of RWMAs and the territory of RWMAs was achieved with Deep 3D Convolution Neural Network.

Details of the RWMA detection model are shown in **Supplementary Figure 2**. The backbone of the model is R2plus1D, which is a time-saved and calculation-saved feature extractor. In order to effectively use the information extracted by the R2plus1D feature extractor, three fully connected layers are added to the model (24). There is a Batch Normalization layer, an activation layer (LeakyReLU) and a 50% dropout layer following each full-connected layer. Batch Normalization can improve the efficiency of model training, which can save time required to train the video model (25). The output of the

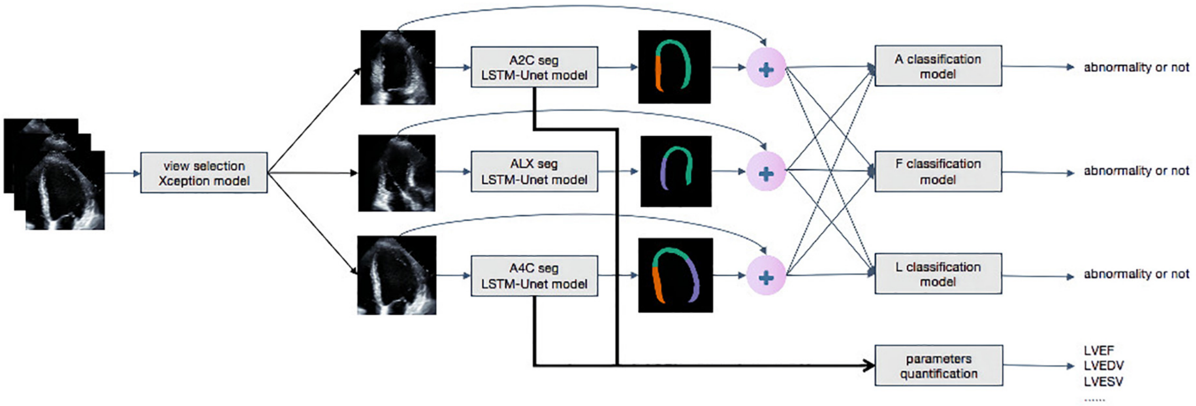


FIGURE 3
The whole work flow of deep learning model. Steps of data processing. The first model achieves view selection on echocardiography. The Xception model generates a confidence level for view selection and selects A4C, A2C, and ALX views whose confidence is higher than 0.9. Secondly, LSTM-Unet segments each frames of outputs of Xception. The segment and the original video are concatenated as inputs of classification models to detect regional wall motion abnormality. The outputs of LSTM-Unet with A4C and A2C are calculated important parameters, such as LVEDV, LVESV, and LVEF.

RWMA detection model contains two values (two red neurons as shown in **Supplementary Figure 2**). These values (the score of no abnormality and the score of an abnormality) are derived from the full-connected layers transform information extracted by R2plus1D.

The RWMAs detection models were trained using two Graphics Processing Units (GPU), NVIDIA Tesla P100. Each model contains about 1 million parameters. All the parameters are trained in the direction of minimizing cross entropy, which is an error function to calculate how far the models' outputs is from real label. The models are trained with Stochastic Gradient Descent Momentum (SGDM) with 0.9 momentum and 1e−4 weight decay. The learning rate starts from 1e−5 and increases linearly with epoch until 1e−4 at epoch 10, which is called warm-up (26). Then learning rate decline linearly from 1e−4 at epoch 10 to 5e−5 at epoch 50.

In order to improve the generalization of RWMAs detection models, spatiotemporal video augmentation methods are adopted. The left subplot in **Supplementary Figure 3** shows the clipping in the time dimension and the right subplot shows the spatial cropping. At the training state, each video is randomly cropped and clipped in the spatial and temporal dimensions, so as to enhance the diversity of data and the generalization of the model. In the test and validation phase, the videos are divided into non-overlapping 8-frame video segments and each video segment is inferred three times with three spatial crops. For example, a 32-frame video is divided into 4 video segments, each of which contain 8 frames, and each video segment generates three crops, which means the 32-frame sample generate 12 results in total. Majority voting combines 12 results. The models are performed with Python 3.6.8 and PyTorch 1.4.0. The code will be released in GitHub.

Quantification of key metrics

The key metrics derived from the model include left ventricle ejection fraction (LV EF), end-diastolic volume (LV EDV), end-systolic volume (LV ESV), end-diastolic transversal dimension (LV EDTD), left atrial end-systolic transversal dimension (LA ESTD), right ventricular end-diastolic transversal dimension (RV EDTD), and right atrial end-systolic transversal dimension (RA ESTD). We calculated these metrics based on the output of segmentation model and the 2015 guidelines of the American Society of Echocardiography and the European Association of Echocardiography (19). In order to enhance the interpretability of deep learning, we adopted the segmentation model to segment the area of four chambers, and then used Simpson biplane method to calculate LVEDV, LVESV, and EF. The long short-term memory (LSTM) can effectively extract the time information from the video.

Statistical analysis

Analyses were performed using algorithms written in Python 3.6 from the libraries of Numpy, Pandas, and Scikit-learn. Continuous variables were expressed as mean ± standard deviation, median and interquartile range, or counts and percentage, as appropriate. Comparisons of reports and machine algorithm performances were performed using one-way analysis of variance (ANOVA), followed by the least significant difference (LSD) *t*-test. The detection models were assessed according to the area under the receiver operating characteristic (AUROC) curves which plotted sensitivity versus 1−specificity derived from the model's prediction confidence score. Results were regarded as statistically significant when $P < 0.05$. All calculations were performed by using IBM SPSS version 23.0.

Result

Study population

For the internal training and validation dataset, a total of 2,274 transthoracic echocardiographic examinations were divided between standard and bedside ultrasound. As specified in the echocardiography and clinical reports, MIs and RWMAs were present in 1,137 of the 2,274 studies (50%), 62% of which were from bedside ultrasound. In the internal test dataset, MIs and RWMAs were present in 374 of 1,868 cases (20%), 52% of which were examined by bedside ultrasound. In the external test dataset, MIs and RWMAs were present in 849 of 3,026 cases (28%), 37% of which were examined by bedside ultrasound. The clinical and echocardiographic characteristics of the included populations are summarized in Table 1 (training dataset) and Supplementary Table 1 (internal and external test datasets). As expected, significant differences in baseline characteristics existed between normal subjects and those with a myocardial infarction.

View selection and segmentation

As summarized in Supplementary Figure 4, the deep-learning architecture identified the apical 4-chamber, 2-chamber and long-axis views with a high degree of accuracy: 94, 99, and 95%, respectively. The quality control model achieved an average 95% consistency compared with expert in identifying qualified images (Supplementary Figure 1). As for segmentation, the model provided good agreement with manual segmentation with an average Dice of 0.89 (Table 2). Although the performance of the model for segmenting bedside ultrasound images was slightly lower than that in standard ultrasound, our model was applicable with both machines.

Detection of regional wall motion abnormalities

For the detection of a regional wall motion abnormalities in the internal test dataset, the deep learning model had an average AUROC of 0.91 for images obtained with standard echocardiographic equipment compared to 0.88 for images obtained with beside equipment. Youden's Index was used to evaluate model performance, which yielded sensitivities of 85.4% vs. 85.2% and specificities of 83.2% vs. 78.2% for standard versus beside equipment, respectively. In the external test dataset, the model achieved an average AUROC of 0.90 vs. 0.85 for standard versus bedside ultrasound, with corresponding sensitivities of 81.6% vs. 78.3% and specificities of 83.7% vs. 78.1%. The model had a similar performance for detecting anterior, inferior and lateral wall motion abnormalities in both bedside and standard ultrasound (Figure 4 and Table 3). Overall, these results corresponded to comparable accuracies in detecting RWMAs in the three territories: 0.83 for anterior, 0.81 for inferior and 0.85 for lateral walls.

To test the advantages of this tool for experts and beginners, we randomly selected 100 cases from both MI and control cases captured from standard and bedside equipment. In total, 3 experts and 5 beginners participated in the test, where the first reads were based on their own judgments, while they had access to the AI results for the second reads. The second

TABLE 1 Baseline characteristics of the training and validation dataset.

	Training and validation dataset			
	Standard		Bedside	
	MI	Normal	MI	Normal
Echo number	430	947	707	190
Age	65 (55,73)	60 (53,76)	67 (54,77)	58 (50,66)
Male patients(%)	353 (83.3)	590 (62.3)	460 (65.2)	118 (62.1)
Comorbidities (%)				
Hypertension	115 (35.1)	249 (26.3)	270 (42.4)	37 (19.5)
Hyperlipidemia	217 (66.2)	148 (14.6)	326 (51.3)	10 (5.3)
Diabetes	124 (38.0)	103 (10.9)	324 (50.8)	21 (11.1)
Renal insufficiency	65 (17.4)	79 (8.3)	228 (35.8)	6 (3.2)
Ischemic stroke history	53 (17.4)	96 (10.1)	121 (21.5)	17 (8.9)
Echo parameters				
LV EF (%)	46 (41,54) *	62 (60,64)	43 (36,48) †	60 (59,62)
LV EDV (mm^2)	119 (98,144) *	87 (80,100)	106 (88,129) †	84 (75,98)
LV ESV (mm^2)	62 (48,82) *	33 (30,37)	59 (46, 76) †	33 (30,39)
LV EDTD (mm)	49 (45,53) *	43 (41,45)	47 (43,51) †	42 (40,45)
LA ESTD (mm)	40 (38,43) *	36 (34,38)	41 (38,43) †	36 (34,38)
RV EDTD (mm)	32 (30,34) *	30 (29,32)	31 (29,33) †	30 (28,32)
RA ESTD (mm)	32 (30,34) *	30 (28,32)	32 (29,34) †	30 (28,31)
Territories of RWMAs				
Multiple walls	168 (39.1)		319 (45.1)	
A	291 (67.7)		529 (74.8)	
F	220 (51.2)		363 (51.3)	
L	154 (35.8)		268 (37.9)	

Values are median (IQR) or n (%). *$p < 0.05$ vs. normal subjects in standard group. †$p < 0.05$ vs. normal subjects in bedside group. BMI, Body Mass Index; LVEF, left ventricular ejection fraction; LVEDV, left ventricular end-diastolic volume; LVESV, left ventricular end-systolic volume; LV EDTD, left ventricular end-diastolic transversal dimension; LA ESTD, left atrial end-systolic transversal dimension; RV EDTD, right ventricular end-diastolic transversal dimension; RA ESTD, right atrial end-systolic transversal dimension; MI, myocardial infarction; RWMAs, regional wall motion abnormalities; A, apical, anterior and anteroseptal walls; F, inferior and inferoseptal walls; L, anterolateral and inferolateral walls.

TABLE 2 Performance of the segmentation model.

	Segmentation (Dice)									
	LV endocardium		LV myocardium		LA endo		RV endo		RA endo	
	Standard	Bedside	Standard	Bedside	Standard	Bedside	Standard	Bedside	Standard	Bedside
A4C	0.94	0.95	0.84	0.81	0.94	0.93	0.89	0.90	0.94	0.93
A2C	0.93	0.93	0.79	0.77	0.93	0.91				
ALX	0.93	0.92	0.82	0.78	0.93	0.93				

A4C, apical 4-chamber; A2C, apical 2-chamber; ALX, apical long axis; LV, left ventricle; LA, left atrium; RA, right atrium; RV, right ventricle; A4C, apical 4-chamber; A2C, apical 2-chamber; endo, endocardium.

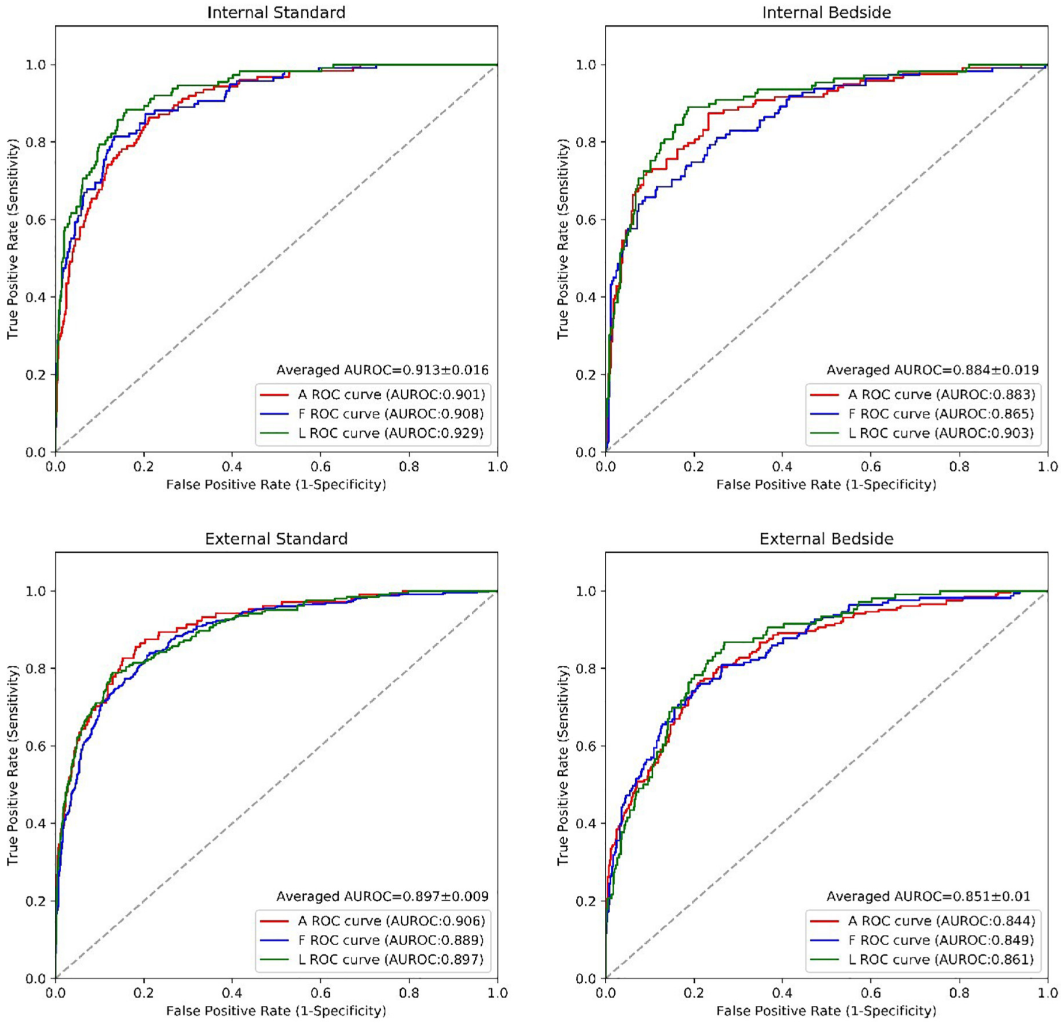


FIGURE 4
The performance of the RWMAs detection model. The performance of the RWMAs detection model for bedside vs. standard cases in retrospective invalidation dataset and prospective testing dataset. Abbreviations as in Figure 2.

reads were performed at a separate time without access to the results of the first read. The comparison of results of the first and second reads are summarized in Supplementary Figure 5 and Supplementary Table 2. The AI models did not significantly improve the accuracy of experts, but was very helpful for beginners, with average accuracy improving by 9.8, 7.4, and 12.8% for the A, F and L territories, respectively (Supplementary Table 2).

Quantification of metrics of chamber sizes and function

The output of our segmentation model was used to compute chamber dimensions and ejection fraction based on the biplane method of disks summation (modified Simpson's rule) (19). As above, this analysis was performed on studies which passed the automated quality control algorithm. Results of the Bland–Altman analysis comparing parameter values provided by the AI algorithm and from the clinical reports are summarized in Table 4. For each of the 7 parameters, the mean bias and LOAs were similar for the analysis performed on studies obtained with standard and bedside equipment. Accordingly, data are further summarized in the Bland–Altman plots in Figure 5 (LVEF) and Supplementary Figure 6 (structural parameters) in which results obtained from standard and bedside equipment are pooled.

We further extended our analysis to segregate patients into 3, clinically meaningful discrete LVEF groups: reduced (<40%), midrange (40–50%) and preserved (>50%). The results of this prediction were moderately consistent with that of echocardiographic reports, with an accuracy of 77% (Supplementary Table 3). There was a tendency to underestimate LVEF in our model, especially in higher manual values. Overall, these results indicate that the degree of accuracy of the automatic quantification of these key metrics was within the bounds of normal clinical practice.

Discussion

Prompt recognition of RWMAs by echocardiography is an important tool for timely diagnosis and treatment of myocardial infarction in patients presenting with chest pain, especially in the emergency department. However, accurate diagnosis relies on technical expertise in image acquisition, intrinsic quality of the imaging equipment, and significant experience in image interpretation. Technological advances in portable echocardiographic equipment are making high quality imaging more readily available. However, availability of appropriately trained physicians for on-demand interpretation is limited in most hospitals and analyses performed by less experienced physicians may lead to misdiagnoses which can adversely impact clinical care. Our tool provides a fully automated pipeline for all routine aspects of interpreting echocardiograms. For example, echocardiographic images obtained from a patient admitted to the emergency department with chest pain can be submitted electronically to the model which automatically assess for the presence of RWMAs and also quantifies cardiac function, enabling high-efficient serial primary care. In non-emergent settings, this tool can be used to assess temporal changes of regional and global heart function during repeated echocardiographic videos in patients during follow up for a myocardial infarction, for monitoring cardiotoxicity during chemotherapy and in patients receiving cardiac rehabilitation. Our model makes analysis of these echocardiograms less burdensome to the system while maintaining (or even enhancing) reliability and reproducibility.

Our study is the first to rigorously demonstrate that deep learning methods can automatically assess image quality and interpret RWMAs with a high degree of accuracy and to provide a comparison of results from standard and portable echocardiographic equipment. The first steps in our pipeline involve automated view selection, quality control and image segmentation. Each of these steps was performed with a high degree of accuracy. The importance of automated image quality assessment cannot be overstated. In order to mimic clinical practice, we did not apply any initial screening of image quality for inclusion since physicians are also faced with images of varied quality. Our algorithm excluded unqualified images from which detection of RWMAs would be inappropriate, even by experienced clinicians (see Supplementary Figure 1 for examples). Interestingly, 2.7% of standard cases and 14.7% of portable bedside cases were excluded. As summarized Supplementary Figure 7, when the deep learning model was applied to detect regional wall motion abnormalities in these unqualified images, the AUCs, sensitivities and specificities were all markedly decreased and the bias and LOAs for each of the 7 parameters of chamber sizes and function were significantly larger. In our cohort, most portable bedside studies were obtained in the emergency room in patients presenting with chest pain. Thus, the higher rate of exclusion of bedside cases

TABLE 3 Performance of model for identifying the presence and territories of RWMAs.

	Internal test dataset		External test dataset	
	Standard	Bedside	Standard	Bedside
AUC				
A	0.901	0.883	0.906	0.844
F	0.908	0.865	0.889	0.849
L	0.929	0.903	0.897	0.861
Average	0.913	0.884	0.897	0.851
Sensitivity				
A	86.3%	87.4%	82.7%	76.80%
F	81.4%	79.3%	83.3%	76.10%
L	88.4%	89.0%	78.9%	82.10%
Average	85.4%	85.2%	81.6%	78.30%
Specificity				
A	78.8%	76.8%	84.9%	78.7%
F	86.7%	76.4%	79.3%	78.8%
L	84.0%	81.3%	87.0%	76.9%
Average	83.2%	78.2%	83.7%	78.1%

A, apical, anterior and anteroseptal walls; F, inferior and inferoseptal walls; L, anterolateral and inferolateral walls.

may reflect factors such as the critical nature of the patients under which images are obtained and perhaps less availability of experienced sonographers in this setting. Under such urgent conditions, less consideration may be given to image quality. Therefore, availability of an AI model that can provide feedback in real time can promote acquisition of high-quality images and ensure that measurements and detection of RWMAs are based on qualified images.

Another important feature of our model that analyzed RWMAs was that it focused analysis on the left ventricle by excluding the other cardiac chambers. As such, the segmentation model (which achieved an average Intersection Over Union value of 80.9%) was able to divide the left ventricle into three regions corresponding to coronary artery perfusion territories (15, 16). This division was based on the current guidelines and, in addition to its intrinsic clinical utility, could provide the foundation for subsequent research. The model exhibited a high performance with average Intersection Over Union value of 80.9%, but relatively lower for epicardium due to the obscure borders near the edge of imaging area.

Wall motion abnormality detection and classification

Overall, the deep learning model exhibited good performance with similar accuracies for detecting RWMAs in all 3 regions of the left ventricle in both internal and

TABLE 4a The measurements of the corresponding clinical metrics for the RWMAs made by physicians and predicted by AI in internal test dataset.

Parameters	Equipment	Median value from clinical report (IQR)	Bland–Altman analysis (Physicians vs. AI)		
			Bias	Upper LOA	Lower LOA
LV EF	Standard	60 (59,62)	4.0	15	−11
	Bedside	58 (51,60)	4.7	15	−9
LV EDV	Standard	92 (81,108)	6.0	50	−40
	Bedside	85 (77,101)	6.4	45	−39
LV ESV	Standard	36 (31,43)	−1.1	19	−23
	Bedside	35 (31,47)	−1.2	21	−30
LV EDTD	Standard	44 (42,47)	0.8	8.0	−5.9
	Bedside	42 (38,46)	1.5	11	−6.2
LA ESTD	Standard	38 (35,40)	2.6	14	−7.5
	Bedside	36 (31,41)	2.7	15	−8.0
RV EDTD	Standard	31 (29,33)	−0.9	8.1	−9.5
	Bedside	31 (29,33)	0.9	10	−8.4
RA ESTD	Standard	31 (29,33)	0.5	11	−9.0
	Bedside	32 (29,33)	1.5	11	−10

TABLE 4b The measurements of the corresponding clinical metrics for the RWMAs made by physicians and predicted by AI in external test dataset.

Parameters	Equipment	Median value from clinical report (IQR)	Bland–Altman analysis (Physicians vs. AI)		
			Bias	Upper LOA	Lower LOA
LV EF	Standard	59 (51,63)	3.4	17	−7.7
	Bedside	47 (37,58)	4.6	16	−4.1
LV EDV	Standard	103 (95,114)	14	41	−20
	Bedside	108 (90,136)	4.7	58	−43
LV ESV	Standard	59 (55,62)	6.6	16	−12
	Bedside	55 (39,83)	−2.2	18	−24
LV EDTD	Standard	48 (45,50)	1.9	12	−6.4
	Bedside	48 (43,53)	−1.1	7.0	−10
LA ESTD	Standard	36 (32,39)	−1.4	10	−12
	Bedside	40 (36,44)	0.5	12	−14
RV EDTD	Standard	35 (32,38)	1.8	11	−7.6
	Bedside	35 (21,38)	1.2	8.9	−7.5
RA ESTD	Standard	35 (32,38)	1.4	9.8	−5.9
	Bedside	35 (31,39)	−0.1	13	−9.7

LVEF, left ventricular ejection fraction; LVEDV, left ventricular end-diastolic volume; LVESV, left ventricular end-systolic volume; LV EDTD, left ventricular end-diastolic transversal dimension; LA ESTD, left atrial end-systolic transversal dimension; RV EDTD, right ventricular end-diastolic transversal dimension; RA ESTD, right atrial end-systolic transversal dimension; IQR, interquartile range; LOA, limits of agreement.

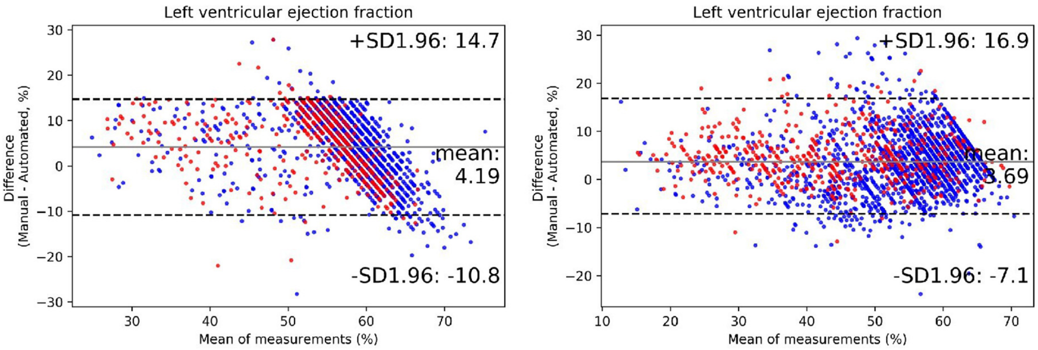


FIGURE 5
The performance of the automated quantification model. Bland–Altman plots of left ventricular ejection fraction in repeated measurements using the exact same video clips of internal **(left plot)** and external **(right plot)** testing dataset. The red dots represent cases acquired from portable bedside ultrasound; the blue dots represent cases acquired from standard ultrasound. The black lines represent limits of agreement.

external test datasets (**Table 3**). Performance in the external test dataset was only slightly lower (by ∼3%) than in the external test dataset. Also importantly, results achieved from images obtained with the bedside ultrasound were comparable to those of the standard equipment with the difference of average AUC between equipment of only 0.04 in internal and external test datasets (**Figure 4**). Our primary motivation is that the model will assist, not replace, physician decision making. Therefore, the AI model will save experts' time without influence his or her judgment and proved to be very helpful for beginners, with average accuracy improving by 9.8, 7.4, and 12.8% respectively for the A, F, and L territories. With the advantage of objectivity and consistency, the AI model may become an educational tool for beginners to improve their skill in image acquisition and interpretation.

We also analyzed the incorrect cases of each models using Logistic regression. After multivariable adjustment, correlation between the accuracy and age was statistically significant in all models (**Supplementary Table 4**). The violin plot showed that the average age of incorrect cases was older than that of correct cases (**Supplementary Figure 8**). This finding is consistent with our experience in clinical practice. Because the degree of wall motion in older patients was generally lower than young patients, which makes it more difficult for models to distinguish MI and normal cases.

Automatic quantification of cardiac function

In addition to detection of territories with RWMAs, patient care is influenced by parameters of cardiac size and function. Accordingly, our model also automatically and reliably quantified the relevant parameters derived from end-diastolic and end-systolic images. Since it is intended that our deep learning model be used in conjunction with bedside echocardiographic devices without ECG capabilities, end-diastolic and end-systolic images need to be selected based on endocardial areas determined from the segmentation model; this approach is similar to those employed by Zhang et al. (10) and Ouyang et al. (9). Finally, the deep learning model was reasonably consistent with physicians' classification of reduced (<40%), midrange (40∼50%) and preserved (>50%) LVEFs, which has important implications for treatment and prognosis of patients with heart failure (3, 5).

Related work

Automated detection of RWMAs have been described in two recent studies. Kusunose et al constructed a deep learning model that utilized 3 mid-level short-axis static images acquired at the end-diastolic, mid-systolic, and end-systolic phases to detect the presence and territories of RWMAs (15). The highest AUC produced by the model is 0.97, which is similar to the AUC by cardiologist and significantly higher than the AUC by resident readers. However, the pipeline was semiautomatic in that the initial input requires manual selection for echocardiogram views and cycle phases. As such, that model was based on analysis of static images, which does not parallel how RWMAs are detected in clinical practice which rely on dynamic videos. Finally, the study lacked external test dataset.

Huang et al also developed a deep learning model for detection of RWMA that directly analyzed dynamic videos, first by performing automated view selection and segmentation (16). The AUC for the external dataset was 0.89. However, the dataset of RWMAs was relatively small (*n* = 576) and 84% (*n* = 486) of studies included RWMAs of multiple walls; thus the ability to detect cases with single wall RWMAs was not fully evaluated. In contracts, nearly 50% of cases in our study have single wall RWAMs. In addition, to meet the stringent quality control, the author excluded up to a third of the examinations despite only using standard echocardiographic equipment. This model is therefore not suitable for widespread use in clinical practice, especially for analysis of bedside ultrasound in the emergency department. In contrast, our quality control algorithm was based on assessments by expert echocardiographers and therefore more closely mimicked clinical practice. Accordingly, our algorithm excluded only 2.7% of studies from standard equipment and 14.7% of studies from portable bedside studies. Despite having excluded a smaller percent of cases, our overall model performed was comparable to that of this prior study. Thus, our fully automatic pipeline can be applied to both standard and bedside ultrasound for detection of RWMAs and measurement of cardiac function, even in the emergency department.

Study limitations

The results of our study need to be considered within the context of several limitations. First, the distribution of echocardiography machines differed between MI and control cases, because the MI cases are more likely to have been performed with portable bedside equipment in an intensive care or emergency department, while control cases were mainly obtained by standard equipment in dedicated ultrasound rooms. Second, like other deep learning studies, we face the "black box problem" related to unexplained model features and how they contribute to the final result. To limit this problem to some degree, we removed irrelevant areas (e.g., RV free wall region) during the segmentation process so that the model focused on LV myocardium. Third, our model detected the presence and RWMAs rather than the severity of motion abnormalities, because wall motion score indexes recommended by society guidelines have interobserver and intra-observer variabilities. Instead, we developed an automated model to quantify cardiac function in real time. Although our model achieved good performance in the external test set, testing of the pipeline in a prospective RCT cohort is warranted.

Conclusion

We developed and validated a fully automated echocardiography pipeline applicable to both standard and portable bedside ultrasound with various functions, including view selection, quality control, segmentation, detection of the region of wall motion abnormalities and quantification of cardiac function. With high levels of sensitivity and specificity, the model has the potential to be used as a screening tool to aid physician in identifying patients with RWMAs, particularly in through the use of portable bedside ultrasound in the emergency room and intensive care units.

Data availability statement

The original contributions presented in this study are included in the article/**Supplementary material**, further inquiries can be directed to the corresponding author.

Author contributions

KH, XXL, and FY designed this study, analyzed and interpreted the patient data, drafted the manuscript, and guaranteed that all aspects of the work was investigated and resolved. DB and KH critically revised the important intellectual content of the manuscript. WW, XuC, QW, LZ, HG, BL, LY, and XL collected and analyzed and interpreted the data. YC, XiC, HP, PZ, and ZW designed the network architecture and performed the data preparation and analysis. All authors read and approved the final manuscript.

Funding

KH received support from Ministry of Industry and Information Technology of China (2020-0103-3-1). FY received support from the Beijing Natural Science and Technology Foundation (7202198).

Conflict of interest

YC, XiC, HP, PZ, and ZW were employees of BioMind. DB was a consultant to BioMind.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcvm.2022.903660/full#supplementary-material

References

1. Shengshou H. Report on cardiovascular health and diseases in China 2021: an updated summary. *Chin Circ J.* (2022) 37:26.

2. Xu H, Yang Y, Wang C, Yang J, Li W, Zhang X, et al. Association of hospital-level differences in care with outcomes among patients with acute st-segment elevation myocardial infarction in China. *JAMA Netw Open.* (2020) 3:e2021677. doi: 10.1001/jamanetworkopen.2020.21677

3. Vogel B, Claessen BE, Arnold SV, Chan D, Cohen DJ, Giannitsis E, et al. St-segment elevation myocardial infarction. *Nat Rev Dis Primers.* (2019) 5:39. doi: 10.1038/s41572-019-0090-3

4. Prastaro M, Pirozzi E, Gaibazzi N, Paolillo S, Santoro C, Savarese G, et al. Expert review on the prognostic role of echocardiography after acute myocardial infarction. *J Am Soc Echocardiogr.* (2017) 30:431–43.e2. doi: 10.1016/j.echo.2017.01.020

5. Prasad SB, Lin AK, Guppy-Coles KB, Stanton T, Krishnasamy R, Whalley GA, et al. Diastolic dysfunction assessed using contemporary guidelines and prognosis following myocardial infarction. *J Am Soc Echocardiogr.* (2018) 31:1127–36. doi: 10.1016/j.echo.2018.05.016

6. Amsterdam EA, Wenger NK, Brindis RG, Casey DE Jr., Ganiats TG, Holmes DR Jr., et al. 2014 Aha/Acc guideline for the management of patients with non-st-elevation acute coronary syndromes: a report of the American college of cardiology/American heart association task force on practice guidelines. *J Am Coll Cardiol.* (2014) 64:e139–228. doi: 10.1016/j.jacc.2014.09.017

7. Ibanez B, James S, Agewall S, Antunes MJ, Bucciarelli-Ducci C, Bueno H, et al. 2017 esc guidelines for the management of acute myocardial infarction in patients presenting with st-segment elevation: the task force for the management of acute myocardial infarction in patients presenting with st-segment elevation of the European Society of Cardiology (Esc). *Eur Heart J.* (2018) 39:119–77. doi: 10.1093/eurheartj/ehx393

8. Parisi AF, Moynihan PF, Folland ED, Feldman CL. Quantitative detection of regional left ventricular contraction abnormalities by two-dimensional echocardiography. Ii. Accuracy in coronary artery disease. *Circulation.* (1981) 63:761–7. doi: 10.1161/01.cir.63.4.761

9. Ouyang D, He B, Ghorbani A, Yuan N, Ebinger J, Langlotz CP, et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature.* (2020) 580:252–6. doi: 10.1038/s41586-020-2145-8

10. Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, et al. Fully automated echocardiogram interpretation in clinical practice. *Circulation.* (2018) 138:1623–35. doi: 10.1161/CIRCULATIONAHA.118.034338

11. Narula S, Shameer K, Salem Omar AM, Dudley JT, Sengupta PP. Machine-learning algorithms to automate morphological and functional assessments in 2d echocardiography. *J Am Coll Cardiol.* (2016) 68:2287–95. doi: 10.1016/j.jacc.2016.08.062

12. Kusunose K, Haga A, Inoue M, Fukuda D, Yamada H, Sata M. Clinically feasible and accurate view classification of echocardiographic images using deep learning. *Biomolecules.* (2020) 10:665. doi: 10.3390/biom10050665

13. Silva JF, Silva JM, Guerra A, Matos S, Guerra A, Costa C. Ejection fraction classification in transthoracic echocardiography using a deep learning approach. *Proceedings of the 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*. Karlstad: (2018).

14. Yue Z, Li W, Jing J, Yu J, Yi S, Yan W. Automatic Segmentation of the Epicardium and Endocardium Using Convolutional Neural Network. *Proceedings of theIEEE International Conference on Signal Processing*. Chengdu: IEEE (2017).

15. Kusunose K, Abe T, Haga A, Fukuda D, Yamada H, Harada M, et al. A deep learning approach for assessment of regional wall motion abnormality from echocardiographic images. *JACC Cardiovasc Imaging.* (2020) 13(2 Pt 1):374–81. doi: 10.1016/j.jcmg.2019.02.024

16. Huang MS, Wang CS, Chiang JH, Liu PY, Tsai WC. Automated recognition of regional wall motion abnormalities through deep neural network interpretation of transthoracic echocardiography. *Circulation.* (2020) 142:1510–20. doi: 10.1161/CIRCULATIONAHA.120.047530

17. Sengupta PP, Shrestha S, Berthon B, Messas E, Donal E, Tison GH, et al. Proposed requirements for cardiovascular imaging-related machine learning evaluation (Prime): a checklist: reviewed by the American college of cardiology healthcare innovation council. *JACC Cardiovasc Imaging.* (2020) 13:2017–35. doi: 10.1016/j.jcmg.2020.07.015

18. Yang F, Chen X, Lin X, Chen X, Wang W, Liu B, et al. Automated analysis of doppler echocardiographic videos as a screening tool for valvular heart diseases. *JACC Cardiovasc Imaging.* (2021) 15:551–63. doi: 10.1016/j.jcmg.2021.08.015

19. Lang RM, Badano LP, Mor-Avi V, Afilalo J, Armstrong A, Ernande L, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American society of echocardiography and the European association of cardiovascular imaging. *J Am Soc Echocardiogr.* (2015) 28:1–39.e14. doi: 10.1016/j.echo.2014.10.003

20. Voigt JU, Pedrizzetti G, Lysyansky P, Marwick TH, Houle H, Baumann R, et al. Definitions for a common standard for 2d speckle tracking echocardiography: consensus document of the eacvi/ase/industry task force to standardize deformation imaging. *Eur Heart J Cardiovasc Imaging.* (2015) 16:1–11. doi: 10.1093/ehjci/jeu184

21. Falk T, Mai D, Bensch R, Çiçek Ö, Abdulkadir A, Marrakchi Y, et al. U-Net: deep learning for cell counting, detection, and morphometry. *Nat Methods.* (2019) 16:67–70. doi: 10.1038/s41592-018-0261-2

22. Ibtehaz N, Rahman MS. Multiresunet : rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Netw.* (2020) 121:74–87. doi: 10.1016/j.neunet.2019.08.025

23. Ronneberger O, Fischer P, Brox T. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. Cham: Springer (2015).

24. Du T, Wang H, Torresani L, Ray J, Lecun Y, Paluri M. A closer look at spatiotemporal convolutions for action recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE (2018).

25. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. *Proceedings of the 32nd International Conference on Machine Learning*. Lille: (2015). doi: 10.1007/s11390-020-0679-8

26. Goyal P, Dollár P, Girshick R, Noordhuis P, Wesolowski L, Kyrola A, et al. Accurate, large minibatch sgd: training imagenet in 1 hour. *arXiv[Preprint].* (2017). Available online at: https://doi.org/10.48550/arXiv.1706.02677 (accessed June 8, 2017).

frontiers | Frontiers in Cardiovascular Medicine

TYPE Original Research
PUBLISHED 12 September 2022
DOI 10.3389/fcvm.2022.981901

Check for updates

OPEN ACCESS

EDITED BY
Federico Caobelli,
University Hospital of Basel,
Switzerland

REVIEWED BY
Alexander Van Rosendael,
Leiden University Medical Center
(LUMC), Netherlands
Uxio Hermida,
Faculty of Life Sciences and Medicine,
King's College London, United
Kingdom
Hieu Nguyen,
Johns Hopkins University, United
States

*CORRESPONDENCE
Zhiwei Zhai
z.zhai@amsterdamumc.nl

SPECIALTY SECTION
This article was submitted to
Cardiovascular Imaging,
a section of the journal
Frontiers in Cardiovascular Medicine

RECEIVED 29 June 2022
ACCEPTED 11 August 2022
PUBLISHED 12 September 2022

CITATION
Zhai Z, van Velzen SGM, Lessmann N,
Planken N, Leiner T and Išgum I (2022)
Learning coronary artery calcium
scoring in coronary CTA from
non-contrast CT using unsupervised
domain adaptation.
Front. Cardiovasc. Med. 9:981901.
doi: 10.3389/fcvm.2022.981901

Learning coronary artery calcium scoring in coronary CTA from non-contrast CT using unsupervised domain adaptation

Zhiwei Zhai[1,2]*, Sanne G. M. van Velzen[1,2,3],
Nikolas Lessmann[4], Nils Planken[5], Tim Leiner[6,7] and
Ivana Išgum[1,2,3,5]

[1]Department of Biomedical Engineering and Physics, Amsterdam University Medical Center, Location University of Amsterdam, Amsterdam, Netherlands, [2]Faculty of Science, Informatics Institute, University of Amsterdam, Amsterdam, Netherlands, [3]Amsterdam Cardiovascular Sciences, Heart Failure and Arrhythmias, Amsterdam, Netherlands, [4]Diagnostic Image Analysis Group, Radboud University Medical Center Nijmegen, Nijmegen, Netherlands, [5]Department of Radiology and Nuclear Medicine, Amsterdam University Medical Center, Location University of Amsterdam, Amsterdam, Netherlands, [6]Department of Radiology, Utrecht University Medical Center, University of Utrecht, Utrecht, Netherlands, [7]Department of Radiology, Mayo Clinic, Rochester, MN, United States

Deep learning methods have demonstrated the ability to perform accurate coronary artery calcium (CAC) scoring. However, these methods require large and representative training data hampering applicability to diverse CT scans showing the heart and the coronary arteries. Training methods that accurately score CAC in cross-domain settings remains challenging. To address this, we present an unsupervised domain adaptation method that learns to perform CAC scoring in coronary CT angiography (CCTA) from non-contrast CT (NCCT). To address the domain shift between NCCT (source) domain and CCTA (target) domain, feature distributions are aligned between two domains using adversarial learning. A CAC scoring convolutional neural network is divided into a feature generator that maps input images to features in the latent space and a classifier that estimates predictions from the extracted features. For adversarial learning, a discriminator is used to distinguish the features between source and target domains. Hence, the feature generator aims to extract features with aligned distributions to fool the discriminator. The network is trained with adversarial loss as the objective function and a classification loss on the source domain as a constraint for adversarial learning. In the experiments, three data sets were used. The network is trained with 1,687 labeled chest NCCT scans from the National Lung Screening Trial. Furthermore, 200 labeled cardiac NCCT scans and 200 unlabeled CCTA scans were used to train the generator and the discriminator for unsupervised domain adaptation. Finally, a data set containing 313 manually labeled CCTA scans was used for testing. Directly applying the CAC scoring network trained on NCCT to CCTA led to a sensitivity of 0.41 and an average false positive volume 140 mm^3/scan. The proposed method improved the sensitivity to 0.80 and reduced average false positive volume of 20 mm^3/scan. The results indicate that the unsupervised domain adaptation approach enables automatic CAC scoring in contrast enhanced CT while learning from a large and diverse set of CT scans without contrast. This may allow for better utilization of existing annotated data sets and extend the applicability of automatic CAC scoring to contrast-enhanced CT scans without the need for additional manual annotations. The code is publicly available at https://github.com/qurAI-amsterdam/CACscoringUsingDomainAdaptation.

KEYWORDS

coronary artery calcium scoring, unsupervised domain adaptation, convolutional neural network (CNN), coronary CTA, adversarial learning

1. Introduction

In recent years, deep neural networks have achieved impressive performance on various medical image analysis tasks (1, 2). This success is highly associated with the use of large amounts of representative annotated training data. However, the dependence on such data sets limits the applicability of already trained and well-performing networks to non-representative data sampled from a different distribution, such as images acquired at different sites, on different scanners, and by different acquisition protocols. Hence, generalizing deep neural networks trained on specific data to test data originating from a different domain remains a major challenge.

The domain shift, i.e., differences in data distributions and types of data between training and test domains, can be addressed by unsupervised domain adaptation methods that transfer a model that was trained on the source domain in a supervised manner to the target domain where no labels are available (3, 4). The common idea of unsupervised domain adaptation methods is to align features extracted by a network between two domains, aiming to generate similar feature distributions for both domains (4, 5). To achieve this, an adversarial learning strategy can be used. In this case, the generator network is optimized to extract features with similar distribution for the two domains while the discriminator network is trained to distinguish features from these domains (6).

Several works have investigated methods for unsupervised approaches to domain shift problem for segmentation of cardiac images (7–10). Dou et al. (8) proposed an unsupervised adversarial domain adaptation network to transfer cardiac segmentation network between MRI and CT. In this work the feature distributions of source and target domains were aligned at multiple scales. Chen et al. (7) extended the work of Dou et al. by aligning the domains in both image and feature perspectives. This method was evaluated with cardiac segmentation and abdominal multi-organ segmentation between MRI and CT. Wu et al. (10) presented an unsupervised domain adaptation framework to adapt cardiac segmentation between MRI and CT. In this method, a novel distance metric was proposed to calculate the misalignment of feature distributions in latent space and enable explicit domain adaptation.

In this work, we address detection and quantification of coronary artery calcium (CAC scoring) in contrast-enhanced coronary CT angiography (CCTA). Our aim is to exploit large sets of already annotated data in CT scans without contrast enhancement and extend the applicability of CAC scoring to CCTA. Current CAC scoring protocols are performed in a highly standardized manner without injection of iodinated contrast. Coronary artery calcifications are identified as high density areas of $\geq$ 130 Houndsfield Units (HU) in the coronary artery (11). Manual CAC scoring can be tedious and time-consuming, therefore, automated CAC scoring methods have been proposed (12, 13). Recent methods using deep learning have demonstrated accurate performance (14, 15). Given that CAC scoring is commonly performed in non-contrast CT (NCCT), automated methods have mostly focused on application in these scans. While earlier methods focused on a single type of NCCT scans (16–18) recent studies showed that the methods can generalize to diverse types of NCCT data. In a large-scale study containing data of 7,240 subjects, Van Velzen et al. (19) trained and evaluated a method proposed by Lessmann et al. (16) with different types of NCCT scans including scans from different hospitals, multiple scanners and multiple image acquisition protocols and demonstrated a good agreement between automated and manual scoring. Subsequently, Zeleznik et al. (20) demonstrated the robustness of a deep learning system for automated CAC scoring on routine cardiac gated and non-gated NCCT of 20,084 individuals.

In addition to CAC scoring in NCCT, CAC can be quantified in CCTA (21) and consequently, a number of methods automating the process have been developed (22–25). In a clinical cardiac CT exam, commonly cardiac NCCT is acquired first to determine the calcium score, which is followed by the acquisition of CCTA to detect presence of non-calcified plaque and stenosis in the coronary arteries. However, the amount of calcified plaque extracted from CCTA scans allows accurate cardiovascular risk stratification (22, 24). Hence, when the scan without contrast is not available, calcium scoring in CCTA may allow determination of patient's cardiovascular risk and thus allow better utilization of the already acquired data. Furthermore, performing CAC scoring in CCTA could allow omitting acquisition of the NCCT and thereby reduce the radiation dose to the patient and save scan time (24, 25).

Coronary artery calcium scoring in CCTA differs substantially from scoring in NCCT as the contrast material enhancing the coronary artery lumen typically exceeds the threshold (130 HU) used for CAC scoring in NCCT. Therefore, automatic methods trained on NCCT are not directly applicable to CCTA scans. Training the deep learning method with extra annotated CCTA data may improve its applicability to CCTA. However, manually annotating a large amount of representative training data is tedious and time consuming. To address this, in this study, we investigate the feasibility of adapting a CAC scoring network trained on a large set of labeled NCCT scans (16, 19) to unlabeled CCTA scans using unsupervised domain adaptation. For this, we investigate a cross-domain approach described by Dou et al. (8) to enable CAC scoring in CCTA without annotations while utilizing NCCT with available manual annotations.

2. Materials

2.1. Image data

This study includes three data sets. First, a data set of *labeled* low-dose chest NCCT scans from the National Lung Screening Trail (NLST) was used. The NLST enrolled 53,454 current or former heavy smokers aged 55–74 in the United States (26). In our previous study, a set of 1,687 baseline chest NCCT scans was selected (16). This set was designed to be diverse with respect to scanner model and reconstruction algorithm. The selected scans were acquired on 13 different scanner models in 31 hospitals. These chest NCCT scans were acquired with breath hold after inspiration and using a tube voltage 120 or 140 kVp, depending on the subjects weight. Scans were reconstructed to 0.49–0.98 mm in-plane resolution, 1–3 mm slice thickness, and 0.6–3 mm increment. For our work, all scans were resampled to 3 mm slice thickness and 1.5 mm increment, following earlier studies (16).

Second, a mixed set of *labeled* cardiac NCCT and *unlabeled* CCTA scans was used. Specifically, 200 labeled cardiac NCCT scans were acquired in clinical patient workup at University Medical Center Utrecht, The Netherlands (19, 27) and 200 unlabeled CCTA scans were acquired at Amsterdam University Medical Center location University of Amsterdam, The Netherlands. The cardiac NCCT scans were acquired with a Philips Brilliance iCT 256 scanner, with ECG synchronization and 120 kVp tube voltage. Scans were reconstructed to 0.29–0.49 mm in-plane resolution, 3 mm slice thickness, and 1.5 increment. The CCTA scans were acquired with a Siemens Somatom Force CT Scanner, with ECG synchronization and 70–120 kVp tube voltage. Scans were reconstructed to 0.22–0.46 mm in-plane resolution, 0.6 mm slice thickness, and 0.4 mm increment.

Third, a data set of *labeled* 313 CCTA scans from Amsterdam University Medical Center location University of Amsterdam, The Netherlands was used to evaluate the CAC detection on the target domain (CCTA test set). These CCTA scans were acquired with the Siemens Somatom Force CT Scanner, with ECG synchronization and 70–120 kVp tube voltage. Scans were reconstructed to 0.19–0.77 mm in-plane resolution, 0.6–1 mm slice thickness, and 0.4 mm increment.

2.2. Manual reference annotations

Manual reference labels of CAC were available from previous studies for the low-dose chest NCCT scans in the NLST data set(16) and the cardiac NCCT in the mixed set (19). The labeling was performed semi-automatically: all regions of ≥ 3 adjacent voxels with a CT value above 130 HU were shown as overlay. An observer manually identified lesions and labeled them according to their anatomical location, i.e., left anterior descending artery (LAD), left circumflex artery (LCX), or right coronary artery (RCA) (19). Given that chest CT without ECG synchronization does not allow visualization of the left main (LM) artery, CAC in the LM was labeled as LAD. Examples of chest NCCT slices and manual reference annotations are shown in the Supplementary Figure S1.

For the 200 CCTA scans in the mixed set, reference labels of CAC were not available. Hence, for the CCTA scans from the CCTA test set, CAC was manually annotated with a semi-automated method as either LAD, LCX, or RCA. This was done using an in-house developed software designed in MevisLab 3.2 (28). In agreement with manual labeling in NCCT, CAC in the LM was labeled as LAD. Because the standard 130 HU threshold for CAC detection in NCCT can not be used in CCTA, we used scan specific thresholds, following earlier studies (25, 29). For this, a region of interest (ROI) defined by a bounding box with a size around 35 × 36 × 44 voxels in the ascending aorta at the level of the origin of the left coronary artery was manually selected. Subsequently, the mean $mean_{ROI}$ and standard deviation STD_{ROI} from the CT values of the voxels within the ROI were used to compute a scan specific threshold $mean_{\mathrm{ROI}} + 3STD_{\mathrm{ROI}}$. Using this threshold, each coronary artery calcification was manually identified by a mouse click on the lesion. Subsequently, all connected voxels in the lesion above the scan specific threshold were marked as CAC in LAD, LCX, or RCA using 3D connected component labeling considering six-voxel connectivity. Examples of CCTA slices and manual reference annotations are shown in the Supplementary Figure S2.

In this study, NCCT scans (both chest and cardiac) are considered the source domain and CCTA scans are representing the target domain. The NCCT scans with CAC annotations from the NLST data set were used to train the CAC detection network on the source domain. The mixed set of labeled cardiac NCCT (source domain) and unlabeled CCTA (target domain) was used to train our unsupervised domain adaptation method. The labeled CCTA scans (target domain) in the CCTA test set were only used to evaluate the CAC detection on the target domain. The description of data sets and their usage are illustrated in Table 1.

3. Methods

A CNN is used for detecting CAC candidates in CCTA scans that is followed by false positive (FP) reduction, as shown in Figure 1. The CNN, which is trained on labeled NCCT data is adapted for application in CCTA using unsupervised domain adaptation. False positive reduction is performed by limiting the detected lesions to plausible CAC location and size.

3.1. CAC detection in CCTA with unsupervised domain adaptation

Unsupervised domain adaptation aims to transfer a model trained with data from a source domain with labels $\mathcal{D}_s = (X_s^i, Y_s^i)_{i=1..n_s}$ to a target domain without labels $\mathcal{D}_t = (X_t^i)_{i=1..n_t}$, where $\mathcal{D}$ represents domain, X represents images and Y represents labels. As proposed by Dou et al. (8), we use an adversarial training strategy to adapt the CNN to the target domain. In our application, a large set of chest NCCT scans with CAC labels is available, and hence, we aim to transfer the knowledge from NCCT to CCTA for CAC scoring. Therefore, the CAC scoring CNN trained with *labeled* low-dose chest NCCT scans is transfered to CCTA using adversarial domain adaptation.

We used our previous CAC scoring method described by Lessmann et al. (16) that has been trained and evaluated with a large set of low-dose chest NCCT data. The method consists of two sequential convolutional neural networks (CNN). The first CAC scoring CNN detects CAC candidates and labels them according to their anatomical location, i.e., as CAC in LAD, LCX, or RCA. The second CNN reduces the number of false positive detections. In our current work, only the first CNN is used to transfer knowledge obtained by training the network with NCCT to enable application in CCTA data using unsupervised domain adaptation.

To adapt the CAC detection network(16) from the source domain to the unlabeled target domain, we aim to align the distributions of extracted features from the two domains following the work by Dou et al. (8). For this, we divide the CAC detection network into two parts: a feature generator $G(\cdot)$ and a classifier $C(\cdot)$, as shown in Figure 1. The $G(\cdot)$ maps input images into feature representations in the latent space and the $C(\cdot)$ predicts the output class from the feature representations. The early layers of the network which are used for feature extraction are mostly related to the domain, while the deeper layers are mostly task-specific and learn semantic-level features for conducting the predictions (8, 30). Hence, we adapt the feature generator $G(\cdot)$ trained with NCCT to enable application in CCTA with adversarial domain adaptation, and we reuse the classifier $C(\cdot)$ as originally trained.

To enable adversarial learning, we design a discriminator $D(\cdot)$ to identify whether the features are from the source domain or the target domain. While the feature generator $G(\cdot)$ aims to extract features with similar distributions for both domains, the $D(\cdot)$ discriminates between the two domains (Figure 2). The adversarial loss based on the differences in feature distribution between the two domains is formulated as:

$$\mathcal{L}_{adv} = E_{x_t \in \mathcal{D}_t} log(D(G(x_t))) - E_{x_s \in \mathcal{D}_s} log(D(G(x_s))) \qquad (1)$$

where $G(\cdot)$ is optimized to minimize the adversarial loss, and $D(\cdot)$ is optimized to maximize the same loss. The generator $G(\cdot)$ is optimized based on the objective function calculated from the discriminator $D(\cdot)$, which can lead to an incorrect optimization forgetting the classification task. That means the features extracted by the trained $G(\cdot)$ can fool the $D(\cdot)$. However, these features are not beneficial for the final classification task $C(G(\cdot))$. For cross domain learning with *paired* data, the alignment loss in feature space, such as $L1(G(x_s), G(x_t))$ or $L2(G(x_s), G(x_t))$, can be used as a constraint for the generator optimization (31). For cross-domain learning with *unpaired* training data as in our case, such an alignment loss in feature space can not be used as a constraint for the generator optimization. In this work, the images were not registered to a common space either. Instead, as proposed in the work by Chen et al. (4), we use a classification loss in the source domain $\mathcal{D}_s$ as constraint to stabilize the training and avoid catastrophic forgetting.

The classification loss is formulated as:

$$\mathcal{L}_{cls} = L_{CE}(C(G(X_s)), Y_s) \qquad (2)$$

where L_{CE} is the cross-entropy loss, X_s and Y_s are the images and the corresponding reference labels on the source domain.

TABLE 1 Description of data and corresponding usage.

Scan type	#Scans	Reference	Domain	Usage
Chest NCCT	1,687	✓	Source	Training CAC scoring on source domain
Cardiac NCCT	200	✓	Source	Training unsupervised domain adaptation
CCTA	200	✗	Target	
CCTA	313	✓	Target	Testing CAC scoring on target domain

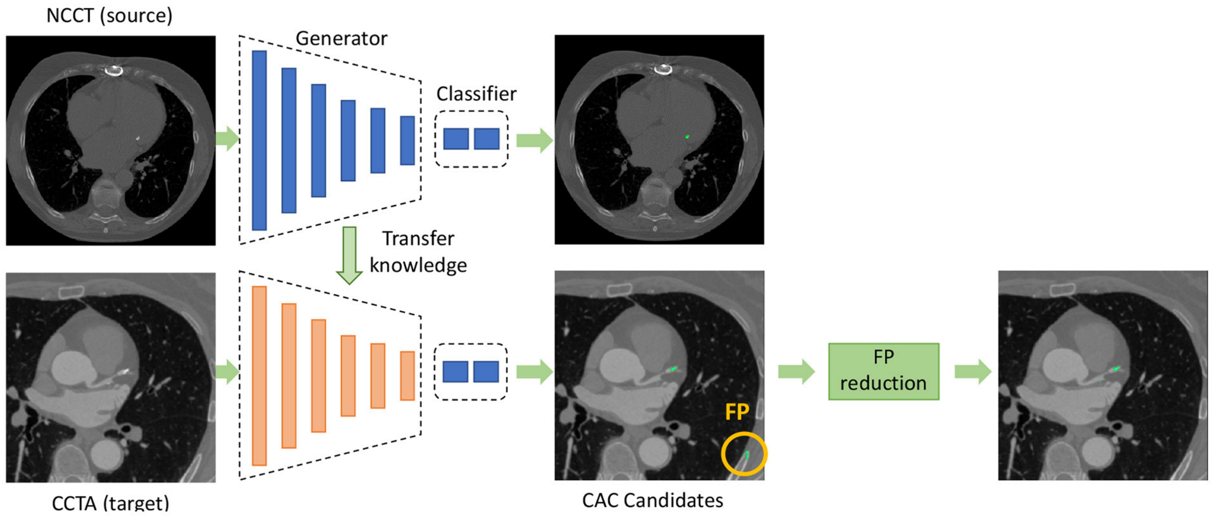


FIGURE 1
Overview of the proposed method for coronary artery calcium (CAC) detection in CCTA. The CNN for CAC detection is divided into a feature generator and a classifier. The feature generator is trained on source domain and is adapted to the target domain using unsupervised domain adaptation. The classifier in the target domain is reused from the source domain. After detection of CAC candidates using the CNN, false positive (FP) reduction is applied to remove FP detections.

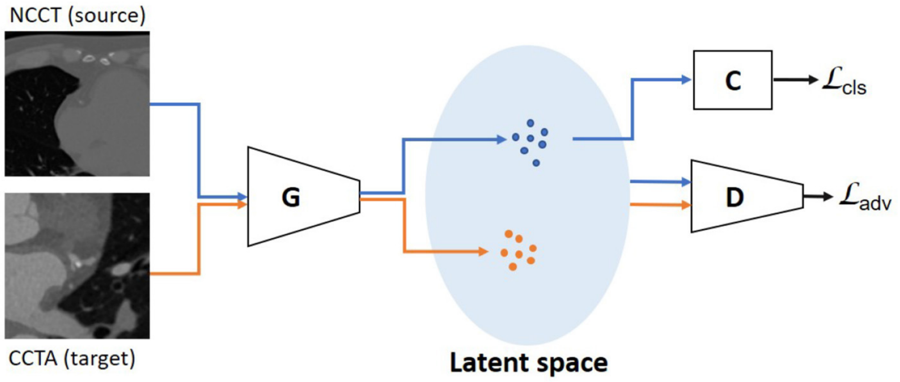


FIGURE 2
Unsupervised domain adaptation with *unpaired* data is performed using an adversarial learning strategy. The discriminator is optimized to distinguish the features from NCCT (source) domain and CCTA (target) domain. The generator is trained to extract features with similar distributions for the two domains. The blue dots in latent space represent features from the source domain, the orange ones from the target domain. The $\mathcal{L}_{adv}$ is used as the objective function and the $\mathcal{L}_{cls}$ is used as a constraint, which is determined on the source domain using the classifier.

During training, the $D(\cdot)$ is trained to maximize the objective of $\mathcal{L}_{adv}$, while the $G(\cdot)$ is optimized to minimize the objective of $\mathcal{L}_{adv}$ and $\mathcal{L}_{cls}$. These are formulated as:

$$\begin{aligned} &\max_D \mathcal{L}_{adv} \\ &\min_G \mathcal{L}_{adv} + \alpha \mathcal{L}_{cls} \end{aligned} \tag{3}$$

where α is a hyper-parameter for balancing the two loss terms. It is set to 2.0 in this work, based on a grid search strategy.

3.2. FP reduction

To identify CAC lesions, 3D connected component labeling is performed from the detected voxels and the scan specific threshold (25, 29). To remove potential false positive detections, detected lesions smaller than 1 mm^3 are discarded as those are likely noise voxels. Similarly, detected lesions larger than 500 mm^3 are discarded as those exceed the expected CAC volume (27). In addition, lesions detected outside the heart are discarded. For this, the heart volume is defined by segmentation of cardiac chambers, as described by Bruns et al. (32) which was trained with CCTA scans of 12 patients scanned for transcatheter aortic valve implantation (SOMATOM Force, Siemens, 70–120 kVp, 310–628 mAs, in-plane resolution 0.31–0.61 mm, slice thickess 0.31–0.61 mm, slice increment 0.45 mm). No additional changes or fine tuning for the data in this current study was performed. Subsequently, the segmentation of cardiac chambers was dilated by a sphere as a structuring element with diameter of 10.0 mm to ensure the heart wall and coronary arteries are included in the segmentation.

3.3. Evaluation

To evaluate the performance of CAC scoring on CCTA, the volume-wise and lesion-wise performance was determined by comparing automatically detected CAC with the manually annotated reference. Since the typically used Agatston score (11) is not applicable for CAC quantification in CCTA, the volume score was used. The evaluation was performed for total CAC and separately for CAC in LAD, LCX, and RCA. Both the volume-wise and lesion-wise performance was evaluated using sensitivity, false-positive (FP) rate, and F1 score (16). The agreement of calcium volume and number of lesions between the automatic detection and the reference labels was determined with Spearman correlation coefficients. Finally, the agreement between automatic volume scores and manual reference volume scores was assessed by examining Bland-Altman plots including 95% limits of agreement. Since errors tend to increase with increasing CAC volume, the variation of absolute differences between automatic and manual scores was modeled using regression for nonuniform differences(33). Because the absolute differences have a half-normal distribution, the modeled absolute differences were multiplied by $1.96 \times (\pi/2)^{0.5}$ to obtain the 95% limits of agreement.

4. Experiments and results

4.1. CAC scoring on CCTA

First, we retrained the two-stage CNNs for CAC detection (16) with the *labeled* chest NCCT data as the source domain. For this, the 1,687 NCCT scans in the NLST data set were randomly divided into 60% training set (1,012 scans), 10% validation set (169 scans), and 30% test set (506 scans). As originally reported (16), during the training, categorical cross-entropy was used as loss function, Adam was used as optimizer with a learning rate of 5×10^{-4}. The first CNN was trained with three orthogonal (axial, sagittal and coronal) patches of 155×155 pixels and the second CNN with three orthogonal patches of 65×65 pixels (16). Randomized patch extraction was used as augmentation for training.

Next, to stabilize adversarial training in the unsupervised domain adaptation, the generator was initialized with the weights of the CAC scoring model trained with the chest NCCT data from the NLST dataset. The unsupervised domain adaptation method was trained with the mixed dataset of *labeled* cardiac NCCT data from source domain and *unlabeled* CCTA data from target domain. When performing unsupervised domain adaptation with mixed data containing *labeled* cardiac NCCT and *unlabeled* CCTA scans the method achieved sensitivity of 0.78 in CCTA (Table 2). For comparison, the sensitivity of 0.53 was achieved when unsupervised domain adaptation was performed with mixed data containing *labeled* chest NCCT and *unlabeled* CCTA scans. *Labeled* cardiac NCCT data was chosen because these scans resemble CCTA scans more than chest NCCT. *Unlabeled* CCTA were used as *unlabeled* data from the target domain. To obtain a reliable discriminator, the discriminator was solely pretrained for 1,000 iterations first. Thereafter, the generator and discriminator were optimized together by training alternately. Specifically, the generator was optimized one iteration after every 20 iterations of the discriminator, according to the heuristic rules of training a Wasserstein GAN (34). Following the standard for adversarial training (34, 35), the discriminator was kept in a compact space. To enforce this constraint, the weights were clipped between $[-0.1, 0.1]$. The RMSProp optimizer was used to optimize the discriminator with a learning rate of 5×10^{-4}, and the generator with a learning rate of 5×10^{-5}, respectively (36). The optimal hyperparameters were determined by grid search. The adversarial learning was trained for 200 epoch. The networks were implemented in PyTorch (37). All the training was trained on NVIDIA GeForce RTX 2080 Ti.

To establish the performance of the CNN adapted from NCCT to CCTA, the network was evaluated with the 313 labeled CCTA test scans. The adapted CNN obtained an average volume-wise sensitivity of 0.78, an average FP volume per scan of 73.9 mm^3 and an F1-score of 0.41. After the FP reduction, the proposed method achieved an average volume-wise sensitivity of 0.80 with an average FP volume per scan of 19.8 mm^3, and F1 of 0.66. There were 36 patients without CAC but with FP detected by the proposed method, with an average FP volume per scan of 40 mm^3. The Spearman correlation between automatically detected and reference CAC volume was 0.73. The Bland-Altman plots comparing automatically detected CAC volume with manually annotated reference are illustrated in Figure 3.

Coronary CT angiography slices and corresponding automatic CAC detections for two outliers cases (marked orange in Figure 3) are shown in Figures 4a,b. In addition, two representative cases from the labeled CCTA test set are shown in Figures 4c,d. For lesion-wise evaluation, the proposed method achieved an average sensitivity of 0.79 and FP lesion per scan of 1.06. The correlation between the number of automatically detected and manually annotated reference lesions was 0.69.

TABLE 2 Results of the automatic CAC scoring evaluated by volume-wise sensitivity, FP volume per scan, and F1-score between automatic detection and manual reference.

		NCCT [506]	CCTA [313]			
	$\mathcal{L}_{adv}$	✗	✓	✓	✓	✗
	$\mathcal{L}_{cls}$	✗	✓	✓	✗	✗
	FP reduction	✗	✓	✗	✗	✗
	Sensitivity	0.89 (0.25)	0.80 (0.32)	0.78 (0.33)	0.68 (0.38)	0.41 (0.48)
CAC	FP volume/scan	73.6 (141)	19.8 (60.6)	64.5 (150)	25.8 (70)	132 (205)
	F1	0.66 (0.37)	0.66 0.38	0.41 (0.40)	0.49 (0.41)	0.16 (0.36)
	Sensitivity	0.92 (0.21)	0.89 (0.27)	0.86 (0.28)	0.79 (0.33)	0.47 (0.48)
LAD	FP volume/scan	31.6 (79.6)	13.9 (45.5)	44.5 (118)	20.2 (54.4)	55.8 (90.5)
	F1	0.79 (0.34)	0.74 (0.37)	0.48 (0.42)	0.56 (0.42)	0.24 0.41
	Sensitivity	0.88 (0.29)	0.74 (0.44)	0.71 (0.45)	0.71 (0.46)	0.66 (0.48)
LCX	FP volume/scan	19.7 (55.6)	0.13 (1.13)	0.17 (1.01)	0.02 (0.31)	1.60 (0.30)
	F1	0.67 (0.42)	0.74 (0.44)	0.69 (0.46)	0.70 (0.46)	0.66 (0.48)
	Sensitivity	0.89 (0.26)	0.87 (0.30)	0.87 (0.31)	0.80 (0.38)	0.67 (0.47)
RCA	FP volume/scan	30.1 (73.4)	6.80 (35.6)	21.3 (78.1)	6.64 (35.6)	77.6 (157)
	F1	0.65 (0.42)	0.73 (0.41)	0.52 (0.46)	0.68 (0.44)	0.31 (0.46)

The method with different settings (using adversarial loss and classification loss in the CAC detection network, and false positive reduction stage) is tested on chest NCCT data and CCTA data. FP volume/scan is given in mm^3. The results are shown as average (standard deviation) for total CAC as well as for LAD, LCX, and RCA separately. $\mathcal{L}_{adv}$, adversarial loss; $\mathcal{L}_{cls}$, classification loss; CAC, coronary artery calcification; LAD, left anterior descending artery; LCX, left circumflex artery; RCA, right coronary artery.

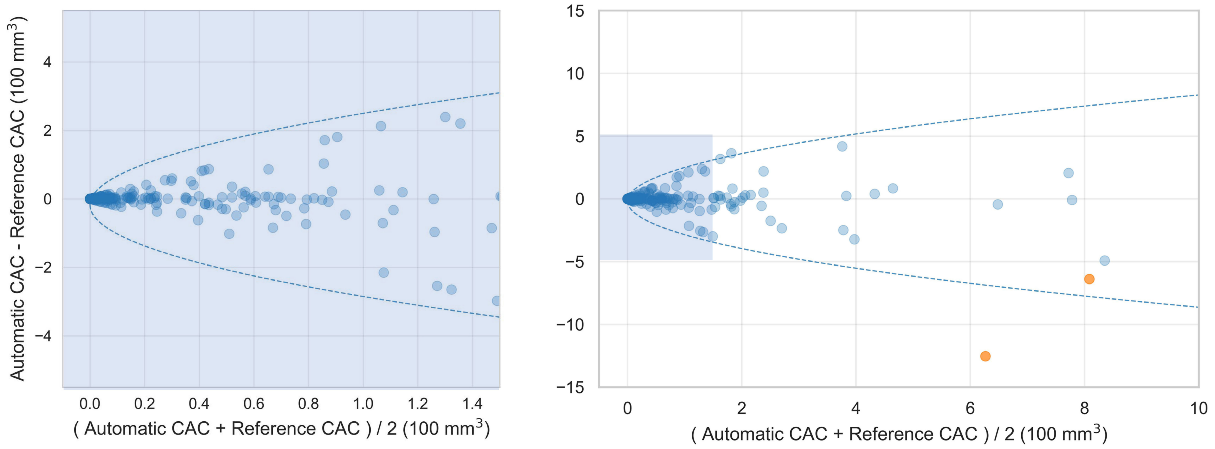


FIGURE 3
Bland-Altman plots comparing automatically detected CAC volume with the manual reference volume. 95% limits of agreement are represented by the formula: $Difference = \pm 1.96 \times (\pi/2)^{0.5} \times (b + a \times Mean^{0.5})$, with $a = 10.9$ and $b = -17.8$. Two outlier cases are colored orange. The Bland-Altman plot of lesions with volume less than 150 mm^3 is shown on the left and all lesions is shown on the right.

4.2. Ablation study

To establish whether our retraining of the original CAC scoring network on the source domain led to adequate performance, the CAC scoring network was evaluated on NLST test set (Section CAC scoring on CCTA) and compared with the originally reported results (16). Results are listed in Table 2 (column 3 showing NCCT results). Our retained network obtained a sensitivity of 0.89, an average FP volume of 73.6 mm^3 per scan and F1 of 0.66. The sensitivity is in agreement with the results (0.84 - 0.91) reported in the original work (16), while the originally reported FP rate (40.7–62.8 mm^3) and therefore F1 (0.84–0.89) slightly outperform our results.

To evaluate the performance of the two-stage CAC scoring networks trained on NCCT to CCTA, the trained CNNs was directly applied to CCTA test scans without adversarial domain adaptation learning. This led to an average sensitivity of 0.41, an average FP volume per scan of 139.7 mm^3, and F1 of 0.16

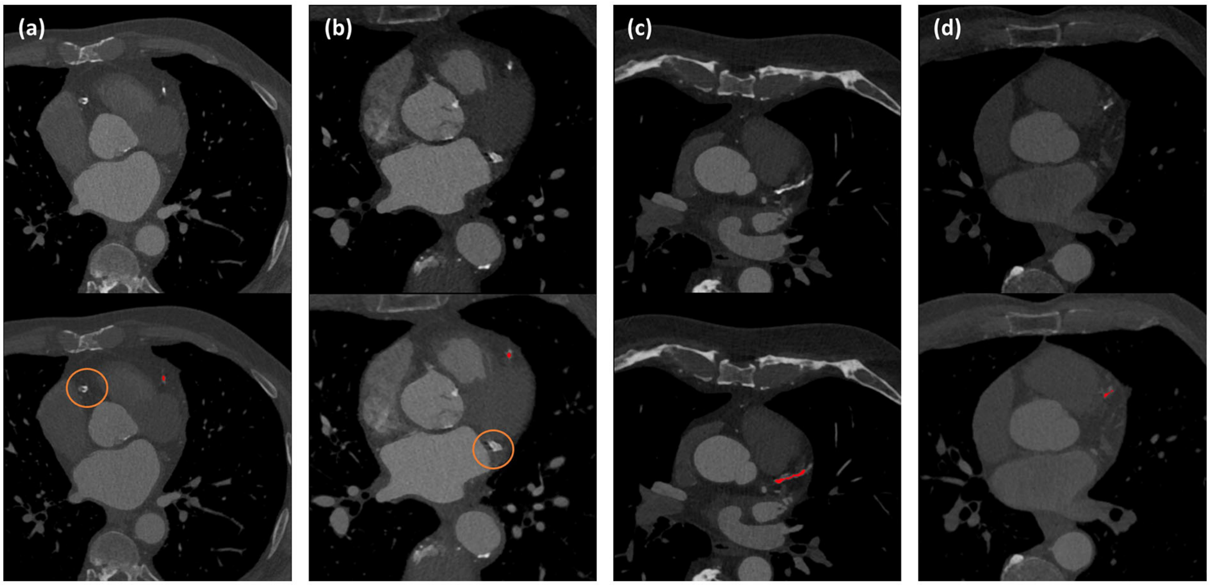


FIGURE 4
Automated CAC detection results in CCTA scans of four patients. The images in the first row show CCTA slices and the detected CACs are shown as overlay in the second row. Panels **(a)** and **(b)** illustrate the two largest outliers shown by orange dots in Figure 3, and false negative CAC are indicated by orange circles. Panels **(c)** and **(d)** show two cases with correct automatic CAC detections.

(Table 2, column 7 showing the CCTA results). Subsequently, adding FP reduction led to an average sensitivity of 0.43, an average FP volume of 0.58 mm^3 and F1 of 0.41. Note that FP reduction stage slightly improved the sensitivity as the region-growing algorithm (38) used to define the lesions from the voxels detected by the CNN may improve lesion segmentation and lead to better agreement with manual reference that used the region-growing algorithm to define CAC lesions.

To investigate the benefit of using the adversarial loss and classification loss for domain adaptation, and FP reduction, additional experiments were performed. The proposed method obtained a volume-wise sensitivity of 0.80, average FP volume per scan of 19.8 mm^3, and F1 of 0.66. Without FP reduction, the volume-wise sensitivity decreased to 0.78, average FP volume per scan increased to 64.5 mm^3 and consequently, F1 score decreased to 0.41. Furthermore, removing the classification loss $\mathcal{L}_{cls}$ from the objective function resulted in the volume-wise sensitivity of 0.68, average FP volume per scan of 25.8 mm^3, and F1 of 0.49. Finally, as described above, removing the adversarial loss $\mathcal{L}_{adv}$ (i.e., without adversarial domain adaptation learning) led to sensitivity of 0.41, FP volume of 139.7 mm^3 per scan, and F1 of 0.16. Detailed results are listed in Table 2 columns 4–7.

4.3. Comparison with previous work

The performance of the proposed method was compared with previously published methods that use deep learning for CAC scoring in CCTA scans (22–25). Wolterink et al. (25) proposed a method that employed paired CNNs for CAC scoring. The first CNN was used to identify CAC-like voxels and the second CNN was used to reduce CAC-like negatives. Fischer et al. (22) proposed a method that firstly detected the coronary artery centerlines and then identified CAC in cross-sectional images along the detected centerlines using long short-term memory (LSTM). In the study by Liu et al. (23), a vessel focused 3D CNN was proposed for CAC detection. The coronary arteries were firstly extracted and straightened volumes were reformed along the coronary arteries. Thereafter, a CNN was used for CAC detection. The results as reported in the original work are listed in Table 3. These demonstrate that our unsupervised method achieved competitive performance. Given that the original implementations of these earlier studies are not publicly available, the compared methods the results should be used as indication only.

5. Discussion

In this work, we have utilized an unsupervised domain adaptation method described by Dou et al. (8) employing a CNN architecture which enables CAC scoring in CCTA while learning from annotated non-representative CT scans without contrast and representative CCTA without reference annotations. For this, the first-stage CNN as previously designed by Lessmann et al. for CAC scoring (16) is divided into a feature generator and a classifier. The feature generator is adapted from NCCT

TABLE 3 Comparison with previously published results on automated coronary artery calcium scoring on CCTA.

			Lesion-wise evaluation			Volume-wise evaluation		
Method	**# train**	**# test**	**Sensitivity**	**FP lesion**	**F1**	**Sensitivity**	**FP volume**	**F1**
Wolterink et al. (25)	150	100	0.71	0.48	–	–	–	–
Liu et al. (23)	80	20	–	–	–	0.85	–	0.83
Fischer et al. (22)	232	194	0.92	0.20	–	–	–	–
Ours	–	313	0.79	1.06	0.66	0.80	19.8	0.66

The number of labeled CCTA scans used for training (# train) and testing (# test) are listed. Performance [sensitivity, false positives (FP) per scan and F1-score] using CAC lesions and volume are given.

to CCTA through adversarial unsupervised domain adaptation and the classifier trained on NCCT is reused. An adversarial loss and classification loss on source domain are used as the objective function. The results demonstrate that the method achieves a competitive performance.

Like previous methods for automatic calcium scoring, our method consists of two distinct stages. In the first stage, a CNN for CAC detection and labeling in non-contrast chest CT from previous work (16) is adjusted for the CAC scoring in CCTA. The ablation study showed that our retraining of the CAC detection CNN did not lead to the same performance reported in the original manuscript (16). However, there are several differences. First, although training and test scans originate from the same set, exact division on the scans into training and test set differs. Second, the original work reported results separately for sharp and soft kernel CT reconstructions, while we did not distinguish between these. Like in the original work, a second stage is used to reduce the number of false positives. Using the described approach for CAC scoring in CCTA, simple image processing (restricting allowed volume of CAC, limiting the analysis to the volume of interest) substantially reduced false positive detections. Nevertheless, retrospective analysis showed that occasionally false positives remain inside heart and in the coronary arteries with high HU value. Visual analysis of the results showed small false positive detections in the distal RCA representing contrast material. This is also reflected in the limited Spearman correlation coefficient between the detected and reference lesions. This might be due to the varying contrast levels of CCTA, where parts of the coronary artery lumen had a very high HU value. Likely, locally defined threshold for the extraction of CAC would alleviate the problem. Future research should investigate whether this would would benefit the overall performance. In few cases false positive detections were representing extra-coronary calcifications. Those were aortic calcification in the vicinity of the coronary ostia or calcifications in the aortic valves, which is not uncommon to automatic calcium scoring methods(19).

Retrospective analysis of the outliers shown in Figures 3 and 4 showed that in one case, a large CAC in the RCA (625 mm^3) was detected by the CNN but removed in the FP reduction stage because its volume exceeded the maximum expected CAC volume. In the other case, large CAC in LCX (313 mm^3) was not detected by the CNN. In our training set, median (Q1, Q3) CAC was 7.1 (1.6, 29.2) mm^3 and 95th percentile was 188 mm^3. This shows that the volumes of our false negatives substantially exceeded CAC examples in the training set. Adding examples of large CAC lesions in the training set or learning specifically focused on rare CAC examples might improve the performance.

To train the CNN for detection and labeling of CAC, three different data sets were used. First, we reused the CNN trained on a large set of labeled chest CTs without contrast enhancement. To achieve unsupervised domain adaptation, non-representative labeled cardiac CT without contrast and representative unlabeled CCTA were used. Future work could investigate the optimal size of each set and the optimal way of injecting different data into the training, e.g., training the CNN with different non-contrast CT scan types, refinement with specific data or introducing different data in the domain adaptation stage.

To make the cross domain training stable with unpaired data, the classification loss on the source domain was used. For cross domain learning with paired data, a feature-wise loss could be used (31). Given that we don't have paired data or register the images to a common space, this kind of loss is not applicable in our study. In our work, the feature generator was adapted from source domain to target domain, however, the classifier was directly reused. This could be done even though the input images to feature generator are from different domains because the classifier performs the same task with aligned feature distributions.

To transfer the knowledge of CAC detection from NCCT to CCTA, unsupervised domain adaptation was used. When a limited set of annotated training data from the target domain is available, it is common to pretrain the network with labeled data from the source domain and fine-tune the network with this small set (30, 39). In our case, annotated training data from the target domain is not available and unsupervised domain adaptation allows the training with labeled data from the source domain and unlabelled data from the target domain. Future work could investigate whether a small set of annotated images from the target domain may benefit the performance, possibly also by combining transfer learning approaches with unsupervised domain adaptation.

In this study, following the work by Dou et al. (8), the knowledge about CAC detection was transferred from NCCT to CCTA by aligning the feature distributions between the two domains. However, Chen et al. (7) performed unsupervised domain adaptation by aligning the domains in both image and feature perspectives. The image alignment was used to transform the image appearance and narrow the domain shift between source and target domains. However, we opted for feature alignment only because lack of visible anatomical boundaries in non-contrast scans (arteries, cardiac chambers) to guide the image registration renders image alignment a highly challenging task. Moreover, very small CAC may disappear due to registration, which would not be beneficial for learning.

Comparing the proposed method with previously published deep learning methods on CAC scoring in CCTA scans showed that the proposed method achieved a competitive sensitivity. However, the number of false positive detections did not reach the performance of supervised methods. Methods (22, 23) that limited the ROI for CAC scoring with coronary artery extraction, achieved a lower number of FP detections. Future research could investigate whether limiting the the analysis to the vicinity of the coronary arteries like proposed by Fischer et al. (22) and Liu et al. (23) would be beneficial. For this, tracking the coronary artery centerline (40) could be used.

Bland-Altman plot shown in Figure 3 shows heteroskedastic-like behavior of CAC scores. This behavior is not uncommon for CAC scoring methods, because typically errors tend to increase with higher CAC scores (19, 24). False negative detections tend to be larger in patients with higher calcium burden, possibly because their lesions tend to be larger. Moreover, larger false positive detections often consist of non-coronary calcifications, e.g., aortic calcifications in the vicinity of the coronary ostia or cardiac valves, which are also typically larger in patients with a higher coronary calcium burden. To calculate the 95% confidence intervals of the Bland-Altman plots we accounted for the heteroskedastic behavior by modeling the variation in absolute differences (33).

While CCTA scans are mainly made to provide important information on the presence and the amount of non-calcified plaque and stenosis, cardiac CT scans without contrast enhancement are the reference modality for quantification of calcified coronary artery plaque. Hence, limitation of our method is its ability to quantify calcified plaque in CCTA only. To fully exploit information contained in CCTA, our further work will focus on extending the method to quantification of calcified and non-calcified plaque and stenosis.

In this work, the unsupervised domain adaptation method was trained with 200 NCCT scans and 200 CCTA scans. Like with any machine learning methods, training the unsupervised domain adaptation method with more scans that include more diversity would likely lead to more accurate performance. Finding the optimal set size should be a topic of future research.

In the literature, a wide range in inter-observer agreement for CAC quantification in CCTA has been reported. Specifically, 11% variability in CAC volume when utilizing a scan-specific threshold (41) and 13–25% when using manual delineation of CAC (42). Moreover, correlation of CAC volume between observers of 0.89–0.98 has been reported (42, 43). In the current study the variability between automatic and reference scores was 21%, with a correlation of 0.73. Given that no clinically used risk categories are defined based on CAC volume or other CAC score quantified from CCTA, it remains unclear whether the obtained errors impact clinical decision-making. Therefore, further work needs to investigate the value of the extracted CAC scores for predicting cardiovascular events.

In conclusion, an unsupervised domain adaptation method for CAC scoring that transfers knowledge from NCCT with reference labels to CCTA without reference labels has been presented. The results show that the method achieves a competitive performance. This may allow for better utilization of the existing large and annotated data sets and extend applicability to diverse CT scans without the requirement of extra annotations.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: The data from NLST for this study can be requested at the provider. The cardiac NCCT and CCTA are in-home data. Requests to access these datasets should be directed to https://cdas.cancer.gov/datasets/nlst/.

Ethics statement

The studies involving human participants were reviewed and approved by University Medical Center Utrecht; Amsterdam University Medical Center. The patients/participants provided their written informed consent to participate in this study.

Author contributions

ZZ: conceptualized the study, developed the software, analyzed the data, and drafted the article and revised the manuscript. SV and II: conceptualized the study and drafted and revised the manuscript. NL, NP, and TL: acquired data and revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work is part of the research program Deep Learning for Medical Image Analysis under project number P15-26 project 3 financed by the Dutch Technology Foundation with contribution by Philips Healthcare.

Acknowledgments

The authors thank the National Cancer Institute for access to NCI's data collected by the National Lung Screening Trial. The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI.

Conflict of interest

Author II reports institutional research grants by Pie Medical Imaging, Esaote, Dutch Technology Foundation with participation of Pie Medical Imaging and Philips Healthcare (DLMedIA P15-26). Author TL reports institutional research grants by Pie Medical Imaging, Dutch Technology Foundation with participation of Pie Medical Imaging and Philips Healthcare (DLMedIA P15-26).

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcvm.2022.981901/full#supplementary-material

References

1. Ker J, Wang L, Rao J, Lim T. Deep learning applications in medical image analysis. *IEEE Access.* (2017) 6:9375–89. doi: 10.1109/ACCESS.2017.2788044

2. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal.* (2017) 42:60–88. doi: 10.1016/j.media.2017.07.005

3. Ben-David S, Blitzer J, Crammer K, Pereira F. Analysis of representations for domain adaptation. In: Schölkopf B, Platt J, Hoffman T, editors. *Advances in Neural Information Processing Systems.* Vol. 19. Barcelona 2006). p. 137–44.

4. Chen M, Zhao S, Liu H, Cai D. Adversarial-learned loss for domain adaptation. *Proc AAAI Confer Artif Intell.* (2022) 34:3521–8. doi: 10.1609/aaai.v34i04.5757

5. Wei G, Lan C, Zeng W, Chen Z. Metaalign: Coordinating domain alignment and classification for unsupervised domain adaptation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* Nashville, TN (2021). p. 16643–53. doi: 10.1109/CVPR46437.2021.01637

6. Wei G, Lan C, Zeng W, Zhang Z, Chen Z. ToAlign: task-oriented alignment for unsupervised domain adaptation. *35th Conference on Neural Information Processing Systems (NeurIPS 2021).* (2021) 13834–46.

7. Chen C, Dou Q, Chen H, Qin J, Heng PA. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *IEEE Trans Med Imaging.* (2020) 39:2494–505. doi: 10.1109/TMI.2020.2972701

8. Dou Q, Ouyang C, Chen C, Chen H, Glocker B, Zhuang X, et al. PnP-Adanet: plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation. *IEEE Access.* (2019) 7:99065–76. doi: 10.1109/ACCESS.2019.2929258

9. Guan H, Liu M. Domain adaptation for medical image analysis: a survey. *IEEE Trans Biomed Eng.* (2021) 69:1173–85. doi: 10.1109/TBME.2021.3117407

10. Wu F, Zhuang X. CF distance: a new domain discrepancy metric and application to explicit domain adaptation for cross-modality cardiac image segmentation. *IEEE Trans Med Imaging.* (2020) 39:4274–85. doi: 10.1109/TMI.2020.3016144

11. Agatston AS, Janowitz WR, Hildner FJ, Zusmer NR, Viamonte M, Detrano R. Quantification of coronary artery calcium using ultrafast computed tomography. *J Amer Coll Cardiol.* (1990) 15:827–32. doi: 10.1016/0735-1097(90)90282-T

12. Hecht HS. Coronary artery calcium scanning: past, present, and future. *JACC Cardiovasc Imaging.* (2015) 8:579–96. doi: 10.1016/j.jcmg.2015.02.006

13. van Velzen SG, Hampe N, de Vos BD, Išgum I. Artificial intelligence-based evaluation of coronary calcium. In: De Cecco CN, van Assen M, Leiner T, editors. *Artificial Intelligence in Cardiothoracic Imaging. Contemporary Medical Imaging.* Cham: Humana (2022) p. 245–57.

14. Hampe N, Wolterink JM, Van Velzen SG, Leiner T, Išgum I. Machine learning for assessment of coronary artery disease in cardiac CT: a survey. *Front Cardiovasc Med.* (2019) 6:172. doi: 10.3389/fcvm.2019.00172

15. Litjens G, Ciompi F, Wolterink JM, de Vos BD, Leiner T, Teuwen J, et al. State-of-the-art deep learning in cardiovascular image analysis. *JACC Cardiovasc Imaging.* (2019) 12(8 Pt 1):1549–65. doi: 10.1016/j.jcmg.2019.06.009

16. Lessmann N, van Ginneken B, Zreik M, de Jong PA, de Vos BD, Viergever MA, et al. Automatic calcium scoring in low-dose chest CT using deep neural networks with dilated convolutions. *IEEE Trans Med Imaging.* (2017) 37:615–25. doi: 10.1109/TMI.2017.2769839

17. Martin SS, van Assen M, Rapaka S, Hudson Jr HT, Fischer AM, Varga-Szemes A, et al. Evaluation of a deep learning–based automated CT coronary artery calcium scoring algorithm. *Cardiovasc Imaging.* (2020) 13(2_Pt_1):524–6. doi: 10.1016/j.jcmg.2019.09.015

18. van den Oever LB, Cornelissen L, Vonder M, Xia C, van Bolhuis JN, Vliegenthart R, et al. Deep learning for automated exclusion of cardiac CT examinations negative for coronary artery calcium. *Eur J Radiol.* (2020) 129:109114. doi: 10.1016/j.ejrad.2020.109114

19. van Velzen SG, Lessmann N, Velthuis BK, Bank IE, van den Bongard DH, Leiner T, et al. Deep learning for automatic calcium scoring in CT: validation using multiple cardiac CT and chest CT protocols. *Radiology.* (2020) 295:66–79. doi: 10.1148/radiol.2020191621

20. Zeleznik R, Foldyna B, Eslami P, Weiss J, Alexander I, Taron J, et al. Deep convolutional neural networks to predict cardiovascular risk from computed tomography. *Nat Commun.* (2021) 12:1–9. doi: 10.1038/s41467-021-20966-2

21. Al'Aref SJ, Maliakal G, Singh G, van Rosendael AR, Ma X, Xu Z, et al. Machine learning of clinical variables and coronary artery calcium scoring for the prediction of obstructive coronary artery disease on coronary computed tomography angiography: analysis from the CONFIRM registry. *Eur Heart J.* (2020) 41:359–67. doi: 10.1093/eurheartj/ehz565

22. Fischer AM, Eid M, De Cecco CN, Gulsun MA, Van Assen M, Nance JW, et al. Accuracy of an artificial intelligence deep learning algorithm implementing a recurrent neural network with long short-term memory for the automated detection of calcified plaques from coronary computed tomography angiography. *J Thorac Imaging.* (2020) 35:S49–57.doi: 10.1097/RTI.0000000000000491

23. Liu J, Jin C, Feng J, Du Y, Lu J, Zhou J. A vessel-focused 3D convolutional network for automatic segmentation and classification of coronary artery plaques in cardiac CTA. In: *International Workshop on Statistical Atlases and Computational Models of the Heart*. Cham: Springer (2018). p. 131–41. doi: 10.1007/978-3-030-12029-0_15

24. Mu D, Bai J, Chen W, Yu H, Liang J, Yin K, et al. Calcium scoring at coronary CT angiography using deep learning. *Radiology.* (2022) 302:309–16. doi: 10.1148/radiol.2021211483

25. Wolterink JM, Leiner T, de Vos BD, van Hamersvelt RW, Viergever MA, Išgum I. Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks. *Med Image Anal.* (2016) 34:123–36. doi: 10.1016/j.media.2016.04.004

26. Team NLSTR. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med.* (2011) 365:395–409. doi: 10.1056/NEJMoa1102873

27. Wolterink JM, Leiner T, Takx RA, Viergever MA, Išgum I. Automatic coronary calcium scoring in non-contrast-enhanced ECG-triggered cardiac CT with ambiguity detection. *IEEE Trans Med Imaging.* (2015) 34:1867–78. doi: 10.1109/TMI.2015.2412651

28. Ritter F, Boskamp T, Homeyer A, Laue H, Schwier M, Link F, et al. Medical image analysis: a visual approach. *IEEE Pulse.* (2011) 2:60–70. doi: 10.1109/MPUL.2011.942929

29. Mylonas I, Alam M, Amily N, Small G, Chen L, Yam Y, et al. Quantifying coronary artery calcification from a contrast-enhanced cardiac computed tomography angiography study. *Eur Heart J Cardiovasc Imaging.* (2014) 15:210–5. doi: 10.1093/ehjci/jet144

30. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? In: Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ. *Advances in Neural Information Processing Systems 27 (NIPS 2014)*. Montreal, QC (2014) 27:3320–8.

31. van Tulder G, de Bruijne M. Learning cross-modality representations from multi-modal images. *IEEE Trans Med Imaging.* (2018) 38:638–48. doi: 10.1109/TMI.2018.2868977

32. Bruns S, Wolterink JM, van den Boogert TP, Runge JH, Bouma BJ, Henriques JP, et al. Deep learning-based whole-heart segmentation in 4D contrast-enhanced cardiac CT. *Comput Biol Med.* (2022) 142:105191. doi: 10.1016/j.compbiomed.2021.105191

33. Sevrukov AB, Bland JM, Kondos GT. Serial electron beam CT measurements of coronary artery calcium: Has your patient's calcium score actually changed? *Amer J Roentgenol.* (2005) 185:1546–53. doi: 10.2214/AJR.04.1589

34. Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: *International Conference on Machine Learning.* PMLR. Sydney (2017). 70:214–23.

35. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of wasserstein GANS. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, CA (2017) 30:5767–77.

36. Graves A. Generating sequences with recurrent neural networks. *arXiv Preprint*. (2013) arXiv:13080850.

37. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems.* Vol. 32. Curran Associates, Inc. (2019). p. 8024–35. Available online at: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

38. Hojjatoleslami S, Kittler J. Region growing: a new approach. *IEEE Trans Image Process.* (1998) 7:1079–84. doi: 10.1109/83.701170

39. Minaee S, Kafieh R, Sonka M, Yazdani S, Soufi GJ. Deep-COVID: predicting COVID-19 from chest X-ray images using deep transfer learning. *Med Image Anal.* (2020) 65:101794. doi: 10.1016/j.media.2020.101794

40. Wolterink JM, van Hamersvelt RW, Viergever MA, Leiner T, Išgum I. Coronary artery centerline extraction in cardiac CT angiography using a CNN-based orientation classifier. *Med Image Anal.* (2019) 51:46–60. doi: 10.1016/j.media.2018.10.005

41. Øvrehus KA, Schuhbaeck A, Marwan M, Achenbach S, Nørgaard BL, Bøtker HE, et al. Reproducibility of semi-automatic coronary plaque quantification in coronary CT angiography with sub-mSv radiation dose. *J Cardiovasc Comp Tomogr.* (2016) 10:114–20. doi: 10.1016/j.jcct.2015.11.003

42. Cheng VY, Nakazato R, Dey D, Gurudevan S, Tabak J, Budoff MJ, et al. Reproducibility of coronary artery plaque volume and composition quantification by 64-detector row coronary computed tomographic angiography: an intraobserver, interobserver, and interscan variability study. *J Cardiovasc Comp Tomogr.* (2009) 3:312–20. doi: 10.1016/j.jcct.2009.07.001

43. Dey D, Cheng VY, Slomka PJ, Nakazato R, Ramesh A, Gurudevan S, et al. Automated 3-dimensional quantification of noncalcified and calcified coronary plaque from coronary CT angiography. *J Cardiovasc Comp Tomogr.* (2009) 3:372–82. doi: 10.1016/j.jcct.2009.09.004

frontiers | Frontiers in Cardiovascular Medicine

TYPE Original Research
PUBLISHED 23 September 2022
DOI 10.3389/fcvm.2022.983091

Check for updates

OPEN ACCESS

EDITED BY
Xiang Li,
Massachusetts General Hospital and Harvard Medical School, United States

REVIEWED BY
Jianlong Yang,
Shanghai Jiao Tong University, China
Eung Joo Lee,
Massachusetts General Hospital and Harvard Medical School, United States

*CORRESPONDENCE
Víctor M. Campello
victor.campello@ub.edu

SPECIALTY SECTION
This article was submitted to Cardiovascular Imaging, a section of the journal Frontiers in Cardiovascular Medicine

RECEIVED 30 June 2022
ACCEPTED 01 September 2022
PUBLISHED 23 September 2022

CITATION
Campello VM, Xia T, Liu X, Sanchez P, Martín-Isla C, Petersen SE, Seguí S, Tsaftaris SA and Lekadir K (2022) Cardiac aging synthesis from cross-sectional data with conditional generative adversarial networks. *Front. Cardiovasc. Med.* 9:983091. doi: 10.3389/fcvm.2022.983091

Cardiac aging synthesis from cross-sectional data with conditional generative adversarial networks

Víctor M. Campello[1]*, Tian Xia[2], Xiao Liu[2], Pedro Sanchez[2], Carlos Martín-Isla[1], Steffen E. Petersen[3,4,5,6], Santi Seguí[1], Sotirios A. Tsaftaris[2,6] and Karim Lekadir[1]

[1]Artificial Intelligence in Medicine Lab (BCN-AIM), Universitat de Barcelona, Barcelona, Spain, [2]Institute for Digital Communications, School of Engineering, University of Edinburgh, Edinburgh, United Kingdom, [3]William Harvey Research Institute, NIHR Barts Biomedical Research Centre, Queen Mary University London, London, United Kingdom, [4]Barts Heart Centre, St Bartholomew's Hospital, Barts Health NHS Trust, London, United Kingdom, [5]Health Data Research UK, London, United Kingdom, [6]The Alan Turing Institute, London, United Kingdom

Age has important implications for health, and understanding how age manifests in the human body is the first step for a potential intervention. This becomes especially important for cardiac health, since age is the main risk factor for development of cardiovascular disease. Data-driven modeling of age progression has been conducted successfully in diverse applications such as face or brain aging. While longitudinal data is the preferred option for training deep learning models, collecting such a dataset is usually very costly, especially in medical imaging. In this work, a conditional generative adversarial network is proposed to synthesize older and younger versions of a heart scan by using only cross-sectional data. We train our model with more than 14,000 different scans from the UK Biobank. The induced modifications focused mainly on the interventricular septum and the aorta, which is consistent with the existing literature in cardiac aging. We evaluate the results by measuring image quality, the mean absolute error for predicted age using a pre-trained regressor, and demonstrate the application of synthetic data for counter-balancing biased datasets. The results suggest that the proposed approach is able to model realistic changes in the heart using only cross-sectional data and that these data can be used to correct age bias in a dataset.

KEYWORDS

aging heart, generative adversarial network, magnetic resonance imaging, synthesis, data augmentation

Introduction

Understanding the effects of the aging process is becoming more important as the life expectancy increases worldwide. Aging has crucial implications for health and age is the main risk factor for the development of cardiovascular disease (1, 2). Insights into the aging mechanism can be very valuable to inform new interventions to delay the occurrence of possible adverse events and for improving health of the elderly.

According to the medical literature, age is positively related to morphological changes in the heart such as increased left atrial diameter (3), increased wall thickness in the left ventricle (LV) and reduced LV dimensions (1, 2). These changes are associated with atrial fibrillation and heart failure with preserved ejection fraction (2, 4). Females and males show differences in the aforementioned changes with increased LV wall thickness being more prevalent in women (5). Also, a marked increase in epicardial adipose tissue deposition has been observed with age (5).

Collecting longitudinal data is very time-consuming and requires repeated visits of participants with the associate chance of dropouts along the duration of the study. Two longitudinal studies have analyzed cardiac health with more than three decades of measurements. These are the Framingham Heart Study (FHS) (6) and the Baltimore Longitudinal Study on Aging (7). However, imaging is only available for the FHS and only for echocardiography. Imaging with higher spatial resolution may be found in two other longitudinal studies, the UK Biobank (ukbiobank.ac.uk) and the Multi-Ethnic Study of Atherosclerosis (MESA) study (8), where participants are scanned using magnetic resonance imaging (MRI). However, only a subset of participants have repeated scans adquired in the next 1–10 years after the first scan visit. Thus, modeling the aging process in the heart with good spatial resolution is restricted to 10 years or less if one relies only on longitudinal data, and the analysis may be limited by small differences in the patient positioning between visits. A potential data-driven approach, however, that leverages cross-sectional data, i.e., data from different participants with different age, to synthetically age or rejuvenate a real image could boost the efficient use of such a large cohort. In recent years, models based on generative adversarial networks (GANs) (9) have been proposed for this task.

Deep learning models for synthesizing an aged version of an input image have been proposed for several applications, but especially for face aging. For example, Zhang et al. (10) was one of the first works to propose learning a manifold of images, *via* cross-sectional data, that can be navigated for increasing or decreasing the apparent age of a human face. The authors used an autoencoder and adversarial training to generate photorealistic images of a younger and older version of an input face. Later, Liu et al. (11) used a GAN-based model that included also attribute conditioning such as race or sex to enforce attribute preservation, highlighting the importance of covariates for the modeling. Contrary to Zhang et al. (10), their model had a last layer responsible for fusing the input image with the generated features, so that the model did not need to generate the whole image as output.

In medical imaging, a recent study by Xia et al. (12) proposed a conditional GAN (cGAN) (13) that considered age and disease status for generating an aged brain MRI using only cross-sectional data. Other works have modeled the changes in the brain due to aging with autoencoders and adversarial training (14, 15) or with normalizing flows (16), although the image quality was worse in these cases. Finally, cGANs have also been applied recently to synthesize future fundus images given a lesion probability map and a vessel segmentation (17).

In this work, we propose a conditional generative model for extracting longitudinal patterns related to aging from cross-sectional data and apply it to cardiac imaging for the first time, to the best of our knowledge. Moreover, we demonstrate the model applicability for counter-balancing biased datasets with respect to age. Finally, we analyze the modeling ability of the proposed approach for two other tasks: apparent body mass index modification and end-systolic phase synthesis from end-systolic frames.

Materials and methods

Dataset

For this work, MRI studies from the UK Biobank were used. These studies contain short- and long-axis views of 43,352 participants (including 23,508 female subjects). The participants were scanned at ages between 45 and 82 years old (mean age 64.1 ± 7.7). The scanner used was a MAGNETOM Aera, syngo MR D13A (Siemens, Erlangen, Germany) with a field strength of 1.5 Tesla [see (18) for further details about the imaging protocol]. Only the four chamber view was used in this work for simplicity and in order to include information from all heart chambers during the modeling. The end-diastolic phase for each subject was identified and used in this work, given that the model was two-dimensional. A total of 14,788 subjects were selected for training the generative models. No preprocessing was applied to the images.

Additionally, 764 ground-truth annotations of the four chamber long axis view performed by expert cardiologists from the Barts Heart Centre were made available to the authors from a previous work (19). The regions of interest annotated were the left and right ventricular cavities, the left ventricular myocardium and the left and right atria. We also delineated the aorta in 50 samples. Automatic segmentations for the rest of participants were generated for the four chambers, the myocardium and the aorta by training a U-Net model (20) (details in Supplementary material).

Conditional generative modeling

In order to generate synthetic images of the heart depending on a given covariate, a conditional generative adversarial network is proposed, as depicted in Figure 1. The two components of the model are a generator and a discriminator. Specifically, the generator is responsible for

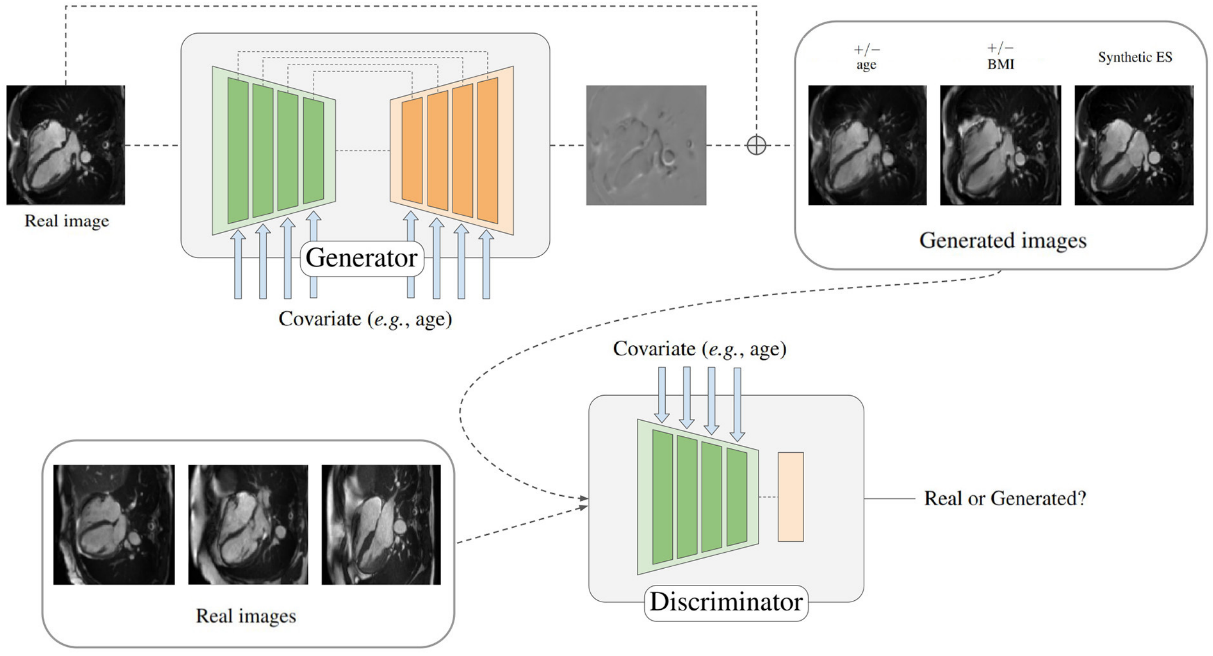


FIGURE 1
Depiction of the proposed model for generating synthetically aged and rejuvenated heart images. The covariate is combined with the model features by using conditional biasing. Heart scans reproduced by kind permission of UK Biobank ©.

creating the mapping that will be applied to the input image when conditioned on different covariates to obtain a target image, while the discriminator is trained to tell apart real and synthesized images given some covariates.

The architecture of the generator follows a typical U-Net (20) encoding-decoding scheme where each layer is composed of stacks of two residual blocks with an intermediate attention block. It is based on the generator used in recent state-of-the-art diffusion models (21, 22) that was first introduced by Ho et al. (23). The discriminator consists of an encoder, just as the one used for the generator, and an adaptive pooling layer. Each residual block along the networks is conditioned on the input variable by using *conditional biasing*, i.e., by transforming the variable into a vector of varying dimension and adding one value per intermediate feature channel prior to a group normalization step. Figure 2 depicts the conditional biasing mechanism in more detail. This type of conditioning allows for a better conservation of the input information throughout the network by consistently introducing the conditional variable on each layer. Additionally, the network is able to fit the different parameters used to compute the conditioning vector on each layer separately, enhancing the ability of the model to learn different distributions at different resolutions.

As covariates, age and body mass index (BMI) were considered for two separate tasks. The generator was conditioned with the difference between the age (respectively BMI) of the input image and the desired output age (respectively BMI). The discriminator, however, was conditioned on the actual age (respectively BMI) of the input image (real or synthesized). The covariate was specified to the model using the Transformer sinusoidal embedding (24).

Training details

The underlying framework for training the model relied on the Wasserstein-GAN with a gradient penalty term (WGAN-GP) (25, 26), that achieved better results than usual GANs, by minimizing the Wasserstein-1 distance (also called Earth-Mover distance).

The generator (G) and the discriminator (D) were trained using the adversarial objective loss for WGAN-GP:

$$\mathcal{L}_{\text{WGAN-GP}} = \mathbb{E}_{\tilde{\mathbf{x}}\sim P_{gen.}}[D(\tilde{\mathbf{x}},\mathbf{a}_t)] - \mathbb{E}_{\mathbf{x}\sim P_{real}}[D(\mathbf{x},\mathbf{a}_t)] \\ +\lambda_{GP}\mathbb{E}_{\hat{\mathbf{x}}\sim P_{\hat{\mathbf{x}}}}[(||\nabla^2_{\hat{\mathbf{x}}}D(\hat{\mathbf{x}},\mathbf{a}_t)||-1)^2], \tag{1}$$

where $\mathbf{a}_s$ and $\mathbf{a}_t$ stand for source and target age, respectively, $\mathbf{x}$ is the input image, $\tilde{\mathbf{x}} = G(\mathbf{x},\mathbf{a}_d)$ is the generated sample with age gap $\mathbf{a}_d = \mathbf{a}_t - \mathbf{a}_s$, and $\hat{\mathbf{x}} = \epsilon\mathbf{x} + (1-\epsilon)\tilde{\mathbf{x}}$, with $\epsilon \sim \mathcal{U}(0,1)$, is a random point along the line connecting the real and generated samples. P_{real} and P_{gen} represent the distributions of real and generated images, respectively. The gradient penalty factor, λ_{GP}, was set to 10 in all experiments following the original work by Gulrajani et al. (26).

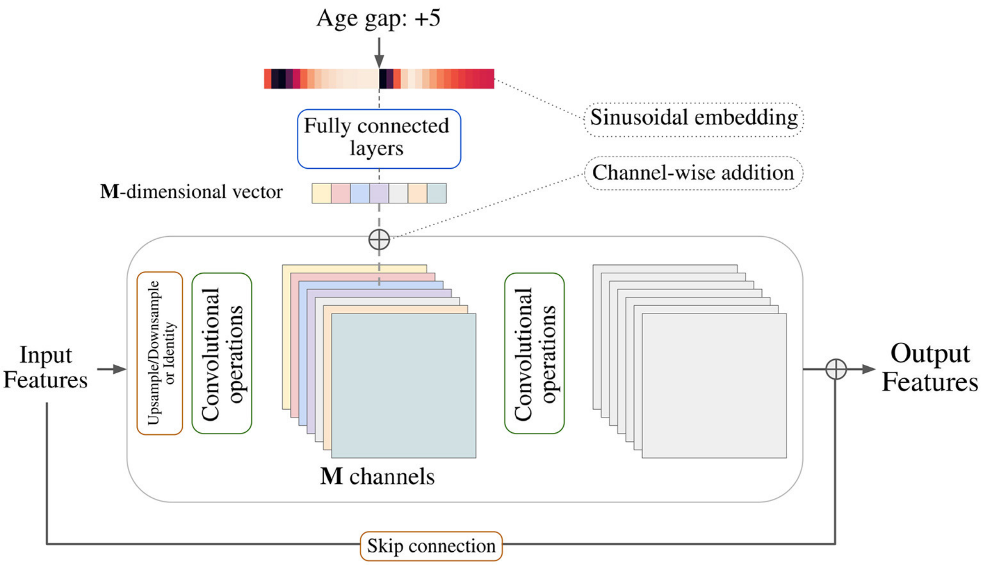


FIGURE 2
Depiction of a general residual block and the conditional biasing mechanism used to pass covariates to the model. A sinusoidal embedding is applied to transform the scalar covariate into a vector that is later converted to a vector with the same dimensionality as the number of intermediate feature channels. Each value in the vector is then added to its corresponding channel in order.

In addition to the adversarial loss, a cycle-consistency term was considered to enforce the reconstruction of the original image after two generator steps, one for aging (rejuvenating) and one for rejuvenating (aging) the subject back to the original state. In detail, the difference between the transformed image (after adding $\mathbf{a}_d$ years) and the reconstructed image (after subtracting $\mathbf{a}_d$ years to the transformed image) was minimized. This term is formally written as

$$\mathcal{L}_{cc} = \mathop{\mathbb{E}}_{\mathbf{x}\sim P_{real}} [||\mathbf{x} - G(G(\mathbf{x}, \mathbf{a}_d), -\mathbf{a}_d)||_1]. \tag{2}$$

Overall, the final objective loss was

$$L = \min_G \max_D \left(\mathcal{L}_{\text{WGAN-GP}} + \lambda_{cc}\mathcal{L}_{cc}\right), \tag{3}$$

where the weight λ_{cc} was empirically set to 1 based on model performance.

During training, WGAN-GP requires the discriminator performance to be close to optimal. For this reason, the first 20 epochs were used as a warm-up period, and the discriminator was updated 50 times for every generator update. For the remaining epochs, the discriminator was updated five times for every generator update. The AdamW (27) optimizer was used for both networks with a learning rate and weight decay of 10^{-4} and first and second moments equal to 0.9 and 0.999, respectively. Data augmentation was used to increase the variability in the input images appearance. The transformations considered were random bias field addition, random histogram shift and random contrast adjustment (28). The images were cropped along the x axis by 90 pixels, resized to 128^2 pixel size and the intensities rescaled to the $[0, 1]$ range. The generated mapping is an array of shape 128^2 with values clipped to the range $[-1, 1]$ (the maximum modification allowed to the input image). After the addition of the mapping, the resulting image was again clipped to $[0, 1]$. The whole training process took ∼90 h in a Nvidia 3090 GPU for 300 epochs and with a batch size of 12 images. PyTorch (version 1.10.0) was used for the implementation.

Results

Given the lack of real longitudinal data with a time span between visits larger than 10 years and the added factor of morphological variations attributed to different patient positioning, we propose to evaluate the current model using two proxy approaches that circumvent the limited time span and the potential disalignment between scans. First, we assess the resulting synthetic images *via* age accuracy and image quality, and compare the proposed model against two baselines. Second, we train age regressors with an imbalanced dataset augmented with synthetic samples. Moreover, in order to demonstrate the modeling capabilities of the current approach, two alternative tasks with an easier interpretation are considered: (1) BMI modification and (2) end-diastolic to end-systolic phase transformation.

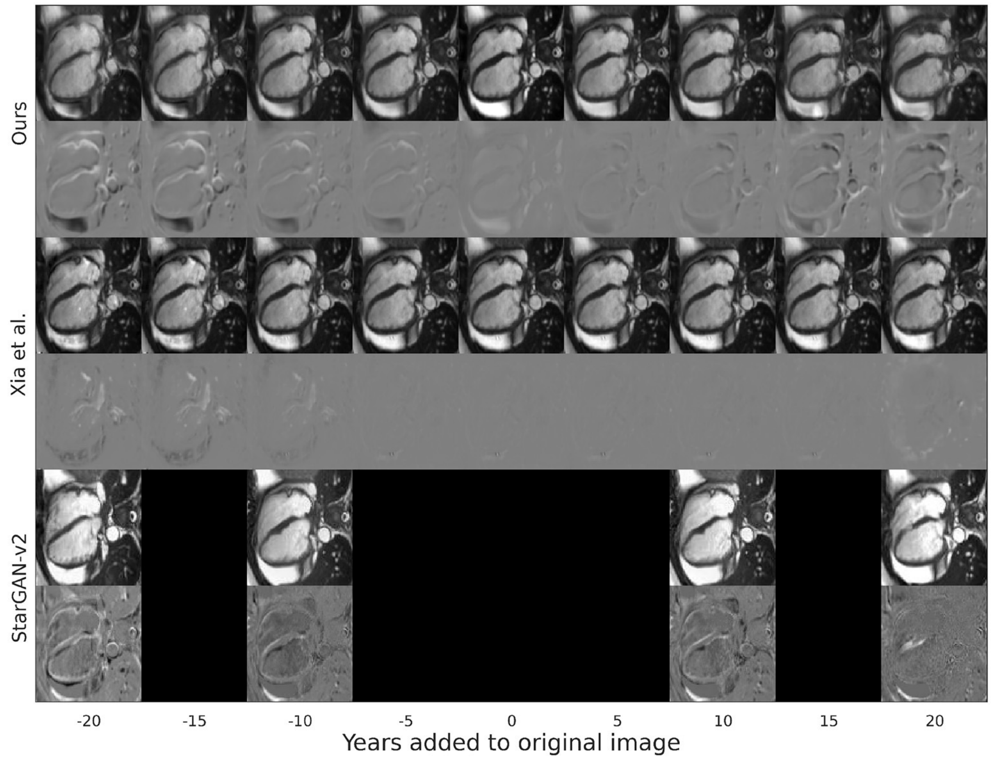


FIGURE 3
Synthetic aged and rejuvenated images for a randomly selected subject from the test set for the current proposal and the two baselines. Each pair of rows contain the generated image and the mapping applied to the original image to obtain it. The column with age gap equal to zero represents the reconstructed image. Reproduced by kind permission of UK Biobank ©.

TABLE 1 Mean absolute error (MAE) for age prediction of images generated synthetically from a random testing set of 907 subjects with varied age and grouped in age gaps of 5 years.

	MAE for predicted age (|predicted age − target age|) ↓							
Age gap	**−20**	**−15**	**−10**	**−5**	**5**	**10**	**15**	**20**
Ours	**$7.2_{4.3}$**	**$5.8_{3.8}$**	**$6.0_{4.0}$**	**$4.7_{3.6}$**	**$4.8_{3.4}$**	**$5.8_{3.8}$**	**$5.8_{3.7}$**	**$6.2_{3.6}$**
Xia et al. (12)	$12.8_{4.7}$	$9.4_{4.8}$	$6.8_{4.4}$	$5.7_{4.2}$	$5.7_{3.9}$	$8.7_{4.7}$	$10.6_{4.9}$	$12.8_{5.0}$
Zero order	$16.4_{4.3}$	$12.6_{4.7}$	$8.7_{4.7}$	$5.6_{4.1}$	$6.0_{4.1}$	$9.2_{4.8}$	$12.8_{4.9}$	$16.6_{4.7}$

Results are presented for models with the ability to generate a variable age gap: our proposal and the adapted model by Xia et al. (12). The Zero order shows the prediction error for unmodified images. The best results are shown in bold face and standard deviations as subscripts.

Qualitative results

At a qualitative level, as presented in Figure 3, the changes of our proposal tend to be more localized in space than the modifications introduced by the other baseline models. These modifications focus mostly on the interventricular septum and the aorta with opposed transformations for opposite age gaps. In detail, for increased age, the interventricular septum is enlarged toward the LV cavity and the aorta is enlarged. Finally, although most of the changes occur in the heart, some modifications are observed in surrounding areas.

Quantitative assessment of generated images

Assessment *via* predicted age

The apparent age of synthesized images was assessed using a pre-trained ResNet18 (29) age regressor (MAE: 4.6 ± 3.2 years for males and 3.9 ± 3.1 years for females). The hypothesis was that images aged (respectively rejuvenated) by the model should have a target age greater (respectively lower) than the original images. Tables 1, 3 show the mean absolute error (MAE) for age predictions using the pre-trained regressor when tested on the

TABLE 2 Image quality in terms of Fréchet inception distance (FID) and peak signal-to-noise ratio (PSNR) for images generated synthetically from a random testing set of 907 subjects with varied age and grouped in age gaps of 5 years.

	FID ↓								PSNR (dB) ↑							
Age gap	**−20**	**−15**	**−10**	**−5**	**5**	**10**	**15**	**20**	**−20**	**−15**	**−10**	**−5**	**5**	**10**	**15**	**20**
Ours	1.0	1.0	1.0	0.9	1.1	1.2	1.8	2.2	19.7	21.2	24.4	25.9	26.6	25.0	21.2	19.2
Xia et al. (12)	0.4	0.4	0.2	0.0	0.0	0.0	0.2	0.5	28.0	28.4	30.7	49.8	63.8	58.7	44.0	31.2

Results are presented for models with the ability to generate a variable age gap: our proposal and the adapted model by Xia et al. (12).

TABLE 3 Mean absolute error (MAE) for age prediction and image quality metrics (FID and PSNR) for synthetically generated images from a subset of subjects with 60 and 70 years old (for an age gap of 10) and with 55 and 75 years old (for an age gap of 20).

	MAE for predicted age ↓				FID ↓				PSNR (dB) ↑			
Age gap	**−20**	**−10**	**10**	**20**	**−20**	**−10**	**10**	**20**	**−20**	**−10**	**10**	**20**
Ours	$4.7_{4.1}$	$\mathbf{5.0_{3.1}}$	$\mathbf{4.1_{3.3}}$	$5.4_{3.0}$	0.9	1.2	1.2	1.5	19.4	24.5	25.0	19.1
Xia et al. (12)	$10.8_{4.1}$	$5.6_{3.2}$	$7.8_{4.4}$	$12.6_{4.6}$	0.4	0.2	0.0	0.5	28.1	29.7	58.8	30.9
StarGAN-v2	$\mathbf{4.2_{4.3}}$	$5.2_{3.0}$	$9.2_{4.0}$	$\mathbf{5.1_{4.7}}$	12.5	5.0	12.4	9.9	9.8	10.3	10.0	9.7
Zero order	$16.4_{4.3}$	$8.7_{4.7}$	$9.2_{4.8}$	$16.6_{4.7}$	–	–	–	–	–	–	–	–

Best results are shown in bold face and standard deviations as subscripts.

images generated by the proposed approach and compare it to a model adapted from the work of Xia et al. (12) that generates images with a controllable age gap and to StarGAN-v2 (30) that transforms images between fixed age gaps (gaps of 10 and 20 years), respectively.

The last row (Zero order) corresponds to the results obtained when the original images are not modified at all, i.e., the predicted age is always the same but the target age changes according to the desired age gap. We find that the proposed model with residual and attentional blocks outperforms the model based on the work by Xia et al. (12) and it obtains comparable results to StarGAN-v2, while StarGAN-v2 can only translate images between fixed domains. The proposed approach presents a significant improvement in MAE when compared to the Zero order.

Assessment *via* image quality

Image quality was assessed *via* the Fréchet inception distance (FID) (31) and the peak signal-to-noise ratio (PSNR). The FID gives a sense of how different two datasets are in terms of features extracted from a pre-trained deep learning model (better quality corresponds to lower values). An InceptionV3 model (32) was trained on the UK Biobank for this purpose (further details in Supplementary material). PSNR, on the other hand, evaluates the amount of corruption or noise in the generated images by directly comparing them to the original ones (better quality corresponds to higher values). As observed in Table 2, both metrics were coherent and showed better image quality for the model based on the work on Xia et al. (12), while StarGAN-v2 obtained images with significantly worse image quality (see Table 3).

This can be attributed to Xia et al.'s model introducing less modifications in the image (see Figure 3), resulting in synthetic images that are more similar to the original images but that do not represent the target age accurately as shown when computing the predicted age error in Table 1. On the other hand, StarGAN-v2 is introducing more modifications in the image, degrading its quality (see Figure 3), while maintaining a competitive predicted age error in Table 3.

Volumetric analysis

In order to quantify the specific changes performed in the heart, the LV size, the interventricular septum width and the ejection fraction are derived from automatically generated segmentations of the original and synthesized images (more details about the segmentation model are provided in Supplementary material). The normalized variation for these metrics after adding or subtracting 20 years to the original subjects is presented in Figure 4 separated by sex.

As observed in the figure, the model shows a clear tendency for decreased LV size with age, going from a 5% increase for rejuvenated subjects to a 5% decrease for aged subjects (with respect to the original sample) for both sexes. With respect to the interventricular septum average width, a decrease around 25% for males and 15% for females is observed for rejuvenated images, while the aged images also show a decrease

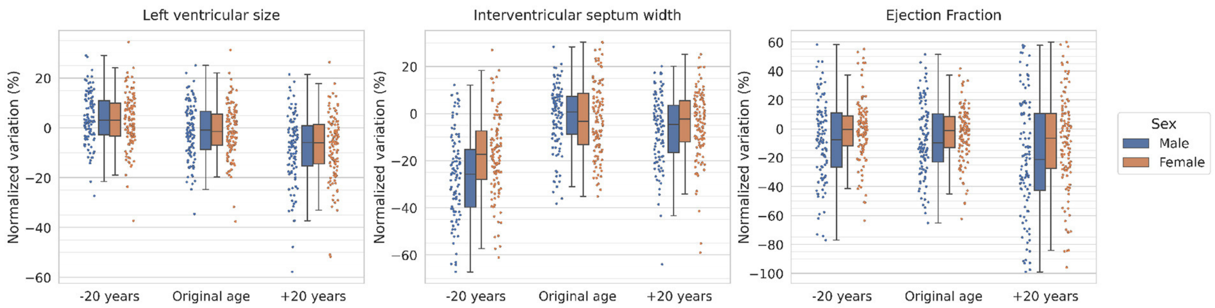


FIGURE 4
Normalized volumetric variations of synthesized images for the left ventricle, the interventricular septum and the ejection fraction. Every dot represents a subject and the boxes represent the interquartile range. The normalized variation is computed as the normalized difference of the selected metric with respect to the mean value for the original distribution (i.e., the "Original age" distribution). The results are obtained for a subset of participants aged 60.

of around 5–10% for both sexes. Finally, the ejection fraction shows a similar distribution for rejuvenated and original images while the aged subjects present a larger variability and an overall mean decrease of around 20% for males and 5% for females.

Synthetic images as data augmentation

Finally, in order to assess the utility of generated images for data augmentation, several age regressors (ResNet18) were trained with two datasets created from a new sample of 1,000 subjects with a particular age imbalance. Dataset one (D1) consisted of an imbalanced dataset with 90% of subjects younger than 70 years old. Dataset two (D2) was constructed to manifest an imbalance for younger patients, with 90% of the subjects being older than 60 years. These datasets were gradually augmented with 1, 5, 10, and 25% of synthetically aged (for D1) or rejuvenated (for D2) subjects. The results are presented in Table 4. A clear reduction in prediction error is observed when using synthetically age (or rejuvenated) subjects and the error when using 10 or 25% of synthetic images is comparable to the error obtained with a balanced dataset (12.7 ± 8.9).

TABLE 4 Mean absolute error (MAE) of ResNet18 age regressors when trained with two imbalanced datasets and different proportions of added synthetic images.

	MAE ↓				
	0%	**1%**	**5%**	**10%**	**25%**
D1 (10% of older subjects)	$14.5_{9.0}$	$13.3_{8.6}$	$14.2_{9.3}$	$12.8_{8.6}$	$11.0_{7.6}$
D2 (10% of younger subjects)	$18.0_{9.7}$	$17.0_{9.7}$	$17.8_{10.0}$	$15.4_{9.4}$	$13.9_{9.2}$
Balanced dataset	$12.7_{8.9}$	–	–	–	–

Standard deviations are presented as subscripts.

Alternative tasks

In order to showcase the capabilities of the proposed approach, the same model is used for modifying the BMI of an input patient and for transforming an image in the end-diastole (ED) time frame to an image in end-systole (ES).

Table 5 compares the prediction error (MAE) between apparent BMI and the target BMI, as obtained from a pre-trained ResNet18 BMI regressor (MAE 1.4 ± 1.1 for males and 1.6 ± 1.4 for females), for images generated with the proposed model and for images that were not modified at all (Zero order). The MAE increases slightly with higher BMI differences between input and synthetic images, although it shows a significant improvement as compared to the Zero order error, indicating a relative increase (respectively decrease) for positive (respectively negative) gaps in the apparent BMI of the subject.

With regards to the transformation of cardiac time frames, the model obtained a root mean square error between generated images and the real ES frames of 0.06 (±0.01), when compared at the whole image level. Figure 5 shows some qualitative results obtained for this task that include the generated mapping and the pixel-wise absolute difference between the generated frames and the real ones. As observed in the figure, the model captured the thickening of the myocardium, the contraction of the right ventricle as well as the smaller changes in size in the atria between ED and ES. However, several hallucinations were also introduced (highlighted with orange arrows) by the model that are not clinically accurate.

Discussion

A conditional generative model is proposed that allows for the modification of a cardiac image in two directions, i.e., for

TABLE 5 Mean absolute error (MAE) for apparent BMI of generated images, obtained from a pre-trained ResNet18 BMI regressor.

	MAE ↓										
BMI gap	**−8**	**−6**	**−4**	**−2**	**−1**	**0**	**1**	**2**	**4**	**6**	**8**
Ours	$2.3_{1.6}$	$2.0_{1.5}$	$1.8_{1.4}$	$1.7_{1.3}$	$1.6_{1.3}$	$1.5_{1.2}$	$1.5_{1.3}$	$1.7_{1.4}$	$2.1_{1.6}$	$2.3_{1.7}$	$2.7_{1.8}$
Zero order	$7.2_{2.0}$	$5.6_{1.9}$	$3.8_{1.8}$	$2.2_{1.4}$	$1.7_{1.3}$	$1.5_{1.2}$	$1.7_{1.4}$	$2.3_{1.6}$	$4.1_{1.8}$	$6.0_{1.8}$	$7.9_{1.7}$

Standard deviations are presented as subscripts.

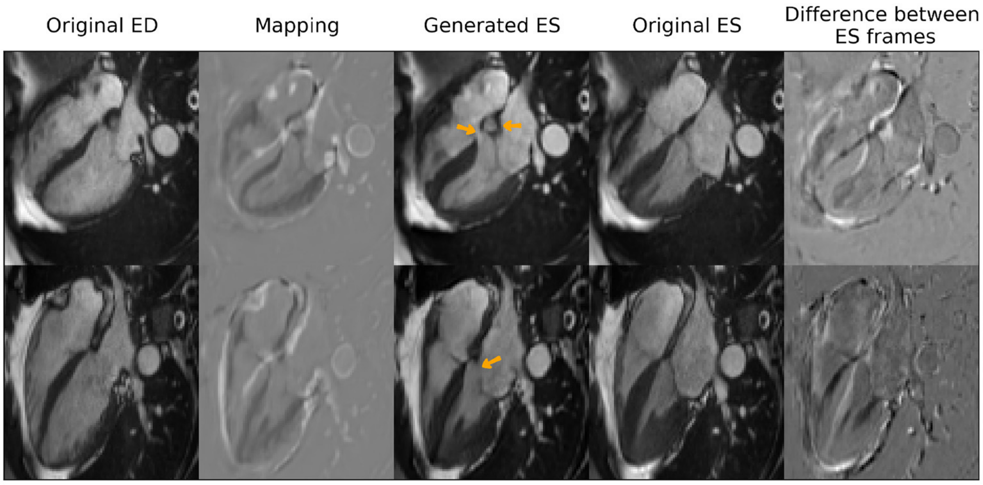


FIGURE 5
Qualitative results for end-systole (ES) frame generation from end-diastole (ED) frames. Orange arrows highlight clinical inaccuracies of the generated images such as incomplete interventricular septum or mitral valve or an extra "blob" in between the atria. Reproduced by kind permission of UK Biobank ©.

increased and decreased age. This is the first approach, to the best of our knowledge, for modeling the aging heart trained only on cross-sectional data. Realistic modifications are obtained without the need of complicated pre-processing steps, such as image registration or histogram matching, or of manual subdivision of the dataset in age groups. The accuracy and image quality of the results is comparable to state-of-the-art GAN methods, such as StarGAN-v2 (30), while the current model allows for a controllable target age and does not need to train several models for the aging task.

The results obtained for increasing age show in general a qualitative thickening of the interventricular septum, with the associated reduction of the LV cavity size, and an enlargement of the aorta. These changes are observed in the opposite direction for rejuvenated hearts. Quantitatively, there is a clear tendency for reduced LV size with age that is consistent with the literature (1, 2). The interventricular septum average width however, is reduced for both increased and decreased age, with the literature signaling this region as the most affected by the asymmetrical concentric LV hypertrophy observed with age (2). The ejection fraction suffers a small decrease in the mean value, while the literature states that it is preserved with age (1), and the distribution becomes wider with age, which might be related to a potential larger group of pathological subjects for increased age and the introduction of uncontrolled bias in the model which should be investigated in future works. Finally, an increased diameter is also observed in the aorta with increased age in both sexes according to the literature (33). Notably, the aorta and the interventricular septum are important areas also for age predictors based on deep learning, according to a recent study (34).

A potential application for this method has been showcased by counter-balancing biased datasets which improves the accuracy of age regression models trained on them. Recent works in the literature (35, 36) also demonstrate the feasibility of synthetic data augmentation. Such augmentation may be especially interesting for counter-balancing datasets with an age bias between healthy and diseased patients or when there are simply not enough control subjects.

Finally, two alternative tasks are presented to demonstrate the model ability to synthesize images given cross-sectional data. On one hand, the model was able to successfully increase and decrease the apparent BMI of subjects in an analogous manner to the aging task. On the other hand, four chamber images in the ES cardiac frame were synthesized from ED frames with a relatively low error when compared to the real frame.

Limitations

The proposed model presents several limitations. First of all, the model has not been validated against real longitudinal data. This validation is particularly challenging, since repeated visits may have images acquired at slightly different slice positions which may then introduce changes in the heart morphology not associated with age. Additionally, the time gap between visits needs to be sufficiently large in order to observe visible changes, while current longitudinal datasets have a time span of $<$10 years between scans.

Secondly, the model is observed to produce images with clinical inaccuracies, as observed in Figure 5, where the synthetic images present an incomplete interventricular septum, an extra "blob" in between the right and left atria or a partially missing mitral valve. One possible approach to avoid incomplete structures is to use deformable maps, instead of modifying directly the pixel intensities, at the expense of preventing the appearance of new structures that are not present in the original image in the first place.

Conclusions

This work proposes a conditional generative model to extract longitudinal patterns using only cross-sectional data. Such a model may be applied to compare population groups, such as subjects following a specific treatment vs. a control group, that are spread in time in a cross-sectional dataset, without the need of acquiring a cost- and time-expensive longitudinal dataset. Moreover, we demonstrate the feasibility of using the generated images for dataset balancing.

Data availability statement

The code used in this study is available at the following link: https://github.com/vicmancr/CardiacAging. The data used in this study belongs to the UK Biobank initative and is available after a successful application process at https://www.ukbiobank.ac.uk/. The data is not open-sourced. It belongs to the UK Biobank intiative. They are the data holders.

Ethics statement

Ethical review and approval was not required for this study in accordance with the local legislation and institutional requirements. Written informed consent was not required for this study in accordance with the local legislation and institutional requirements.

Author contributions

VC, TX, XL, PS, ST, SS, and KL contributed to the conceptualization of this work. VC, TX, XL, PS, CM-I, and SS designed the methodology. VC developed the software tools necessary for preprocessing and analysing image files and for training the models. All authors reviewed the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research has been conducted using the UK Biobank Resource under Application Number 2964. This work was partly funded by the European Union's Horizon 2020 research and innovation program under grant agreement number 825903 (euCanSHare project). SP acknowledges the British Heart Foundation for funding the manual analysis to create a cardiovascular magnetic resonance imaging reference standard for the UK Biobank imaging resource in 5000 CMR scans (www.bhf.org.uk; PG/14/89/31194). SP acknowledges support from the National Institute for Health Research (NIHR) Biomedical Research Centre at Barts. SP acknowledges support from and from the SmartHeart EPSRC programme grant (www.nihr.ac.uk; EP/P001009/1). This article was supported by the London Medical Imaging and Artificial Intelligence Centre for Value Based Healthcare (AI4VBH). This work was supported by Health Data Research UK. ST acknowledges the support of Canon Medical and the Royal Academy of Engineering and the Research Chairs and Senior Research Fellowships scheme (grant RCSRF1819\8\25).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcvm.2022.983091/full#supplementary-material

References

1. Lakatta EG, Levy D. Arterial and cardiac aging: major shareholders in cardiovascular disease enterprises. *Circulation*. (2003) 107:346–54. doi: 10.1161/01.CIR.0000048894.99865.02

2. Steenman M, Lande G. Cardiac aging and heart disease in humans. *Biophys Rev*. (2017) 9:131–7. doi: 10.1007/s12551-017-0255-9

3. Obas V, Vasan RS. The aging heart. *Clin Sci*. (2018) 132:1367–82. doi: 10.1042/CS20171156

4. McManus DD, Xanthakis V, Sullivan LM, Zachariah J, Aragam J, Larson MG, et al. Longitudinal tracking of left atrial diameter over the adult life course: clinical correlates in the community. *Circulation*. (2010) 121:667–74. doi: 10.1161/CIRCULATIONAHA.109.885806

5. Keller KM, Howlett SE. Sex differences in the biology and pathology of the aging heart. *Can J Cardiol*. (2016) 32:1065–73. doi: 10.1016/j.cjca.2016.03.017

6. Benjamin EJ, Levy D, Anderson KM, Wolf PA, Plehn JF, Evans JC, et al. Determinants of Doppler indexes of left ventricular diastolic function in normal subjects (the Framingham heart study). *Am J Cardiol*. (1992) 70:508–15. doi: 10.1016/0002-9149(92)91199-E

7. Shock NW, Greulich RC, Costa PTJr, Andres R, Lakatta EG, Anernberg D, et al. *Normal Human Aging: The Baltimore Longitudinal Study on Aging*. Washington, DC: NIH Publication (1984).

8. Liu CY, Lai S, Kawel-Boehm N, Chahal H, Ambale-Venkatesh B, Lima JAC, et al. Healthy aging of the left ventricle in relationship to cardiovascular risk factors: the Multi-Ethnic Study of Atherosclerosis (MESA). *PLoS ONE*. (2017) 12:e179947. doi: 10.1371/journal.pone.0179947

9. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. *Commun. ACM*. (2020) 63:139–144. doi: 10.1145/3422622

10. Zhang Z, Song Y, Qi H. Age progression/regression by conditional adversarial autoencoder. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2017). p. 5810–8. doi: 10.1109/CVPR.2017.463

11. Liu Y, Li Q, Sun Z, Tan T. A 3 GAN: an attribute-aware attentive generative adversarial network for face aging. *IEEE Trans Inform Forens Secur*. (2021) 16:2776–90. doi: 10.1109/TIFS.2021.3065499

12. Xia T, Chartsias A, Wang C, Tsaftaris SA. Learning to synthesise the ageing brain without longitudinal data. *Med Image Anal*. (2021) 73:102169. doi: 10.1016/j.media.2021.102169

13. Mirza M, Osindero S. Conditional generative adversarial nets. *arXiv preprint arXiv:14111784*. (2014).

14. Ravi D, Alexander DC, Oxtoby NP. Degenerative adversarial neuroimage nets: generating images that mimic disease progression. In: *Lecture Notes in Computer Science*. Springer International Publishing (2019). p. 164–72. doi: 10.1007/978-3-030-32248-9_19

15. Ravi D, Blumberg SB, Ingala S, Barkhof F, Alexander DC, Oxtoby NP. Degenerative adversarial neuroimage nets for brain scan simulations: application in ageing and dementia. *Med Image Anal*. (2022) 75:102257. doi: 10.1016/j.media.2021.102257

16. Wilms M, Bannister JJ, Mouches P, MacDonald ME, Rajashekar D, Langner S, et al. Bidirectional modeling and analysis of brain aging with normalizing flows. In: Kia SM, Mohy-ud-Din H, Abdulkadir A, Bass C, Habes M, Rondina JM, Tax C, Wang H, Wolfers T, Rathore S, Ingalhalikar M, editors. *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-Oncology*. Lima: Springer International Publishing (2020). p. 23–33. doi: 10.1007/978-3-030-66843-3_3

17. Ahn S, Pham QTM, Shin J, Song SJ. Future image synthesis for diabetic retinopathy based on the lesion occurrence probability. *Electronics*. (2021) 10:726. doi: 10.3390/electronics10060726

18. Petersen SE, Matthews PM, Francis JM, Robson MD, Zemrak F, Boubertakh R, et al. UK Biobank's cardiovascular magnetic resonance protocol. *J Cardiovasc Magn Reson*. (2015) 18:8. doi: 10.1186/s12968-016-0227-4

19. Petersen SE, Aung N, Sanghvi MM, Zemrak F, Fung K, Paiva JM, et al. Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort. *J Cardiovasc Magn Reson*. (2017) 19:18. doi: 10.1186/s12968-017-0327-9

20. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: navar N, Hornegger J, Wells WM, Frangi AF, editors. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 Lecture Notes in Computer Science*. Munich: Springer International Publishing (2015). p. 234–41. doi: 10.1007/978-3-319-24574-4_28

21. Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis. *Adv Neural Inform Process Syst*. (2021) 34:8780–94. Available online at: https://proceedings.neurips.cc/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf

22. Nichol A, Dhariwal P, Ramesh A, Shyam P, Mishkin P, McGrew B, et al. Glide: towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:211210741*. (2021).

23. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. *Adv Neural Inform Process Syst*. (2020) 33:6840–51. Available online at: https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

24. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: *Advances in Neural Information Processing Systems*. (2017). p. 30.

25. Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: *International Conference on Machine Learning*. (2017). p. 214–23.

26. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of wasserstein GANs. In: *Advances in Neural Information Processing Systems*. (2017). p. 30.

27. Loshchilov I, Hutter F. Decoupled weight decay regularization. *arXiv preprint arXiv:171105101*. (2017).

28. MONAI Consortium. *MONAI: Medical Open Network for AI*. Zenodo (2022).

29. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2016). p. 770–8. doi: 10.1109/CVPR.2016.90

30. Choi Y, Uh Y, Yoo J, Ha JW. StarGAN v2: Diverse image synthesis for multiple domains. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. (2020) p. 8185–94. doi: 10.1109/CVPR42600.2020.00821

31. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. (2017).

32. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2016). p. 2818–26. doi: 10.1109/CVPR.2016.308

33. Komutrattananont P, Mahakkanukrauh P, Das S. Morphology of the human aorta and age-related changes: anatomical facts. *Anat Cell Biol*. (2019) 52:109. doi: 10.5115/acb.2019.52.2.109

34. Goallec AL, Prost JB, Collin S, Diai S, Vincent T, Patel CJ. Dissecting heart age using cardiac magnetic resonance videos, electrocardiograms, biobanks, and deep learning. *medRixv*. (2021). doi: 10.1101/2021.06.09.21258645

35. Pombo G, Gray R, Cardoso J, Ourselin S, Rees G, Ashburner J, et al. Equitable modelling of brain imaging by counterfactual augmentation with morphologically constrained 3D deep generative models. *arXiv [Preprint]*. (2021). arXiv: 2111.14923. Available online at: https://arxiv.org/pdf/2111.14923.pdf

36. Li R, Bastiani M, Auer D, Wagner C, Chen X. Image augmentation using a task guided generative adversarial network for age estimation on brain MRI. In: *Medical Image Understanding and Analysis*. Springer International Publishing (2021). p. 350–60. doi: 10.1007/978-3-030-80432-9_27

frontiers | Frontiers in Cardiovascular Medicine

TYPE Review
PUBLISHED 08 November 2022
DOI 10.3389/fcvm.2022.1016032

Clinician's guide to trustworthy and responsible artificial intelligence in cardiovascular imaging

Liliana Szabo[1,2,3]*, Zahra Raisi-Estabragh[1,2], Ahmed Salih[1,2], Celeste McCracken[4], Esmeralda Ruiz Pujadas[5], Polyxeni Gkontra[5], Mate Kiss[6], Pal Maurovich-Horvath[7], Hajnalka Vago[3], Bela Merkely[3], Aaron M. Lee[1,2], Karim Lekadir[5] and Steffen E. Petersen[1,2,8,9]

[1]William Harvey Research Institute, NIHR Barts Biomedical Research Centre, Queen Mary University of London, London, United Kingdom, [2]Barts Heart Centre, St Bartholomew's Hospital, Barts Health NHS Trust, London, United Kingdom, [3]Semmelweis University Heart and Vascular Center, Budapest, Hungary, [4]Division of Cardiovascular Medicine, Radcliffe Department of Medicine, National Institute for Health Research Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, University of Oxford, Oxford, United Kingdom, [5]Departament de Matemàtiques i Informàtica, Artificial Intelligence in Medicine Lab (BCN-AIM), Universitat de Barcelona, Barcelona, Spain, [6]Siemens Healthcare Hungary, Budapest, Hungary, [7]Department of Radiology, Medical Imaging Centre, Semmelweis University, Budapest, Hungary, [8]Health Data Research UK, London, United Kingdom, [9]Alan Turing Institute, London, United Kingdom

OPEN ACCESS

EDITED BY
Hui Xue,
National Heart, Lung, and Blood Institute (NIH), United States

REVIEWED BY
James Howard,
Imperial College London,
United Kingdom
Graham Cole,
Imperial College London,
United Kingdom

*CORRESPONDENCE
Liliana Szabo
sz.liliana.e@gmail.com;
l.szabo@qmul.ac.uk

SPECIALTY SECTION
This article was submitted to Cardiovascular Imaging, a section of the journal Frontiers in Cardiovascular Medicine

RECEIVED 10 August 2022
ACCEPTED 11 October 2022
PUBLISHED 08 November 2022

CITATION
Szabo L, Raisi-Estabragh Z, Salih A, McCracken C, Ruiz Pujadas E, Gkontra P, Kiss M, Maurovich-Horvath P, Vago H, Merkely B, Lee AM, Lekadir K and Petersen SE (2022) Clinician's guide to trustworthy and responsible artificial intelligence in cardiovascular imaging. *Front. Cardiovasc. Med.* 9:1016032. doi: 10.3389/fcvm.2022.1016032

A growing number of artificial intelligence (AI)-based systems are being proposed and developed in cardiology, driven by the increasing need to deal with the vast amount of clinical and imaging data with the ultimate aim of advancing patient care, diagnosis and prognostication. However, there is a critical gap between the development and clinical deployment of AI tools. A key consideration for implementing AI tools into real-life clinical practice is their "trustworthiness" by end-users. Namely, we must ensure that AI systems can be trusted and adopted by all parties involved, including clinicians and patients. Here we provide a summary of the concepts involved in developing a "trustworthy AI system." We describe the main risks of AI applications and potential mitigation techniques for the wider application of these promising techniques in the context of cardiovascular imaging. Finally, we show why trustworthy AI concepts are important governing forces of AI development.

KEYWORDS

artificial intelligence, cardiovascular imaging, machine learning (ML), trustworthiness, AI risk

Introduction

In recent years, several artificial intelligence (AI) based systems have been developed in cardiology. This trend is driven by the increasing need to deal with the vast amount of clinical and imaging data produced in the field and with the ultimate aim to advance patient care, diagnosis and prognostication (1, 2). It is not a question anymore whether AI will transform healthcare but rather how it will do so (3). Transformative measures have already impacted many areas of cardiovascular medicine, from smart devices promising to diagnose arrhythmias based on single-lead ECG (4) to automatic image segmentation tools shortening manual image analysis (5, 6). However, there is a critical gap between the development and deployment of AI tools. To date only 24 AI-driven cardiovascular imaging products have received FDA approval (7), suggesting there remain critical challenges in building and implementing these models into everyday practice.

It is easy to scare away busy clinicians with endless legal documentation and specialized terms from philosophy, law and data science. On the other hand, expecting the data science community to be up to date with their field, understand complex medical concepts and consider the ethical ramifications of AI is the recipe for serious unintended consequences (8). Indeed, the discussion around the ethical issues of AI should be inclusive of all participants, from funding agencies to the patients.

The promise of AI revolutionizing cardiovascular imaging could not be delivered without achieving the trust of the end-users and patients. Currently, there are several ethical frameworks for AI applications. One of the most universal guideline was proposed by the European Commission in 2019 (9). This document provides a detailed technical summary and general guidance for dealing with the ethical questions of AI. However, it was written by senior data scientists, consequently does not focus on issues of healthcare applications (10). Indeed, to date, little is accessible to healthcare professionals without an in-depth understanding of the technical terms of the ethical questions embedded in AI applications. Notably, the document written by the European and North American Societies in Radiology detailing potential AI ethics issues can work as a primer for other societies in medicine (11). More recently, the first comprehensive guideline for assessing the trustworthiness of AI-based systems in medical imaging was developed, named FUTURE-AI (12). This technical framework promises to transform AI development in medical imaging and will help create an environment for safe clinical implementation of novel methods (https://future-ai.eu/).

In this narrative review, we aim to summarize the main risks of AI application and potential mitigation techniques in plain language. We provide an overview of ongoing efforts to improve the "trustworthiness" of AI in cardiovascular imaging. Finally, we aim to provide key questions to help initiate dialogue within research groups.

The basic concepts of AI

Several dedicated publications describe AI's definitions and main applications within cardiology in great detail (13–16). Here we restrict ourselves to those basic concepts essential for further discussion of AI trustworthiness.

AI is an umbrella term within data science, incorporating a wealth of models, use cases and aiding methodologies to mimic human thought processes and learning patterns (8). Within AI, the most commonly used models are *machine learning (ML)*-based in medical research; therefore, several important source documents handle AI and ML almost synonymously (1). An overly simplified definition of ML is computer algorithms that "learn" from data. ML methods use pre-processed (e.g., anthropometric data derived from patients) and raw data (e.g., raw imaging files). *Deep learning (DL)* is a subset of ML that deals with algorithms inspired by the structure and function of the human brain. DL algorithms use neural networks to transform the raw data into an abstract level, refine accuracy and adjust when encountering new data (17).

We can differentiate between supervised and unsupervised learning based on the type of data fed into an AI algorithm. In *supervised learning*, humans curate and label data before training, and the model is optimized for accuracy with known inputs and outputs. The following models are used for: classification (putting data into categories) and regression (predicting continuous variables within the concept of ML). On the other hand, *unsupervised learning* deals mainly with unlabelled data, with the ultimate goal of identifying novel patterns in a dataset such as clustering (14).

A critically important step in ML model development is a large and consistently labeled data set—the diverse quality of data and the inconsistent labeling could reduce the accuracy of AI model. Another important step is data splitting: datasets are generally split into training, validation and test sets. Training and validation sets used to train and fit the model, more specifically the validation provides an estimation of the model fit for model selection or tuning of parameters, whilst the test set is reserved to evaluate the final model (18). Given the degradation in performance reported for deep-learning algorithms for medical imaging, it is of paramount importance that the test set consists of independent cohorts to allow for *external validation*, a key requirement for ensuring the trustworthiness of AI systems (19, 20). Moreover, the external validation should be performed by independent parties to ensure objectiveness. Please note, that validation in the original dataset is not synonymous with external validation, which is performed on a separate dataset.

The concept of trust and trustworthiness of AI in medicine

It is easy to get lost in a philosophical discussion about how to define trust or if it is even possible outside the human realm (21, 22). From a practical standpoint, these questions are confusing rather than helpful. For decades AI was part of the scientific discussion, existing in research environments, and science fictions. Therefore, the question of whether to trust AI tools in healthcare was merely a discussion for scholars in the ethical and data science fields. However, with novel tools emerging daily, we are forced to reconsider the potential ramifications embedded in AI.

The questions we face today are highly practical and directly affect the field's development. Can we trust the CMR segmentation provided by the AI tool? Are we confident that the new artifact-removing algorithm does not mask any important clinical clue? Should we rely on the novel diagnosis support toolkit? What does it mean to trust the judgement of an automatic tool? How do we communicate the uncertainties embedded in a novel predictions score? Are we holding AI to a higher standard than clinical judgement based on intuition and experience?

As an example, left ventricular ejection fraction (LVEF) measured using echocardiography is a long-standing "trusted" parameter in cardiovascular medicine. Because with years of development, validation, and experience, we learned to comprehend the signals that link it to disease and outcome, and communicate the findings to the patients so that they trust their practitioners to understand echocardiography (23, 24). Although the information it provides is far from complete and prone to errors, the usefulness of knowing the EF of a patient in a clinical situation is beyond question; even when clinicans use eyeballing (25). On the other hand an AI application based on the idea that the human eye and brain can learn with experience how to estimate EF without measuring ventricular volumes and making calculations is more controversial, as this approach does not allow the revision of the ventricular contours in case of seemingly disparate results (26).

Wynants et al. reviewed multivariable COVID-19-related prediction models at the beginning of the pandemic. They found that the 232 models identified in the study all reported moderate to excellent predictive performance, but all were appraised to have a high or uncertain risk of bias owing to a combination of poor reporting and poor methodological conduct for participant selection, predictor description, and statistical methods used (27). The most sobering conclusion was that none of the proposed models proved to be of much help in clinical practice. The same conclusion was drawn from the investigation by the Alan Turing Institute (28) and others (29).

Main AI applications within cardiovascular imaging

Within cardiovascular imaging, the main areas of AI application are: (1) image acquisition and reconstruction—which helps to reduce the scan time, (2) improving the imaging workflow and efficiency of time-expensive tasks such as segmentation, (3) improving the diagnosis-making process, (4) evaluation of disease progression and prognosis, (5) assessment of treatment effectiveness, and (6) generation of new knowledge. Examples illustrating key areas of AI applications from non-invasive cardiovascular imaging is summarized in Table 1, further examples in can be found in dedicated publications (15, 16, 18, 49, 50).

There has been a steep increase in publications using ML in cardiovascular imaging in the past 5 years. This trend was driven by the increasing availability of high computational power, large datasets (16), and the discovery of the computational effectiveness of convolutional neural network architecture (AlexNet) (51).

It has been envisioned that AI tools will take over or at least substitute the work of radiologists and cardiovascular imagers to a great extent and consequently necessitate fewer human resources creating cheaper and more accurate care in the future (52). Roughly a decade into the area of accessible AI innovation, we can see that changes are less rapid, and the results are beneath our expectations (53). No segmentation is used unchecked, no diagnosis is made without human supervision and approval, and the need for well-trained imagers has increased (54). Notably, only a small proportion of the proposed methods, models and tools gain approval from the appropriate authorities (FDA or European Medicines Agency), and reach the clinical application stage. Should we then just conclude that AI is pointless and we must not use it? On the contrary, these experiences and setbacks should motivate the research into more robust AI models and rigorous validation standards. Only by learning from the critical issues raised by researchers and end-users can we move forward in the field of AI.

Unintended consequences of AI applications in cardiovascular imaging and mitigation strategies

To understand why AI applications are not approved and used to the rate it was predicted during the height of the ML hype in 2016, we have to look into the potential limitations of these tools. Here we provide an introduction to the main risks of AI applications within cardiovascular imaging: (1) lack of robustness and reliability causing patient harm, (2) issues of AI usability and the misuse of tools, (3) bias and lack of fairness within the AI application which can perpetuate existing inequities, (4) privacy and security issues, (5) lack of transparency, (6) gaps in explainability, (7) gaps in accountability, and (8) obstacles in implementation (Table 2). In each section, we describe the main attributes of each risk, provide relevant examples within cardiovascular imaging and illustrate potential mitigation strategies.

TABLE 1 Examples of AI applications from non-invasive cardiovascular imaging.

AI application	Purpose	Modality	References
Image acquisition and reconstruction	Improving image quality, decreasing image artifacts	CCTA CMR	Wolterink et al. (30) Oksuz et al. (31)
	Lowering radiation dose	CT	Benz et al. (32)
	Increasing imaging speed	CMR	Caballero et al. (33)
	Improving non-expert usage (e.g., view classification, automated planning)	Echocardiography CMR	Zhang et al. (34) Edalati et al. (35)
Improving the imaging workflow and efficiency of time-expensive tasks	Automatization of previously manual tasks for increased speed, effectiveness, and potentially improved standardization (e.g., image segmentation)	Echocardiography CCTA CMR	Leclerc et al. (36) Huang et al. (37) Bai et al. (5)
Diagnosis making	Supporting early diagnosis and timely treatement initiation or prevention	Echocardiography CCTA CMR	Sengupta et al. (38) de Vos et al. (39) Zhang et al. (40)
Disease prognostication	Improving the discrimination of high risk imaging features	Echocardiography CCTA CMR	Samad et al. (41) Patel et al. (42) Cheng et al. (43)
Assessment of treatment effectiveness	Monitoring response to medication, device therapy etc.	Echocardiography CCTA	Tokodi et al. (44) Queirós et al. (45)
Generation of new knowledge	Discovering new patterns, cardiovascular phenotypes or disease presentations (key role for unsupervised learning methods)	Echocardiography CCTA CMR	Casaclang-Verzosa et al. (46) Hoshino et al. (47) Zheng et al. (48)

CCTA, coronary computed tomography angiography; CMR, cardiovascular magnetic resonance.

Robustness and reliability

AI robustness is defined as the ability of a system to maintain its performance under changing conditions (56). The promise of a robust AI tool is that it can consistently deliver accurate outputs, even when it encounters unexpected or subquality data. When a model's functionality and accuracy change easily, it is considered "brittle" (8).

Medical imaging encapsulates a wealth of potential sources for AI brittleness (12):

(1) Heterogeneity within imaging types of equipment and vendors.

(2) Image acquisition heterogeneity within imaging centers and operators.

(3) Patient-related heterogeneity (including clinical status and anthropometric peculiarities).

(4) Data labeling and segmentation heterogeneity between annotators.

As mentioned above, ML algorithms play an increasingly important role in the image acquisition of all cardiovascular

TABLE 2 Questions to promote discussion of AI trustworthiness between clinicians and technical experts.

Robustness and reliability

- Did you perform any pre-processing that can potentially affect the reliability of your models?
- Did you use homogenous/ single center data OR heterogenous/multicenter data?

Are there any checkpoints for quality control in your pipeline?

Usability

- Do you have an understanding of the end-users needs in terms of the tool's interface?
- Does the implementation of your tool viable within the clinical workflow?

Bias and fairness

- What fairness means for your application?
- Are there any potential hidden sources of bias?
- Does the algorithm exhibit discrimination toward any group? Is it harmful or beneficial for the use case?
- Did you document and report these potential biases?

Security and safety

- Did you document potential risks of your AI tool? How do you communicate these?
- Does the implementation of your AI can potentially harm patients, worsen outcome or create security breach? If not, how do you know?

Transparency

- Did you document the characteristics of your dataset?
- Did you follow any relevant reporting guideline or checklist?

Explainability

- Do you know what level of explainability your end-users require?
- Can you explain, how your model reaches a certain decision?
- Did you explore complementary explainability methods?

Accountability

- What are the relevant regulations in terms of liability in your use case?
- Who is responsible for errors occurring during the clinical application of the AI tool?
- Who is monitoring the application and how frequently?

Summary of potential questions to support the discussion of AI trustworthiness between clinicians and technical experts adapted from Ammanath (8) and Lekadir et al. (55).

imaging modalities. However, these applications are not without certain limitations. For example, Antun et al. (57) highlighted possible sources of instability of deep learning algorithms at CMR reconstructions. The instabilities usually occur in several forms e.g., undetectable perturbations may result in artifacts in the reconstruction, or a small structure like tumors may not be captured in the reconstruction phase.

The potential brittleness of AI tools is also very well-illustrated by the recent developments in CMR image segmentation (58). Critically, DL-based segmentation tools are often trained and tested on images from single clinical centers, using one vendor with a well-defined protocol resulting in homogenous datasets (59, 60). Furthermore, CMR protocols across prominent multi-center cohort studies are also standardized, prohibiting wider generalizability (5, 61, 62). A notable effort to develop segmentation tools on more heterogeneous datasets to promote robust AI tool development is the Multi-Center, Multi-Vendor and Multi-Disease Cardiac Segmentation (M&Ms) Challenge (63). Investigators of the euCanSHare international project established an open-access CMR dataset (six centers, four vendors, and more than nine phenotype groups) to enable generalizable DL models in cardiac image segmentation. The Society of Cardiovascular Magnetic resonance Imaging (SCMR) registry (64) and Cardiac Atlas project (65) are also aimed at providing diverse databases for similar research ambitions. These efforts are still ongoing, and Campello et al. (63) noted that further research is necessary to improve generalizability toward different scanners or protocols.

Automated coronary computed tomography angiography (CCTA) segmentation faced similar challenges in the past decade. Although the accuracy of the CCTA plaque segmentation tools has been validated against the gold standard invasive methods, the interplatform reproducibility remains disputed (66, 67). Indeed, the time-consuming and labor-intensive nature of quantitative plaque assessment is still responsible for the frequent visual evaluation of coronary artery disease in clinical practice, despite some emerging solutions (68).

Apart from well-curated diverse datasets for benchmarking of segmentation algorithms and the development of novel segmentation tools, the reliability of the output is also a critical to the clinical implementation of these tools. Recently, automated quality control tools have been suggested in high-volume datasets where manual expert inspection is not achievable. Automated quality control tools utilizing different methods, such as Dice similarity coefficient, reverse classification accuracy (RCA) framework, and quality control-driven (QCD) framework, have been implemented within ventricular (69), T1 mapping (70), aortic (71), coronary, and pericardial fat segmentation (72).

AI robustness largely relies on the adaptability of a given model to changing circumstances. A segmentation tool might perform well in a given dataset of healthy hearts, but it might not directly translate into a heterogeneous dataset. The following concepts help promote robustness and reliability in medical imaging applications of AI:

(1) Heterogeneous training data (multi-center, multi-vendor, multiple diseases).
(2) Checking intra- and interobserver variability and whether automated AI tool difference lies within the observer variability.
(3) Applying well-established annotation with powerful annotation software.
(4) Image quality control (to identify artifacts within the data, applying algorithms which help to reduce artifacts).

(5) Applying image harmonization techniques (including the use of phantoms and dedicated harmonization tools such as histogram normalization).
(6) Applying feature harmonization techniques (using test-retest studies and feature selection methods to select stable, robust features for the models).
(7) Data augmentation.
(8) Uncertainty estimation [there is a variety of uncertainty quantification methods, including prediction intervals, Monte Carlo dropout, and ensembling; they are designed to pick up the distance of the new observation to observations the algorithm has already seen Kompa et al. (73)].

Potential issues that can arise during the assessment of robustness and clinical usability is well-illustrated by the adaptation of radiomics in cardiovascular imaging (74). Radiomics enable the extraction of voxel-level information from digital images, promising the quantitative description of tissue shape and texture. The utility of CT radiomics has been demonstrated in identifying vulnerable coronary atherosclerotic plaques (75–77) and linking pericoronary adipose tissue patterns to local inflammation (78, 79). CMR radiomics has also been shown to improve the discrimination of cardiomyopathies (80–82) and improve risk prediction among ST-elevation myocardial infarction patients (83, 84). Despite these advances, the clinical implementation of radiomics is in its infancy. The general critique of the technique lies in the poor repeatability of radiomics features. To improve radiomics usability in CMR, Raisi-Estabragh et al. (85) conducted a multi-center and multi-vendor test-retest study to evaluate the repeatability and reproducibility of CMR radiomics features using cine imaging. The authors reported variable levels of repeatability of the features, which are likely to be clinically relevant. To reduce the radiomics variability introduced by the acquisition center Campello et al. (86) evaluated several image- and feature-based normalization techniques. The authors demonstrated that ComBat, a feature-based harmonization technique, can remove center information, but this does not translate to better algorithmic generalization for classification. The best performing approach in this respect was piecewise linear histogram matching normalization.

Usability

Usability is defined as the extent to which an AI application can be utilized to achieve specific goals by specified users with effectiveness, efficiency and satisfaction (87). As the interaction between healthcare professionals and technology is increasingly important, more and more research effort is aimed at testing clinical usability. However, AI tools are barely tested regarding how they interact with clinicians, and most applications are still in “proof-of-concept” status (88). Key issues of usability include lack of a human-centered approach for the development of the AI technologies, e.g., lack of involvement of the end-user for the definition of the clinical requirements and of multi-stakeholder engagement throughout the production lifecycle.

Bias and fairness

In AI, defining bias and fairness is challenging due to the ever-changing applications we put AI to ISO/IEC TR 24027:2021 (89). Within the healthcare domain, fairness means that AI algorithms should be impartial and maintain the same performance when applied to similarly situated individuals (individual fairness) or different groups of individuals, including under-represented groups (group fairness) (12).

Until now, little data is available regarding the bias and fairness of algorithms in cardiovascular imaging, even though the phenomenon is well-known. As Rajkomar et al. summarized: any type of bias depicted within the dataset is learned and adapted into model performance (90). Overrepresentation of a certain group leads to data collection bias (18), as exemplified by Larrazabal et al. (91). They demonstrated in a large-scale analysis of chest X-ray images that gender imbalance in the training dataset led to incorrect classification of important conditions such as atelectasis, cardiomegaly or effusion. Puyol-Antón et al. performed the first analysis of DL fairness in cardiovascular segmentation using the UKB dataset (92). They found that the segmentation algorithm trained on a dataset balanced regarding participant sex but imbalanced concerning ethnicity resulted in less reliable outcomes for minority groups. It is easy to see how data biases might lead to a less inclusive distribution of resources. Lack of fairness might not only lead to loss of opportunities and worse health outcomes among minority groups but may also reduce public trust in AI applications.

Lekadir et al. identified the main guiding principles for fairness in medical imaging AI (12). Actions to promote AI fairness are not one step but should be implemented throughout the AI lifecycle. Here we summarize the main recommendations from a clinical perspective:

(1) Multi-disciplinarity, which stands for the inclusion of all important stakeholders (AI developers, imaging specialists, patients, and social scientists) in the AI design and implementation.
(2) Context-specific definition of fairness with regards to potential hidden biases in the dataset and data annotators.
(3) Standardization of key variables (e.g., sex, and ethnicity should be collected in a standardized way, because these descriptors of the groups can help test, and verify AI fairness).
(4) The data should be probed for (im)balances, particularly participant age, sex, ethnicity, and social background.

Once we know the potential biases, we have several options to deal with them. There are tools to promote AI fairness on a data collection and curation level, as well as in the model training and testing process.

(1) Data collection process in itself should be transparent and well-documented.
(2) Collecting multi-center data.
(3) Application of specialized statistical methods to evaluate fairness (e.g., true positive rate disparity, statistical parity group fairness, equalized odds, predictive/equality) (93, 94).
(4) Application of specialized statistical methods to mitigate bias (e.g., re-sampling, data augmentation, development of stratified models by sex, or ethnicity).

Exploratory data analysis is also a great tool to probe the dataset for hidden biases (8) and should not be a solitary task for the data scientist. Researchers with a medical background are more adept at picking up chance associations and odd correlations within the dataset. Among other things, data scientists can produce synthetic data to compensate for missing values to create a more balanced dataset. At the same time, we must always stay vigilant to the potential biological meaning of missing data before deciding to make up for it. Therein lies another strong argument for inclusive AI research.

Privacy and security

Any potential breach in healthcare AI systems can seriously undermine the trust of end-users. Therefore, developers should cooperate with cybersecurity experts to protect personal information against bad actors before clinical implementation. In some critical areas, such as data protection, there are firmly outlined rules in place: e.g., EU General Data Protection Regulation (GDPR) or the California Privacy Rights Act (CPRA). However, these regulations can never keep up with the speed of innovation.

Key issues surfacing with the use of clinical data for AI development (95):

(1) Sensitive data being shared without informed consent.
(2) Inappropriate informed consent forms (e.g., information within the consent form is detailed beyond the processing capability of the patient/user, no dedicated time allocated for consent review, and opaque use cases permitting patients from understanding how their data might be used).
(3) Data re-purposing without the patient's knowledge and consent.
(4) Personal data being exposed.
(5) Attacks on AI applications (e.g., data poisoning, adversarial attacks).

For example, the South Denver Cardiology Associates recently confirmed a data breach affecting 287,000 patients. The stolen dataset contained dates of birth, Social Security numbers, driver's license numbers, patient account numbers, health insurance information, and clinical information (96). This leakage might result in identity theft, insurance fraud or other inappropriate use of sensitive data. Moreover, in the field of medical imaging, particular attention is necessary toward dealing with potential adversarial attacks (97), including "one-pixel" attacks (98). These attacks involve slight changes to the input images intending to fool the AI and produce a false result. In other cases details of large scale data sharing agreements remain gray for the public (99), which might lead to data privacy controversies in the future.

Fortunately, several steps can be taken to mitigate these risks on an individual and institutional level:

(1) Increasing the awareness of privacy and security risks, informed consent and cybersecurity through (self)education.
(2) Transparent regulations of data privacy, data re-purposing.
(3) De-centralized, federated learning approaches such as federated learning. Federated learning is an ML setting where many de-centralized clients collaboratively train a model under the arrangement of a central server, keeping the data in several individual locations (100). Despite this, some researchers might be hesitant to use federated learning, because of the potential disclosure of the model. However, the data is never exposed to third parties, not even to the data scientist.
(4) Ongoing cybersecurity research into novel, more secure algorithms.

Transparency and traceability

Transparency in AI is a broad term; it refers to the information about the dataset, processes, uses, and outputs that is a prerequisite for accountability. AI transparency within medicine aims to provide all stakeholders with enough information to join in the discussion in a meaningful way. Two universal requirements guide and promote AI transparency:

(1) Data transparency includes transparent methods and guidelines for data collection, utilization, storage, sharing, and documentation.
(2) Model transparency means we have enough knowledge/information about a model's internal properties to apprehend its output meaningfully.

The goal of traceability is to document the entire development process and to monitor the behaviour and functioning of an AI model or system over time. This approach allows tracking any drift from the original training settings. As clinical practice constantly evolves, images provide greater granularity or novel guidelines emerge; keeping track of the model performance and adapting it to the new circumstances is critical (101). Two main concepts driving a decrease in model performance over time are "concept drift" and "data drift". Concept drift means that some underlying characteristics of variables change (for example, a novel type of cardiomyopathy is distinguished, creating a new class for a classification algorithm), which decreases the accuracy of the model. Data or dataset drift refers to the change in the data, meaning that a difference in the scanning device or image acquisition may directly affect the prediction model deployed (12).

Standardized dataset documentation methods can facilitate ML results' transparency, accountability and repeatability. Recently, Gebru et al. posed a list of questions on how and why data was collected, what is the composition of the data, and how it was curated and labeled in their document entitled "Datasheets for Datasets" (102). Sendak et al. proposed the use of "Model facts cards" for each ML model to ensure that clinicians have a thorough understanding of "how, when, how not, and when not to" incorporate the output into their decisions (103).

Explainability

In terms of an AI system, explainability means that it is possible to comprehend how the output was reached. The greater the explainability of a model, the better we can understand the internal mechanisms of a decision-making tool. However, as Arbelaez Ossa et al. (104) point out, the key issue in AI explainability is the lack of consensus among data scientists, regulators, and healthcare professionals regarding the definition and requirements.

Notably, explainability is not necessary for all ML models. Simple rule-based models applying linear regression or decision trees are inherently explainable. If we can calculate how a given parameter is weighted within the model, it is unnecessary to push the limits of explainability further.

From a strictly clinical end-user perspective, it is also not necessary to understand all steps involved in a complex DL network if the output is readily accessible and visually verifiable by a physician, such as segmentation. On the other hand, if the algorithm promises to deliver a clinical diagnosis or prognostic information based on imaging data or a combination of imaging features, the clinical application needs to reach high levels of intelligibility. A clinician who does not understand how the algorithm reached its conclusion will likely to rely on their own expertise rather than an opaque output.

To deal with the "black box" nature of particularly DL methods, several *post-hoc* explainability algorithms were defined to create more interpretable models. The so-called saliency maps or heat maps are the most widely adopted explainability tools in medical imaging. These color-coded maps show the contribution of each image region to a given model prediction (105, 106). Several distinct approaches can be utilized, such as Gradient-weighted Class Activation Mapping (Grad-CAM) (105) or Dense Captioning (DenseCap) (107), to capture the most crucial image areas. Saliency maps have long been applied in image analysis to understand better the key areas supporting the model's decision. As an example, Candemir et al. (108) trained a 3-dimensional convolutional neural network (CNN) to differentiate between coronary arteries with and without atherosclerosis and has shown the essential features learned by the system on color-coded maps. Saliency maps can also suggest if an algorithm picks up temporal data: Howard et al. (109) applied time distributed CNN model with saliency maps for disease classification based on echocardiography images. The author found that these new architectures more than halve the error rate of traditional CNNs, possibly because of the networks' ability to track the movement of specific structures such as heart valves throughout the cardiac cycle.

Local Interpretable Model-Agnostic Explanations (LIME) is used to explain the model locally for one single subject (110). LIME evaluates a given variable's contribution to the whole of the predictive model. SHapely Additive exPlanations (SHAP) is a model agnostic explainability model (can be used to interpret any model) (111). SHAP is based on game theory and can reveal each predictor's effect on the outcome. It calculates a score for each feature in the model, showing the feature's size and direction effects on the outcome. Al'Aref et al. (112) applied boosted ensemble algorithm (XGBoost) in the participants of the CONFIRM registry and showed that incorporating clinical features (e.g., age, sex, cardiovascular risk factors, laboratory values, and symptoms) in addition to coronary artery calcium score can accurately estimate the pretest likelihood of obstructive coronary artery disease on CCTA. They could supply the 20 most crucial features supporting the model's prediction using the SHAP method. Similarly, Fahmy et al. (113) applied the SHAP to support the interpretation of their model looking into the association between CMR metrics and adverse outcomes (cardiovascular hospitalization and all-cause death) in patients with dilated cardiomyopathy. Many other explainability techniques are also available, and new tools are likely to become more sophisticated and model-specific.

Although these models can improve model interpretation, their understanding requires additional efforts from the physicians. We are yet to see if their outputs can become as acceptable to the community and if they can overcome current limitations. Critiques of current explainability models warn that the performance of the explanations are not routinely quantified, and we can rarely elucidate if a given decision was sensible or

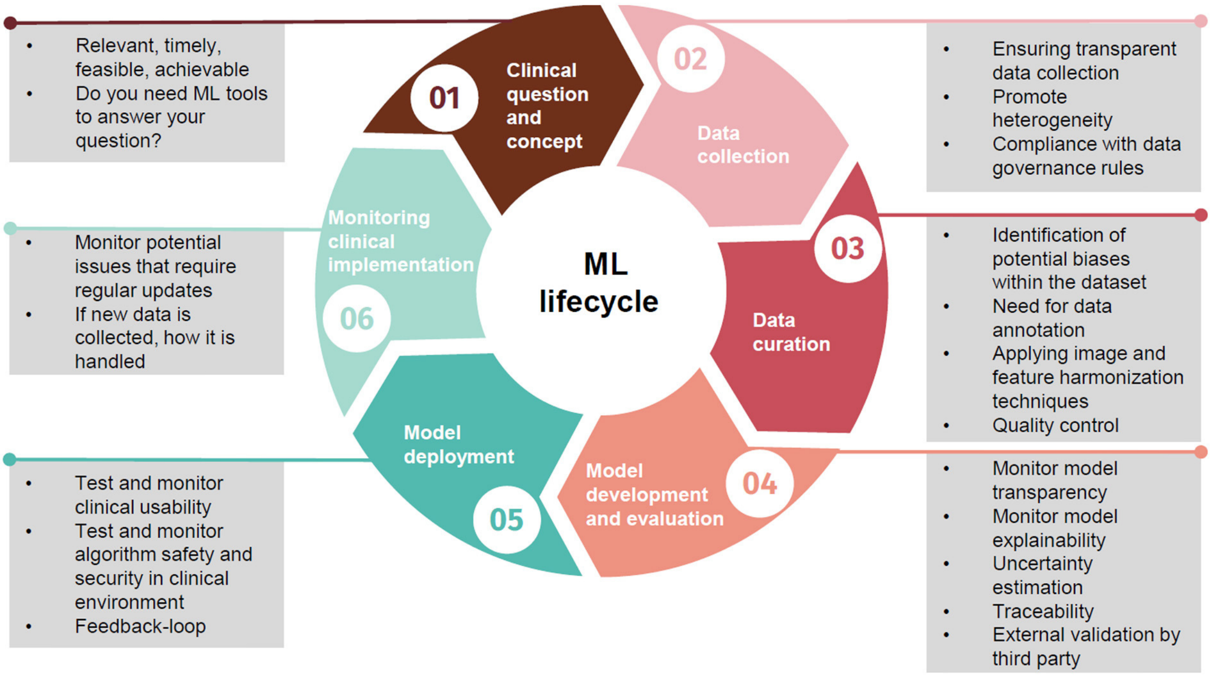

ML lifecycle
01
Clinical question and concept
02
Data collection
03
Data curation
04
Model development and evaluation
05
Model deployment
06
Monitoring clinical implementation
• Relevant, timely, feasible, achievable
• Do you need ML tools to answer your question?
• Ensuring transparent data collection
• Promote heterogeneity
• Compliance with data governance rules
• Identification of potential biases within the dataset
• Need for data annotation
• Applying image and feature harmonization techniques
• Quality control
• Monitor model transparency
• Monitor model explainability
• Uncertainty estimation
• Traceability
• External validation by third party
• Test and monitor clinical usability
• Test and monitor algorithm safety and security in clinical environment
• Feedback-loop
• Monitor potential issues that require regular updates
• If new data is collected, how it is handled

FIGURE 1
Principles of trustworthy AI within the machine learning lifecycle.

not (114). Moreover, they might reduce the complexity of a model to a level that is not representative and promote a false sense of security among users. Ghassemi et al. note that with the currently available methods, our best hope is to go through rigorous internal and external validation and use explainability models for troubleshooting and system audits (114).

Accountability and liability

Accountability refers to the state of being responsible. However, in the context of AI, where algorithms are based on both ML and human ingenuity, the mistakes or errors of the application come from humans developing or using machines (95, 115, 116). On the other hand, it is not clearly defined and regulated yet, with whom the responsibility of AI-powered medical tools lies. Does liability fall on developers, chief executive officers of the developing company, leaders of the healthcare institution buying and authorizing clinical utilization or the doctors using them? Notably, sometimes it is also hard to pinpoint why the AI-related medical error happened (95); therefore, responsibility issues can lead to daunting detective work, steering away the attention from the actual patient care. The main proposed tools to mitigate accountability issues within AI are: (1) the roles and responsibilities of developers and users should be defined, (2) a regulatory framework for accountability should be in place, and (3) dedicated regulatory agencies should be established and monitor AI use.

Clinical implementation

Even if an AI tool complies with all of the criteria mentioned earlier, integrating a new tool into clinical practice hides several expected and unexpected difficulties. The main obstacles to clinical implementation stem from three primary sources: (1) the differences among institutions regarding equipment, staffing, location, financial possibilities, and inner structures of each healthcare institution, (2) change in physician-patient relationship, and (3) difficulties of clinical and technical integration into existing workflows (95, 117).

Medical data, in general, is very noisy and requires human oversight before integration. Cardiovascular imaging data is slightly more structured than clinical records but still lacks interoperability to a great extent (118). Several initiatives already aim at increasing interoperability among healthcare providers [e.g., European Commission (119), Health Data Research UK (120)]. However, it seems fairly evident that medical AI tools will have to adapt to a certain level of data heterogeneity. The physician-patient relationship has been transformed by technical advances and the maturity of social sciences, but it is yet uncertain how AI tools will impact this relationship. Some argue that it will help by easing clinician workload and providing more personalized data for shared decision-making, while others question doctors' role once critical tasks are delegated to sophisticated algorithms (121). Clinical guidelines will need to be updated to consider the potential role of AI tools between healthcare workers and patients (95, 122). Moreover, these guidelines will also need to be updated to integrate novel tools into the clinical workflow without severe disruption of care (123).

Actionable steps

Trustworthy AI is not an obscure concept reserved for technical specialists and scholars of ethical reals (124), but rather a practical set of steps and questions, which, when implemented, can provide us with reliable tools for a new era in healthcare. In an effort to improve the overall quality of the AI prediction models, van Smeden et al. presented 12 critical questions for cardiovascular health professionals to ask (125). Moreover, the use of the Proposed Requirements for Cardiovascular Imaging-Related Machine Learning Evaluation (PRIME) checklist has been suggested by Sengupta et al. (126), a framework that contains a comprehensive list of crucial responsibilities that need to be completed when developing ML models. Here we report key questions to promote discussion of AI trustworthiness between clinicians and technical experts (Table 2), moreover we summarize how these principles fit into the ML lifecycle (Figure 1).

We have to acknowledge that in some instances medical research and consequently medical AI research is plainly inaccurate, but we can rectify these mistakes over time. AI competitions provide an excellent platform for robust validation or rebuttal of results. As an example, a recent competition to predict O(6)-Methylguanine-DNA-methyltransferase (MGMT) promoter methylation from brain magnetic resonance imaging (MRI) scans (127). Overall, 1,555 teams of many thousands of researchers took a large dataset of MRI scans and the results clearly demonstrate that this task is not possible with current approaches, even tough several group claimed to have achieved an ROC scores of up to 0.85 previously (128–130). This suggest that well designed competitions provide and excellent opportunity to improve the quality of AI research.

In order to promote the safe adoption of AI-powered tools in cardiovascular imaging, practicing doctors and future medical professionals need to be properly trained in the technical aspects, potential risks and limitations of the technology (131). McCoy et al. (132) and Grunhut et al. (133) proposed crucial points to improve AI literacy in medical education programs. Furthermore, the involvement and education of the general public are also essential for the broader adoption of these emerging tools.

Embracing the human-in-the-loop principle may offer further benefits where both imagers and ML algorithms fall short (134). It means that we can benefit from the advantages of AI models (i.e., automated segmentation or diagnosis) and having a human at various stages or checkpoints to correct potential errors or use critical thinking where algorithms are not confident in their results. The human can validate or correct the results where the algorithm delivers lower confidence outputs, creating a combined and better decision.

In essence, it does not matter if we call it trustworthy AI, reliable AI or responsible AI—the driving idea is to create an inclusive, collaborative effort in healthcare between all stakeholders. Our task is to consider the possible impact and test our AI tool and all elements of the AI development by posing the right questions relevant to our desired aims.

Author contributions

ZR-E, KL, SEP, and LS contributed to the conceptualization of this review. LS wrote the original draft and produced the table and figure. ZR-E, SEP, PM-H, HV, BM, and AML reviewed the manuscript from a clinical perspective. AS, CM, ER, PG, MK, AML, KL, and SEP reviewed the manuscript from a technical perspective. All authors contributed to the article and approved the submitted version.

Funding

LS received funding from the European Association of Cardiovascular Imaging (EACVI Research Grant App000076437). ZR-E recognizes the National Institute for Health Research (NIHR) Integrated Academic Training Programme which supports her Academic Clinical Lectureship post and was also supported by British Heart Foundation Clinical Research Training Fellowship Grant No. FS/17/81/33318. AS was supported by British Heart Foundation project grant (PG/21/10619). CM was supported by the Oxford NIHR Biomedical Research Center. SEP acknowledges support from the SmartHeart EPSRC programme grant (www.nihr.ac.uk; EP/P001009/1) and also from the CAP-AI programme. CAP-AI is led by Capital Enterprise in partnership with Barts Health NHS Trust and Digital Catapult and was funded by the European Regional Development Fund and Barts Charity. SEP, PG, and KL have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 825903 (euCanSHare project). HV, LS, and BM acknowledge funding from the Technology of Hungary from the National Research, Development, and Innovation Fund, financed under the TKP2021-NKTA funding scheme and Project No. RRF-2.3.1-21-2022-00004 (MILAB) which has been implemented with the support provided by the European Union. AML acknowledges support from the 'SmartHeart' EPSRC programme grant (www.nihr.ac.uk; EP/P001009/1) and from the UKRI London Medical Imaging and Artificial Intelligence Centre for Value Based Healthcare. This article is supported by the London Medical Imaging and Artificial Intelligence Centre for Value Based Healthcare (AI4VBH), which is funded from the Data to Early Diagnosis and Precision Medicine strand of the government's Industrial Strategy Challenge Fund, managed and delivered by Innovate UK on behalf of UK Research and Innovation (UKRI). ER was partly funded from the programme under grant agreement no. 825903 (euCanSHare project) and grant agreement no. 965345 (HealthyCloud project).The funders provided support in the form of salaries for authors as detailed above but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author disclaimer

Views expressed are those of the authors and not necessarily those of the AI4VBH Consortium members, the NHS, Innovate UK, or UKRI.

Conflict of interest

Author SEP provides consultancy to Cardiovascular Imaging Inc, Calgary, Alberta, Canada. Author MK was employed by Siemens Healthcare Hungary.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Quer G, Arnaout R, Henne M, Arnaout R. Machine learning and the future of cardiovascular care. *J Am Coll Cardiol.* (2021) 77:300–13. doi: 10.1016/j.jacc.2020.11.030

2. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J.* (2019) 6:94–8. doi: 10.7861/futurehosp.6-2-94

3. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach.* (2018) 28:689–707. doi: 10.1007/s11023-018-9482-5

4. Rajakariar K, Koshy AN, Sajeev JK, Nair S, Roberts L, Teh AW. Accuracy of a smartwatch based single-lead electrocardiogram device in detection of atrial fibrillation. *Heart.* (2020) 106:665–70. doi: 10.1136/heartjnl-2019-316004

5. Bai W, Sinclair M, Tarroni G, Oktay O, Rajchl M, Vaillant G, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *J Cardiovasc Magn Reson.* (2018) 20:65. doi: 10.1186/s12968-018-0471-x

6. Davies RH, Augusto JB, Bhuva A, Xue H, Treibel TA, Ye Y, et al. Precision measurement of cardiac structure and function in cardiovascular magnetic resonance using machine learning. *J Cardiovasc Magn Reson.* (2022) 24:16. doi: 10.1186/s12968-022-00846-4

7. AI Central. *ACR Data Science Institution AI Central.* AI Central (2022). Available online at: https://aicentral.acrdsi.org/ (accessed July 3, 2022).

8. Ammanath B. *Trustworthy AI: A Business Guide for Navigating Trust and Ethics in AI.* Hoboken, NJ: John Wiley and Sons, Incorporated (2022).

9. High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trusthworthy AI.* High-Level Expert Group on Artificial Intelligence (2019). Available online at: https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html (accessed July 4, 2022).

10. Buruk B, Ekmekci PE, Arda B. A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Med Health Care Philos.* (2020) 23:387–99. doi: 10.1007/s11019-020-09948-1

11. Geis JR, Brady AP, Wu CC, Spencer J, Ranschaert E, Jaremko JL, et al. Ethics of artificial intelligence in radiology: summary of the joint European and north American multisociety statement. *Radiology.* (2019) 293:436–40. doi: 10.1148/radiol.2019191586

12. Lekadir K, Osuala R, Gallin C, Lazrak N, Kushibar K, Tsakou G, et al. FUTURE-AI: guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. *arXiv.* (2021). Available online at: http://arxiv.org/abs/2109.09658 (accessed June 21, 2022).

13. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial intelligence in precision cardiovascular medicine. *J Am Coll Cardiol.* (2017) 69:2657–64. doi: 10.1016/j.jacc.2017.03.571

14. Dey D, Slomka PJ, Leeson P, Comaniciu D, Shrestha S, Sengupta PP, et al. Artificial intelligence in cardiovascular imaging. *J Am Coll Cardiol.* (2019) 73:1317–35. doi: 10.1016/j.jacc.2018.12.054

15. Leiner T, Rueckert D, Suinesiaputra A, Baeßler B, Nezafat R, Išgum I, et al. Machine learning in cardiovascular magnetic resonance: basic concepts and applications. *J Cardiovasc Magn Reson.* (2019) 21:61. doi: 10.1186/s12968-019-0575-y

16. Martin-Isla C, Campello VM, Izquierdo C, Raisi-Estabragh Z, Baeßler B, Petersen SE, et al. Image-Based cardiac diagnosis with machine learning: a review. *Front Cardiovasc Med.* (2020) 7:1. doi: 10.3389/fcvm.2020.00001

17. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* (2015) 521:436–44. doi: 10.1038/nature14539

18. Al'Aref SJ, Singh G, Baskaran L, Metaxas DN, editors. *Machine Learning in Cardiovascular Medicine.* London: Academic Press (2021). 424 p.

19. Alice CY, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. *Radiol Artif Intell.* (2022) 4:e210064. doi: 10.1148/ryai.210064

20. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology.* (2018) 286:800–9. doi: 10.1148/radiol.2017171920

21. Ryan M. In AI we trust: ethics, artificial intelligence, and reliability. *Sci Eng Ethics.* (2020) 26:2749–67. doi: 10.1007/s11948-020-00228-y

22. Lewis PR, Marsh S. What is it like to trust a rock? A functionalist perspective on trust and trustworthiness in artificial intelligence. *Cogn Syst Res.* (2022) 72:33–49. doi: 10.1016/j.cogsys.2021.11.001

23. Feigenbaum H. Evolution of echocardiography. *Circulation.* (1996) 93:1321–7. doi: 10.1161/01.CIR.93.7.1321

24. Lang RM, Badano LP, Mor-Avi V, Afilalo J, Armstrong A, Ernande L, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American society of echocardiography and the european association of cardiovascular imaging. *J Am Soc Echocardiogr.* (2015) 28:1–39.e14. doi: 10.1016/j.echo.2014.10.003

25. Gudmundsson P, Rydberg E, Winter R, Willenheimer R. Visually estimated left ventricular ejection fraction by echocardiography is closely correlated with formal quantitative methods. *Int J Cardiol.* (2005) 101:209–12. doi: 10.1016/j.ijcard.2004.03.027

26. Asch FM, Poilvert N, Abraham T, Jankowski M, Cleve J, Adams M, et al. Automated echocardiographic quantification of left ventricular ejection fraction without volume measurements using a machine learning algorithm mimicking a human expert. *Circ Cardiovasc Imaging.* (2019) 12:e009303. doi: 10.1161/CIRCIMAGING.119.009303

27. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ.* (2020) 369:m1328. doi: 10.1136/bmj.m1328

28. von Borzyskowski I, Mazumder A, Mateen B, Wooldridge M, editors. *Data Science and AI in the Age of COVID-19.* (2021). Available online at: https://www.turing.ac.uk/sites/default/files/2021-06/data-science-and-ai-in-the-age-of-covid_full-report_2.pdf (accessed July 4, 2022).

29. Leslie D, Mazumder A, Peppin A, Wolters MK, Hagerty A. Does "AI" stand for augmenting inequality in the era of covid-19 healthcare? *BMJ.* (2021) 372:n304. doi: 10.1136/bmj.n304

30. Wolterink JM, Leiner T, Viergever MA, Išgum I. Generative adversarial networks for noise reduction in low-dose CT. *IEEE Trans Med Imaging.* (2017) 36:2536–45. doi: 10.1109/TMI.2017.2708987

31. Oksuz I, Ruijsink B, Puyol-Antón E, Clough JR, Cruz G, Bustin A, et al. Automatic CNN-based detection of cardiac MR motion artefacts using k-space data augmentation and curriculum learning. *Med Image Anal.* (2019) 55:136–47. doi: 10.1016/j.media.2019.04.009

32. Benz DC, Ersözlü S, Mojon FLA, Messerli M, Mitulla AK, Ciancone D, et al. Radiation dose reduction with deep-learning image reconstruction for coronary computed tomography angiography. *Eur Radiol.* (2022) 32:2620–8. doi: 10.1007/s00330-021-08367-x

33. Caballero J, Price AN, Rueckert D, Hajnal JV. Dictionary learning and time sparsity for dynamic MR data reconstruction. *IEEE Trans Med Imaging.* (2014) 33:979–94. doi: 10.1109/TMI.2014.2301271

34. Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, et al. Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. *Circulation.* (2018) 138:1623–35. doi: 10.1161/CIRCULATIONAHA.118.034338

35. Edalati M, Zheng Y, Watkins MP, Chen J, Liu L, Zhang S, et al. Implementation and prospective clinical validation of AI-based planning and shimming techniques in cardiac MRI. *Med Phys.* (2022) 49:129–43. doi: 10.1002/mp.15327

36. Leclerc S, Smistad E, Pedrosa J, Østvik A, Cervenansky F, Espinosa F Espeland T, et al. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. *IEEE Trans Med Imaging.* (2019) 38:2198–210. doi: 10.1109/TMI.2019.2900516

37. Huang W, Huang L, Lin Z, Huang S, Chi Y, Zhou J, et al. Coronary artery segmentation by deep learning neural networks on computed tomographic coronary angiographic images. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).* Honolulu, HI: IEEE (2018). p. 608–11. doi: 10.1109/EMBC.2018.8512328

38. Sengupta PP, Huang Y-M, Bansal M, Ashrafi A, Fisher M, Shameer K, et al. Cognitive machine-learning algorithm for cardiac imaging: a pilot study for differentiating constrictive pericarditis from restrictive cardiomyopathy. *Circ Cardiovasc Imaging.* (2016) 9:e004330. doi: 10.1161/CIRCIMAGING.115.004330

39. de Vos BD, Wolterink JM, Leiner T, de Jong PA, Lessmann N, Išgum I. Direct automatic coronary calcium scoring in cardiac and chest CT. *IEEE Trans Med Imaging.* (2019) 38:2127–38. doi: 10.1109/TMI.2019.2899534

40. Zhang N, Yang G, Gao Z, Xu C, Zhang Y, Shi R, et al. Deep learning for diagnosis of chronic myocardial infarction on nonenhanced cardiac cine MRI. *Radiology.* (2019) 291:606–17. doi: 10.1148/radiol.2019182304

41. Samad MD, Ulloa A, Wehner GJ, Jing L, Hartzel D, Good CW, et al. Predicting survival from large echocardiography and electronic health record datasets: optimization with machine learning. *JACC Cardiovasc Imaging.* (2019) 12:681–9. doi: 10.1016/j.jcmg.2018.04.026

42. Patel MR, Nørgaard BL, Fairbairn TA, Nieman K, Akasaka T, Berman DS, et al. 1-Year impact on medical practice clinical outcomes of FFRCT. *JACC Cardiovasc Imaging.* (2020) 13:97–105. doi: 10.1016/j.jcmg.2019.03.003

43. Cheng S, Fang M, Cui C, Chen X, Yin G, Prasad SK, et al. LGE-CMR-derived texture features reflect poor prognosis in hypertrophic cardiomyopathy patients with systolic dysfunction: preliminary results. *Eur Radiol.* (2018) 28:4615–24. doi: 10.1007/s00330-018-5391-5

44. Tokodi M, Schwertner WR, Kovács A, Tosér Z, Staub L, Sárkány A, et al. Machine learning-based mortality prediction of patients undergoing cardiac resynchronization therapy: the SEMMELWEIS-CRT score. *Eur Heart J.* (2020) 41:1747–56. doi: 10.1093/eurheartj/ehz902

45. Queirós S, Dubois C, Morais P, Adriaenssens T, Fonseca JC, Vilaça JL, et al. Automatic 3D aortic annulus sizing by computed tomography in the planning of transcatheter aortic valve implantation. *J Cardiovasc Comput Tomogr.* (2017) 11:25–32. doi: 10.1016/j.jcct.2016.12.004

46. Casaclang-Verzosa G, Shrestha S, Khalil MJ, Cho JS, Tokodi M, Balla S, et al. Network tomography for understanding phenotypic presentations in aortic stenosis. *JACC Cardiovasc Imaging.* (2019) 12:236–48. doi: 10.1016/j.jcmg.2018.11.025

47. Hoshino M, Zhang J, Sugiyama T, Yang S, Kanaji Y, Hamaya R, et al. Prognostic value of pericoronary inflammation and unsupervised machine-learning-defined phenotypic clustering of CT angiographic findings. *Int J Cardiol.* (2021) 333:226–32. doi: 10.1016/j.ijcard.2021.03.019

48. Zheng Q, Delingette H, Fung K, Petersen SE, Ayache N. Pathological cluster identification by unsupervised analysis in 3,822 UK biobank cardiac MRIs. *Front Cardiovasc Med.* (2020) 7:539788. doi: 10.3389/fcvm.2020.539788

49. Liao J, Huang L, Qu M, Chen B, Wang G. Artificial intelligence in coronary CT angiography: current status and future prospects. *Front Cardiovasc Med.* (2022) 9:896366. doi: 10.3389/fcvm.2022.896366

50. Achenbach S, Fuchs F, Goncalves A, Kaiser-Albers C, Ali ZA, Bengel FM, et al. Non-invasive imaging as the cornerstone of cardiovascular precision medicine. *Eur Heart J Cardiovasc Imaging.* (2022) 23:465–75. doi: 10.1093/ehjci/jeab287

51. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM.* (2017) 60:84–90. doi: 10.1145/3065386

52. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med.* (2016) 375:1216. doi: 10.1056/NEJMp1606181

53. Langlotz CP. Will artificial intelligence replace radiologists? *Radiol Artif Intell.* (2019) 1:e190058. doi: 10.1148/ryai.2019190058

54. The Royal College of Radiologists. *Clinical Radiology Census Report 2021.* The Royal College of Radiologists (2021). Available online at: https://www.rcr.ac.uk/clinical-radiology/rcr-clinical-radiology-census-report-2021 (accessed July 31, 2022).

55. Lekadir K, Osuala R, Gallin C, Lazrak N, Kushibar K, Tsakou G, et al. FUTURE-AI: guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. *arXiv [Preprint].* (2021). arXiv: 2109.09658. Available online at: http://arxiv.org/abs/2109.09658 (accessed June 21, 2022).

56. ISO/IEC TR 24029-1:2021. *Artificial Intelligence (AI)—Assessment of the Robustness of Neural Networks.* ISO/IEC TR 24029-1:2021 (2021). Available online at: https://www.iso.org/standard/77609.html (accessed June 12, 2022).

57. Antun V, Renna F, Poon C, Adcock B, Hansen AC. On instabilities of deep learning in image reconstruction - does AI come at a cost? *Proc Natl Acad Sci USA.* (2020) 117:30088–95. doi: 10.1073/pnas.1907377117

58. Galati F, Ourselin S, Zuluaga MA. From accuracy to reliability and robustness in cardiac magnetic resonance image segmentation: a review. *Appl Sci.* (2022) 12:3936. doi: 10.3390/app12083936

59. Petersen SE, Matthews PM, Francis JM, Robson MD, Zemrak F, Boubertakh R, et al. UK biobank's cardiovascular magnetic resonance protocol. *J Cardiovasc Magn Reson.* (2015) 18:8. doi: 10.1186/s12968-016-0227-4

60. Bamberg F, Kauczor H-U, Weckbach S, Schlett CL, Forsting M, Ladd SC, et al. Whole-Body MR imaging in the German national cohort: rationale, design, and technical background. *Radiology.* (2015) 277:206–20. doi: 10.1148/radiol.2015142272

61. Isensee F, Jaeger P, Full PM,Wolf I, Engelhardt S,Maier-Hein KH. *Automatic Cardiac Disease Assessment on Cine-MRI via Time-Series Segmentation and Domain Specific Features.* Cham: Springer (2018). (2018). doi: 10.1007/978-3-319-75541-0_13

62. Budai A, Suhai FI, Csorba K, Toth A, Szabo L, Vago H, et al. Fully automatic segmentation of right and left ventricle on short-axis cardiac MRI images. *Comput Med Imaging Graph.* (2020) 85:101786. doi: 10.1016/j.compmedimag.2020.101786

63. Campello VM, Gkontra P, Izquierdo C, Martin-Isla C, Sojoudi A, Full PM, et al. Multi-Centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge. *IEEE Trans Med Imaging.* (2021) 40:3543–54. doi: 10.1109/TMI.2021.3090082

64. Society for Cardiovascular Magnetic Resonance. *About the SCMR Registry.* Society for Cardiovascular Magnetic Resonance. Available online at: https://scmr.org/page/Registry (accessed July 29, 2022).

65. *Cardiac Atlas Project.* Available online at: http://www.cardiacatlas.org/ (accessed July 29, 2022).

66. Maurovich-Horvat P, Ferencik M, Voros S, Merkely B, Hoffmann U. Comprehensive plaque assessment by coronary CT angiography. *Nat Rev Cardiol.* (2014) 11:390–402. doi: 10.1038/nrcardio.2014.60

67. Lin A, Kolossváry M, Motwani M, Išgum I, Maurovich-Horvat P, Slomka PJ, et al. Artificial intelligence in cardiovascular imaging for risk stratification in coronary artery disease. *Radiol Cardiothorac Imaging.* (2021) 3:e200512. doi: 10.1148/ryct.2021200512

68. Lin A, Manral N, McElhinney P, Killekar A, Matsumoto H, Kwiecinski J, et al. Deep learning-enabled coronary CT angiography for plaque and stenosis quantification and cardiac risk prediction: an international multicentre study. *Lancet Digit Health.* (2022) 4:e256–65. doi: 10.1016/S2589-7500(22)00022-X

69. Robinson R, Valindria VV, Bai W, Oktay O, Kainz B, Suzuki H, et al. Automated quality control in image segmentation: application to the UK biobank cardiovascular magnetic resonance imaging study. *J Cardiovasc Magn Reson.* (2019) 21:18. doi: 10.1186/s12968-019-0523-x

70. Hann E, Popescu IA, Zhang Q, Gonzales RA, Barutçu A, Neubauer S, et al. Deep neural network ensemble for on-the-fly quality control-driven segmentation of cardiac MRI T1 mapping. *Med Image Anal.* (2021) 71:102029. doi: 10.1016/j.media.2021.102029

71. Biasiolli L, Hann E, Lukaschuk E, Carapella V, Paiva JM, Aung N, et al. Automated localization and quality control of the aorta in cine CMR can significantly accelerate processing of the UK Biobank population data. *PLoS ONE.* (2019) 14:e0212272. doi: 10.1371/journal.pone.0212272

72. Bard A, Raisi-Estabragh Z, Ardissino M, Lee AM, Pugliese F, Dey D, et al. Automated quality-controlled cardiovascular magnetic resonance pericardial fat quantification using a convolutional neural network in the UK biobank. *Front Cardiovasc Med.* (2021) 8:677574. doi: 10.3389/fcvm.2021.677574

73. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. *Npj Digit Med.* (2021) 4:4. doi: 10.1038/s41746-020-00367-3

74. Chang S, Han K, Suh YJ, Choi BW. Quality of science and reporting for radiomics in cardiac magnetic resonance imaging studies: a systematic review. *Eur Radiol.* (2022) 32:4361–73. doi: 10.1007/s00330-022-08587-9

75. Kolossváry M, Karády J, Szilveszter B, Kitslaar P, Hoffmann U, Merkely B, et al. Radiomic features are superior to conventional quantitative computed tomographic metrics to identify coronary plaques with napkin-ring sign. *Circ Cardiovasc Imaging.* (2017) 10:e006843. doi: 10.1161/CIRCIMAGING.117.006843

76. Zeleznik R, Foldyna B, Eslami P, Weiss J, Alexander I, Taron J, et al. Deep convolutional neural networks to predict cardiovascular risk from computed tomography. *Nat Commun.* (2021) 12:715. doi: 10.1038/s41467-021-20966-2

77. Lin A, Kolossváry M, Cadet S, McElhinney P, Goeller M, Han D, et al. Radiomics-Based precision phenotyping identifies unstable coronary plaques from computed tomography angiography. *Cardiovasc Imaging.* (2022) 15:859–71. doi: 10.1016/j.jcmg.2021.11.016

78. Oikonomou EK, Williams MC, Kotanidis CP, Desai MY, Marwan M, Antonopoulos AS, et al. A novel machine learning-derived radiotranscriptomic signature of perivascular fat improves cardiac risk prediction using coronary CT angiography. *Eur Heart J.* (2019) 40:3529–43. doi: 10.1093/eurheartj/ehz592

79. Lin A, Kolossváry M, Yuvaraj J, Cadet S, McElhinney PA, Jiang C, et al. Myocardial infarction associates with a distinct pericoronary adipose tissue radiomic phenotype. *JACC Cardiovasc Imaging.* (2020) 13:2371–83. doi: 10.1016/j.jcmg.2020.06.033

80. Izquierdo C, Casas G, Martin-Isla C, Campello VM, Guala A, Gkontra P, et al. Radiomics-based classification of left ventricular non-compaction, hypertrophic cardiomyopathy, and dilated cardiomyopathy in cardiovascular magnetic resonance. *Front Cardiovasc Med.* (2021) 8:764312. doi: 10.3389/fcvm.2021.764312

81. Antonopoulos AS, Boutsikou M, Simantiris S, Angelopoulos A, Lazaros G, Panagiotopoulos I, et al. Machine learning of native T1 mapping radiomics for classification of hypertrophic cardiomyopathy phenotypes. *Sci Rep.* (2021) 11:23596. doi: 10.1038/s41598-021-02971-z

82. Baeßler B, Mannil M, Maintz D, Alkadhi H, Manka R. Texture analysis and machine learning of non-contrast T1-weighted MR images in patients with hypertrophic cardiomyopathy—preliminary results. *Eur J Radiol.* (2018) 102:61–7. doi: 10.1016/j.ejrad.2018.03.013

83. Baessler B, Mannil M, Oebel S, Maintz D, Alkadhi H, Manka R. Subacute and chronic left ventricular myocardial scar: accuracy of texture analysis on nonenhanced cine MR images. *Radiology.* (2018) 286:103–12. doi: 10.1148/radiol.2017170213

84. Rauseo E, Izquierdo Morcillo C, Raisi-Estabragh Z, Gkontra P, Aung N, Lekadir K, et al. New imaging signatures of cardiac alterations in ischaemic heart disease and cerebrovascular disease using CMR radiomics. *Front Cardiovasc Med.* (2021) 8:716577. doi: 10.3389/fcvm.2021.716577

85. Raisi-Estabragh Z, Gkontra P, Jaggi A, Cooper J, Augusto J, Bhuva AN, et al. Repeatability of cardiac magnetic resonance radiomics: a multi-centre multi-vendor test-retest study. *Front Cardiovasc Med.* (2020) 7:586236. doi: 10.3389/fcvm.2020.586236

86. Campello VM, Martín-Isla C, Izquierdo C, Guala A, Palomares JFR, Viladés D, et al. Minimising multi-centre radiomics variability through image normalisation: a pilot study. *Sci Rep.* (2022) 12:12532. doi: 10.1038/s41598-022-16375-0

87. ISO 9241-11:2018. *Ergonomics of Human-System Interaction—Part 11: Usability: Definitions and Concept.* Geneva: ISO (2018)

88. Lara Hernandez KA, Rienmüller T, Baumgartner D, Baumgartner C. Deep learning in spatiotemporal cardiac imaging: a review of methodologies and clinical usability. *Comput Biol Med.* (2021) 130:104200. doi: 10.1016/j.compbiomed.2020.104200

89. ISO/IEC TR 24027:2021. *Artificial Intelligence (AI) — Bias in AI Systems and AI Aided Decision Making.* ISO/IEC TR 24027:2021 (2021). Available online at: https://www.iso.org/standard/77607.html (accessed June 16, 2022).

90. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med.* (2018) 1:18. doi: 10.1038/s41746-018-0029-1

91. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc Natl Acad Sci USA.* (2020) 117:12592–4. doi: 10.1073/pnas.1919012117

92. Puyol-Antón E, Ruijsink B, Mariscal Harana J, Piechnik SK, Neubauer S, Petersen SE, et al. Fairness in cardiac magnetic resonance imaging: assessing sex and racial bias in deep learning-based segmentation. *Front Cardiovasc Med.* (2022) 9:859310. doi: 10.3389/fcvm.2022.859310

93. Barocas S, Hardt M, Narayanan A. Fairness in Machine Learning. *Nips Tutor.* Fairmlbook.Org (2017). Available online at: http://www.fairmlbook.org

94. Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. CheXclusion: Fairness gaps in deep chest X-ray classifiers. In: *Pacific Symposium on Biocomputing, Vol. 26.* (2020). p. 232–43. doi: 10.1142/9789811232701_0022

95. European Parliament. *Directorate General for Parliamentary Research Services. Artificial Intelligence in Healthcare: Applications, Risks, and Ethical and Societal Impacts.* LU: Publications Office (2022). Available online at: https://data.europa.eu/doi/10.2861/568473 (accessed July 29, 2022).

96. HIPAA Journal. *South Denver Cardiology Associates Confirms Data Breach Affecting 287,000 Patients.* HIPAA Journal. Available online at: https://www.hipaajournal.com/south-denver-cardiology-associates-confirms-data-breach-affecting-287000-patients/ (accessed August 1, 2022).

97. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. *Science.* (2019) 363:1287–9. doi: 10.1126/science.aaw4399

98. Sipola T, Kokkonen T. One-pixel attacks against medical imaging: A conceptual framework. In: *World Conference on Information Systems and Technologies.* Terceira Island: Springer (2021) p. 197–203. doi: 10.1007/978-3-030-72657-7_19

99. Heart Flow. *NHS England and NHS Improvement Mandate Adoption of AI-Powered HeartFlow Analysis to Fight Coronary Heart Disease.* Heart Flow. Available online at: https://www.heartflow.com/newsroom/nhs-england-and-nhs-improvement-mandate-adoption-of-ai-powered-heartflow-analysis-to-fight-coronary-heart-disease/ (accessed September 22, 2022).

100. Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Bhagoji AN, et al. Advances and open problems in federated learning. *Found Trends® Mach Learn.* (2021) 14:1–210. doi: 10.1561/2200000083

101. Mora-Cantallops M, Sanchez-Alonso S, Garc?a-Barriocanal E, Sicilia M-A. Traceability for trustworthy ai: A review of models and tools. *Big Data Cogn Comput.* (2021) 5:20. doi: 10.3390/bdcc5020020

102. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Iii HD, et al. Datasheets for datasets. *Commun ACM.* (2021) 64:86–92. doi: 10.1145/3458723

103. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. *NPJ Digit Med.* (2020) 3:41. doi: 10.1038/s41746-020-0253-3

104. Arbelaez Ossa L, Starke G, Lorenzini G, Vogt JE, Shaw DM, Elger BS. Re-focusing explainability in medicine. *Digit Health.* (2022) 8:205520762210744. doi: 10.1177/20552076221074488

105. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis.* (2020) 128:336–59. doi: 10.1007/s11263-019-01228-7

106. Tjoa E, Khok HJ, Chouhan T, Cuntai G. Improving deep neural network classification confidence using heatmap-based eXplainable AI. *arXiv.* (2022). Available online at: http://arxiv.org/abs/2201.00009 (accessed July 3, 2022).

107. Johnson J, Karpathy A, Fei-Fei L. DenseCap: fully convolutional localization networks for dense captioning. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Las Vegas, NV: IEEE (2016). p. 4565–74. doi: 10.1109/CVPR.2016.494

108. Candemir S, White RD, Demirer M, Gupta V, Bigelow MT, Prevedello LM, et al. Automated coronary artery atherosclerosis detection and weakly supervised localization on coronary CT angiography with a deep 3-dimensional convolutional neural network. *Comput Med Imaging Graph.* (2020) 83:101721. doi: 10.1016/j.compmedimag.2020.101721

109. Howard JP, Tan J, Shun-Shin MJ, Mahdi D, Nowbar AN, Arnold AD, et al. Improving ultrasound video classification: an evaluation of novel deep learning methods in echocardiography. *J Med Artif Intell.* (2020) 3:4. doi: 10.21037/jmai.2019.10.03

110. Ribeiro MT, Singh S, Guestrin C. "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.* San Francisco, CA: ACM (2016). p. 1135–44. doi: 10.1145/2939672.2939778

111. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *arXiv [Preprint]. arXiv: 1705.07874.* Available online at: https://arxiv.org/pdf/1705.07874.pdf

112. Al'Aref SJ, Maliakal G, Singh G, van Rosendael AR, Ma X, Xu Z, et al. Machine learning of clinical variables and coronary artery calcium scoring for the prediction of obstructive coronary artery disease on coronary computed tomography angiography: analysis from the CONFIRM registry. *Eur Heart J.* (2020) 41:359–67. doi: 10.1093/eurheartj/ehz565

113. Fahmy AS, Csecs I, Arafati A, Assana S, Yankama TT, Al-Otaibi T, et al. An explainable machine learning approach reveals prognostic significance of right ventricular dysfunction in nonischemic cardiomyopathy. *JACC Cardiovasc Imaging.* (2022) 15:766–79. doi: 10.1016/j.jcmg.2021.11.029

114. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health.* (2021) 3:e745–50. doi: 10.1016/S2589-7500(21)00208-9

115. Kaplan B. How should health data be used?: privacy, secondary use, and big data sales. *Camb Q Healthc Ethics.* (2016) 25:312–29. doi: 10.1017/S0963180115000614

116. Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. *arXiv.* (2020). Available online at: http://arxiv.org/abs/2001.00973 (accessed August 1, 2022).

117. Arora A. Conceptualising artificial intelligence as a digital healthcare innovation: an introductory review. *Med Devices Evid Res.* (2020) 13:223–30. doi: 10.2147/MDER.S262590

118. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. *NPJ Digit Med.* (2019) 2:79. doi: 10.1038/s41746-019-0158-1

119. European Commission. *European Health Data Space.* European Commission. Available online at: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en. (accessed August 1, 2022).

120. Health Data Research UK. *BHF Data Science Centre.* Health Data Research UK. Available online at: https://www.hdruk.org/helping-with-health-data/bhf-data-science-centre/ (accessed August 1, 2022).

121. Matthew Nagy, MPH Bryan Sisk, MD. How will artificial intelligence affect patient-clinician relationships? *AMA J Ethics.* (2020) 22:E395–400. doi: 10.1001/amajethics.2020.395

122. Cohen IG. Informed consent and medical artificial intelligence: what to tell the patient? *George Law J.* (2020) 108:1425–69. doi: 10.2139/ssrn.3529576

123. Meskó B, Görög M. A short guide for medical professionals in the era of artificial intelligence. *NPJ Digit Med.* (2020) 3:126. doi: 10.1038/s41746-020-00333-z

124. Hatherley JJ. Limits of trust in medical AI. *J Med Ethics.* (2020) 46:478–81. doi: 10.1136/medethics-2019-105935

125. van Smeden M, Heinze G, Van Calster B, Asselbergs FW, Vardas PE, Bruining N, et al. Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease. *Eur Heart J.* (2022) 43:2921–30. doi: 10.1093/eurheartj/ehac238

126. Sengupta PP, Shrestha S, Berthon B, Messas E, Donal E, Tison GH, et al. Proposed requirements for cardiovascular imaging-related machine learning evaluation (PRIME): a checklist. *JACC Cardiovasc Imaging.* (2020) 13:2017–35. doi: 10.1016/j.jcmg.2020.07.015

127. Kaggle. *RSNA-MICCAI Brain Tumor Radiogenomic Classification.* Kaggle. Available online at: https://www.kaggle.com/competitions/rsna-miccai-brain-tumor-radiogenomic-classification (accessed Septembet 26, 2022).

128. Sasaki T, Kinoshita M, Fujita K, Fukai J, Hayashi N, Uematsu Y, et al. Radiomics and MGMT promoter methylation for prognostication of newly diagnosed glioblastoma. *Sci Rep.* (2019) 9:14435. doi: 10.1038/s41598-019-50849-y

129. Korfiatis P, Kline TL, Coufalova L, Lachance DH, Parney IF, Carter RE, et al. MRI texture features as biomarkers to predict MGMT methylation status in glioblastomas: MRI texture features to predict MGMT methylation status. *Med Phys.* (2016) 43:2835–44. doi: 10.1118/1.4948668

130. Han L, Kamdar MR. MRI to MGMT: predicting methylation status in glioblastoma patients using convolutional recurrent neural networks. In: *Biocomputing 2018.* World Scientific (2018). p. 331–42. doi: 10.1142/9789813235533_0031

131. Keane PA, Topol EJ. AI-facilitated health care requires education of clinicians. *Lancet.* (2021) 397:1254. doi: 10.1016/S0140-6736(21)00722-4

132. McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? *NPJ Digit Med.* (2020) 3:86. doi: 10.1038/s41746-020-0294-7

133. Grunhut J, Wyatt AT, Marques O. Educating future physicians in artificial intelligence (AI): an integrative review and proposed changes. *J Med Educ Curric Dev.* (2021) 8:238212052110368. doi: 10.1177/23821205211036836

134. Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: humanism and artificial intelligence. *JAMA.* (2018) 319:19–20. doi: 10.1001/jama.2017.19198

frontiers | Frontiers in Cardiovascular Medicine

TYPE Original Research
PUBLISHED 15 November 2022
DOI 10.3389/fcvm.2022.964355

OPEN ACCESS

EDITED BY
Christian Tesche,
Augustinum Klinik München, Germany

REVIEWED BY
Sebastian Haberkorn,
Heinrich Heine University of Düsseldorf, Germany
Liang Zhong,
National Heart Centre Singapore, Singapore
Verena Brandt,
Technical University Munich, Germany

*CORRESPONDENCE
Nils Hampe
n.hampe@amsterdamumc.nl

SPECIALTY SECTION
This article was submitted to Cardiovascular Imaging, a section of the journal Frontiers in Cardiovascular Medicine

RECEIVED 08 June 2022
ACCEPTED 17 October 2022
PUBLISHED 15 November 2022

CITATION
Hampe N, van Velzen SGM, Planken RN, Henriques JPS, Collet C, Aben J-P, Voskuil M, Leiner T and Išgum I (2022) Deep learning-based detection of functionally significant stenosis in coronary CT angiography. *Front. Cardiovasc. Med.* 9:964355. doi: 10.3389/fcvm.2022.964355

Deep learning-based detection of functionally significant stenosis in coronary CT angiography

Nils Hampe[1,2,3]*, Sanne G. M. van Velzen[1,2,3], R. Nils Planken[4], José P. S. Henriques[5], Carlos Collet[6], Jean-Paul Aben[7], Michiel Voskuil[8], Tim Leiner[9,10] and Ivana Išgum[1,2,3,4]

[1]Department of Biomedical Engineering and Physics, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, Netherlands, [2]Amsterdam Cardiovascular Sciences, Heart Failure and Arrhythmias, Amsterdam, Netherlands, [3]Informatics Institute, University of Amsterdam, Amsterdam, Netherlands, [4]Department of Radiology and Nuclear Medicine, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, Netherlands, [5]AMC Heart Center, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, Netherlands, [6]Onze Lieve Vrouwziekenhuis, Cardiovascular Center Aalst, Aalst, Belgium, [7]Pie Medical Imaging BV, Maastricht, Netherlands, [8]Department of Cardiology, University Medical Centre Utrecht, Utrecht, Netherlands, [9]Department of Radiology, University Medical Center Utrecht, Utrecht, Netherlands, [10]Department of Radiology, Mayo Clinic, Rochester, MN, United States

Patients with intermediate anatomical degree of coronary artery stenosis require determination of its functional significance. Currently, the reference standard for determining the functional significance of a stenosis is invasive measurement of the fractional flow reserve (FFR), which is associated with high cost and patient burden. To address these drawbacks, FFR can be predicted non-invasively from a coronary CT angiography (CCTA) scan. Hence, we propose a deep learning method for predicting the invasively measured FFR of an artery using a CCTA scan. The study includes CCTA scans of 569 patients from three hospitals. As reference for the functional significance of stenosis, FFR was measured in 514 arteries in 369 patients, and in the remaining 200 patients, obstructive coronary artery disease was ruled out by Coronary Artery Disease-Reporting and Data System (CAD-RADS) category 0 or 1. For prediction, the coronary tree is first extracted and used to reconstruct an MPR for the artery at hand. Thereafter, the coronary artery is characterized by its lumen, its attenuation and the area of the coronary artery calcium in each artery cross-section extracted from the MPR using a CNN. Additionally, characteristics indicating the presence of bifurcations and information indicating whether the artery is a main branch or a side-branch of a main artery are derived from the coronary artery tree. All characteristics are fed to a second network that predicts the FFR value and classifies the presence of functionally significant stenosis. The final result is obtained by merging the two predictions. Performance of our method is evaluated on held out test sets from multiple centers and vendors. The method achieves an area under the receiver operating characteristics curve (AUC) of 0.78, outperforming other works that do not require manual correction of the segmentation of the artery. This demonstrates that our method may reduce the number of patients that unnecessarily undergo invasive measurements.

KEYWORDS

convolutional neural networks, coronary computed tomography angiography, fractional flow reserve, transformer, coronary artery tree

1. Introduction

Coronary artery disease (CAD) is the leading cause of death worldwide (1, 2). CAD is characterized by a buildup of atherosclerotic plaque in the coronary arteries, potentially leading to a functionally significant stenosis, i.e., stenosis that causes myocardial ischaemia. Currently, invasive fractional flow reserve (FFR) measurements are considered the clinical reference for determining the functional significance of a stenosis. However, invasive FFR is associated with high costs and it constitutes a burden for the patient (3, 4). Hence, identifying patients with functionally significant stenosis prior to the invasive measurements and treatment would be of high value. While visual interpretation of coronary CT angiography (CCTA) by clinical experts enables identification of the vast majority of functionally significant stenoses (high sensitivity), it suffers from a high number of false positives (low specificity) (5, 6). As a consequence, 20–50% of invasive FFR measurements are performed unnecessarily (6). Therefore, predicting FFR non-invasively from CT angiography is a subject of intensive investigations.

For non-invasive FFR prediction from CCTA, several algorithms have been proposed. Currently, most accurate methods are based on computational fluid dynamics (CFD) (7–12). However, CFD-methods are computationally expensive, hampering (real-time) implementation on clinical workstations. Moreover, CFD-based methods rely on the accuracy of the anatomical artery tree model, i.e., artery lumen segmentation and boundary conditions describing aortic pressure and peripheral resistances, which are challenging to obtain.

In addition to development of CFD-based FFR prediction methods, approaches emerged that correlate quantitative indices derived from CCTA with measured FFR value. These clinical indices characterize a coronary artery through e.g., transluminal attenuation gradient (TAG) (13, 14) or plaque volume (15, 16), or describe specific lesions by quantifying degree of stenosis (16, 17) or contrast density difference (CDD) (18, 19). While the mathematical simplicity and intuitive design of the calculated indices enables their interpretation, it limits their capability to model the complex relationship between FFR and the coronary artery characteristics on CCTA. Hence, to improve FFR prediction with clinical indices, machine learning classifiers were employed that combined multiple indices (11, 16, 20–24). This led to a substantial performance increase compared to the performance of a single index. Similarly, using clinical indices describing the local geometry and plaque composition, as well as global features describing the entire artery tree, Itu et al. (21) trained a deep learning classifier for prediction of the pressure gradient caused by each lesion. For training, the authors leveraged hemodynamic simulations in 12,000 artificial coronary anatomies. To enable learning of relationships between lesions, Wang et al. (25) and Gao et al. (26) employed the same features as input to a recursive neural network (RNN). However, these index-based works share a drawback with CFD-based methods: calculating the indices requires accurate segmentation of the coronary artery lumen, which can be highly challenging, especially in the presence of pathology (27). While these methods typically use an automatic segmentation method as a starting point, errors in the automatic segmentation regularly necessitate substantial manual interaction.

To avoid lengthy assessment times, algorithms that apply deep learning technology directly to the CCTA scan have been investigated. Deep learning algorithms have shown the ability to model complex relations of image characteristics in a large number of medical task (28). However, these methods often require a large amount of diverse training data, which may be challenging to obtain in the medical domain. Hence, previous deep learning-based works reduced the complexity of the task by focusing analysis to a relevant region of interest (29–32) or by training separate networks to extract image characteristics (29, 31–33). Given that obstruction in the coronary arteries is expected to lead to underperfusion of the left ventricle (LV), Zreik et al. (29) focused analysis on the LV myocardium by characterizing it using a convolutional autoencoder (CAE). Subsequently, the authors predicted the presence of a functionally significant stenosis using a support vector machine (SVM), which can be strained with limited data due to its small number of parameters. In a subsequent study, Zreik et al. (31) characterized the coronary arteries by training a CAE on multi planar reconstructions (MPRs) of the coronary arteries. Related to this, Denzinger et al. (30) used a CNN in combination with an RNN to classify MPRs. The authors used the clinical revascularization decision as reference label, obtained using functional tests including cardiac stress MRI or MIBI SPECT. To further improve performance, Zreik et al. (32) combined the characterizations of the myocardium and the coronary arteries using a deep learning-based multi instance learning framework. As an alternative to focusing analysis to a region of interest, Kumamaru et al. (33) enhanced lumen-related image features using a difference image between the CCTA scan and a non-contrast cardiac CT, synthesized from CCTA using deep learning. Thereafter, authors trained a 3D ladder network to extract relevant image characteristics. These deep learning-based works were limited by their moderate performance. Unlike other deep learning-based works that applied CNNs to the CCTA scan, Li et al. (34) first used the artery segmentation to extract a point cloud representing the coronary artery geometry. The authors used this point cloud as input to a modified version of the point-net (35), to predict the pressure in the coronary artery tree. However, the authors used hemodynamic simulations as reference labels in training and testing and hence, the performance compared to invasive FFR measurements is unknown.

In this work, we propose a method to non-invasively predict the presence of a functionally significant stenosis in an artery through deep learning-based analysis of CCTA scans. As in previous deep learning works, we focus on a region of interest by first extracting an MPR for the artery of interest. Given that previous research demonstrated the importance of lumen area, its attenuation and plaque volume for predicting FFR, we exploit these characteristics. To circumvent the need for challenging lumen segmentation, during testing, we use a convolutional neural network (CNN) to directly extract these characteristics from the MPR along the artery centerline. Additionally, we extract characteristics directly from the coronary artery tree that indicate per coronary artery centerline point whether it is located in a main artery or side-branch and whether a bifurcation is present at that location. Thereafter, using the extracted characteristics we assess the functional significance of FFR.

For this purpose, we train a second network to perform both regression of the FFR value and classification of the functional significance of an artery. In contrast to previous works that use abstract, high dimensional features, extraction of our specific characteristics is supervised, resulting in targeted information distillation and lower dimensional features. While training of previous deep learning-based works on the limited training data requires compressing the high dimensional features along the artery prior to training the stenosis classification (31, 32), our targeted extraction of artery characteristics enables us to directly use these characteristics along the artery as input to our second network. This second network is designed to exploit the spatial structure encoded in the extracted characteristics through the use of convolutions and self-attention. The so-learned representations are likely more descriptive than unsupervised features characterizing the entire artery. Additionally, using tangible characteristics, instead of the abstract features employed in previous deep learning works (29, 31–33), enables interpretability of our method. We performed experiments on a diverse data set from multiple centers and vendors.

This paper is organized as follows. The data is described in Section 2. Section 3 provides a description of the method, which is followed by a description of our evaluation in Section 4 and by experiments and results in Section 5. We discuss our findings in Section 6 and describe our conclusions in Section 7.

2. Data

2.1. Patients and imaging data

This study retrospectively included 657 patients who underwent CCTA for suspected obstructive CAD. Scans were acquired in three different hospitals: Scans of 263 patients (age 47–79 years) were acquired in the Onze Lieve Vrouwe Ziekenhuis, Aalst, Belgium (Site 1) with a Siemens Somatom Definition Flash CT scanner; Scans of 152 patients (age 34–84 years) were acquired in the University Medical Center Utrecht, the Netherlands (Site 2) with a Philips iCT 256 CT scanner; Scans of 243 patients (age 48–85 years) were acquired in the Amsterdam University Medical Centers—location University of Amsterdam, the Netherlands (Site 3) with a Siemens Somatom Force CT scanner. Patients were only included if all arteries were in the field of view of the CCTA scan. This study was approved (Site 1) or the need for informed consent was waived by the respective institutional review boards (Site 2, Site 3).

During acquisition, contrast medium was injected with a flow rate of 4 to 6 mL/s for a total of 30 to 92 mL iopromide (Ultravist 300 mg I/mL, Bayer Healthcare, Berlin, Germany), depending on the patient weight and test bolus images (29, 36). The tube voltages ranged between 70 and 140 kVp and tube currents between 71 and 901 mAs. All scans were reconstructed to an in-plane resolution ranging from 0.22 to 0.83 mm^2 with 0.3 to 0.5 mm slice increment and 0.5 to 1.0 mm slice thickness.

In total 85 out of 658 patients were excluded because the quality of the CCTA scan was not sufficient due to e.g., severe step-and-shoot artifacts ($n = 22$), severe cardiac motion artifacts ($n = 47$) or artifacts caused by metal implants ($n = 16$; Table 1). Furthermore, patients who underwent stenting or coronary artery bypass grafting (CABG) prior to CCTA acquisition were excluded ($n = 4$). After exclusions, 569 patients remained for further analysis.

For development and validation of the method, 438 arteries with FFR measurements from 302 patients were used. Additionally, for independent evaluation of the method, the performance was evaluated with two held-out test sets. The first set consisted of 76 arteries with FFR measurements in 67 patients randomly sampled from all three sites. It is referred to as $\text{Test}_{\text{Cath}}$. The sets used for development and validation, as

TABLE 1 Patients were excluded due to artifacts, i.e., severe step-and-shoot artifacts, severe cardiac motion artifacts or artifacts caused by metal implants.

	Artifacts			Stenting/	Total
	Step and shoot	Motion	Metal	CABG	excluded
Site 1	13	21	2	1	37
Site 2	8	15	11	3	37
Site 3	1	11	3	0	15
Total	22	47	16	4	89

well as $\text{Test}_{\text{Cath}}$, consist of patients with intermediate degree of anatomical stenosis for which the cardiologist recommended invasive FFR measurement to assess hemodynamic significance of the stenosis. Therefore, these sets are representative of the clinical population that undergoes FFR measurement for suspicion of obstructive CAD in the catheterization laboratory, which represents our primary target population. The second test set consisted of 600 arteries of 200 patients, in which instead of invasive FFR measurement obstructive CAD was ruled out as they were assigned to category zero (absence of stenosis or plaque in all coronary arteries) or one (low degree of anatomical stenosis or plaque in all coronary arteries) according to the Coronary Artery Disease-Reporting and Data System (CAD-RADS) (37). Hence, arteries in this population have a degree of stenosis $< 25\%$. The chances of finding functionally significant stenosis in these patients would be marginal (38). To warrant that our algorithm classifies these arteries correctly, they were used for testing by assuming FFR > 0.8, indicating the absence of functionally significant stenosis. Thus, this second test set is referred to as $\text{Test}_{\text{NoCath}}$ and it is used for evaluation purposes only as no patient with little or no stenosis was sent for invasive FFR measurement.

Analysis in this set was performed for the main arteries, i.e., left anterior descending artery (LAD), left circumflex artery (LCX) and right coronary artery (RCA). CAD-RADS scoring was performed within 3 days of the acquisition of the CCTA scan. Figure 1 and Table 2 show details regarding the data selection.

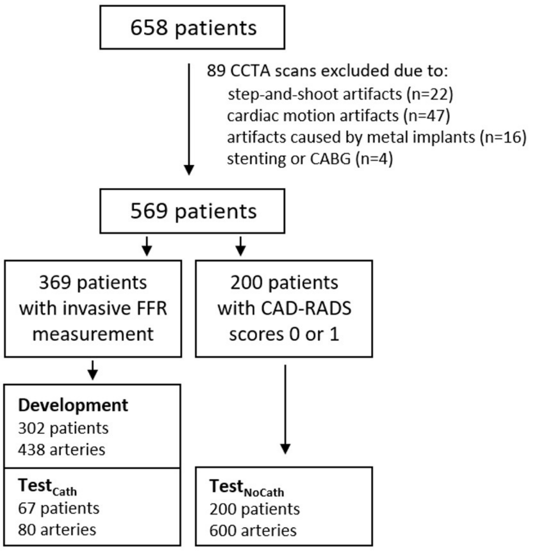

658 patients
89 CCTA scans excluded due to:
step-and-shoot artifacts (n=22)
cardiac motion artifacts (n=47)
artifacts caused by metal implants (n=16)
stenting or CABG (n=4)
569 patients
369 patients with invasive FFR measurement
200 patients with CAD-RADS scores 0 or 1
Development
302 patients
438 arteries
Test_{Cath}
67 patients
80 arteries
Test_{NoCath}
200 patients
600 arteries

FIGURE 1
Data included in the study.

2.2. FFR measurements

Among the 569 patients, 369 underwent invasive FFR measurement in 514 arteries. To measure FFR, a coronary pressure guidewire (Certus Pressure Wire, St. Jude Medical, St. Paul, Minnesota or Pressure wire X, Abbott Vascular, California) was inserted into the distal segment of the coronary vessel, and maximal hyperemia was induced by administration of intravenous adenosine through a central vein. The lowest FFR value measured at the most distal location was chosen for analysis. An FFR pullback was performed to assess the presence of drift. If multiple FFR measurements were available in one artery, the value measured at the most distal location was chosen. The maximum time interval between the acquisition of the CCTA scan and the FFR measurement was 90 days for Site 1 and 1 year for Site 2 and Site 3.

2.3. Reference artery characteristics

To train the network for extraction of artery characteristics, reference annotations of the coronary artery lumen and coronary calcium were required. Given the extensive manual workload of the tasks, these were performed semi-automatically in a subset of 56 arteries, randomly selected from the development data set. First, automatic segmentations of the lumen and calcium were generated in the original CT image volumes using methods previously developed in our group (39, 40). Thereafter, automatic segmentations were

TABLE 2 Data set for training and validation (Development), and two separate test sets, one with FFR measurements ($\text{Test}_{\text{Cath}}$) and one without ($\text{Test}_{\text{NoCath}}$).

		Development	$\text{Test}_{\text{Cath}}$	$\text{Test}_{\text{NoCath}}$
Patients		302	67	200
Arteries		438	76	600
Hospital				
	Site 1	249	58	0
	Site 2	159	14	0
	Site 3	30	4	600
Anatomical segment				
	LAD	221	50	200
	LCX	81	9	200
	RCA	72	10	200
	LM	14	1	0
	SB	50	6	0
FFR statistics				
	% positive	0.42	0.78	0.00
	Mean FFR	0.82	0.70	-
	Std FFR	0.11	0.16	-

In addition to the number of arteries and patients in each set, the table lists the contribution of each site to each set, the anatomical segments in which the invasive measurements were performed and statistics describing the distribution of FFR values. Anatomical segments were categorized into the main branches, i.e., LAD, left anterior descending artery; LCX, left circumflex artery; RCA, right coronary artery; LM, left main; SB, side-branch of the main arteries.

transferred to the MPR of the artery, visually inspected and corrected when needed. Using the segmentations, the reference lumen area and calcium area were generated by summing up the pixels of the respective segmentation in each cross-sectional slice of the MPR perpendicular to the artery centerline. Note that MPRs for all arteries share the same spacings and in-plane resolution. For the average lumen attenuation, the average of the image pixels within the lumen segmentation mask was calculated in each cross-sectional slice of the MPR.

3. Methods

Our method assesses the functional significance of stenosis in an artery from CCTA. First, we extract the coronary artery centerline tree. To analyze the artery of interest, we then reconstruct an MPR. Subsequently, we extract relevant characteristics of the artery along its centerline using a 2D CNN and the characteristics of the artery within the coronary artery tree. Using these characteristics, we assess the presence of a functionally significant stenosis with a dedicated CNN (Figure 2).

3.1. Artery extraction

To localize the coronary arteries in the CCTA image, the coronary artery centerline tree is extracted and anatomical labels are assigned to the tree's segments using our previously developed method (41). Thereafter, the labeled centerline tree is inspected and manually corrected if needed. This is the only manual interaction that might be required for our method at test time. Figure 3 illustrates the pre-processing steps. In most cases this took 1 min, but could take up to at most 5 min when challenged by pathology. For each selected artery centerline, an MPR with 0.1 mm in-plane voxel size and 0.5 mm distance between MPR slices is reconstructed using trilinear interpolation. The in-plane shape of the MPR is 127 x 127 and the number of slices is dependent on the artery length. Finally, image intensities in the MPRs are normalized to zero mean and unit variance across the data set, to ensure training stability of the neural networks.

3.2. Artery characterization

3.2.1. Extraction of coronary artery characteristics

To automatically characterize a coronary artery, we extract the lumen area, its attenuation and the amount of coronary artery calcium from the artery's MPR. Specifically, for each point of the coronary artery centerline, we predict the lumen area, the average lumen attenuation and the calcium area in its cross-section with a 2D CNN (Figure 4). The network analyzes stacks of three cross-sectional slices and consists of four alternating convolutional blocks and pooling operations. The convolutional blocks are comprised of two convolutional layers (kernel size 3, 16 filters), each followed by batch normalization and the ReLU activation function. Finally, three separate output heads regress values for the lumen area, average attenuation in the lumen and calcium area for the central slice of the input stack.

3.2.2. Extraction of coronary tree characteristics

The coronary artery geometry has impact on the characteristics of the blood flow and local appearance of the artery. Therefore, for each point along the coronary artery centerline, we extract two additional characteristics. The first one indicates the presence of bifurcations at the artery centerline point. The second one indicates whether a centerline point belongs to a main branch (i.e., left main (LM), LAD, LCX, RCA) or a side-branch. The locations of bifurcations and side-branches follow from the tree topology and labels. Specifically, for each MPR slice, information about bifurcations and side branches was extracted from the coronary artery centerline point of the tree at that location, i.e., by considering the amount

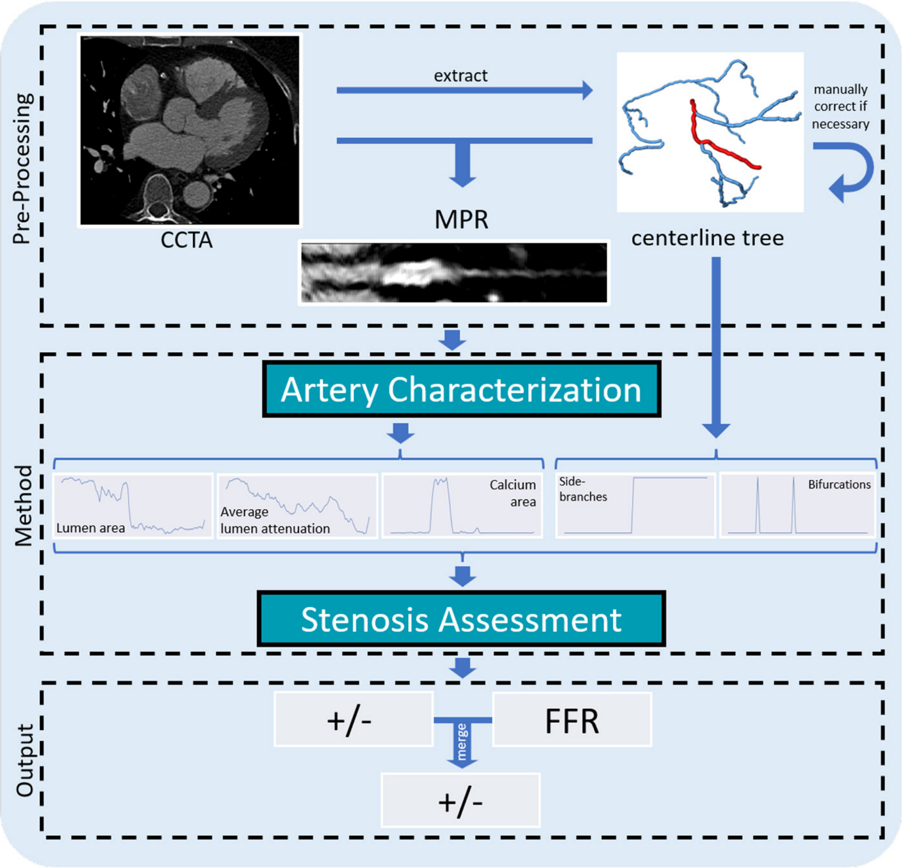


FIGURE 2
Overview of our method for assessing the presence of a functionally significant stenosis in a coronary artery. From the CCTA scan, we extract a coronary artery centerline tree. For each artery we generate an MPR that is further analyzed to predict the lumen area, its average attenuation and the calcium area per centerline point. These characteristics, as well as characteristics indicating the presence of bifurcations and whether the artery is a main branch or to a side branch of a main artery, are fed to the classification network to determine the presence of a functionally significant stenosis in the artery.

of successive centerline points and the label, respectively. We normalize all characteristics to zero mean and unit variance across the training data set.

3.3. Stenosis assessment

To assess the presence of a functionally significant stenosis, we analyze the extracted artery characteristics with a 1D convolutional neural network (Figure 5). The network performs both regression of the FFR value and classification of functionally significant stenosis. To obtain a robust final decision, we merge the predictions.

The network receives the 5 artery characteristics (lumen area, average lumen attenuation, calcium area, bifurcations and side-branches) as input. To focus on changes in lumen area and its attenuation rather than their absolute values, we calculate their percentage difference at each location in the artery with respect to the previous location. Because the relevant features in the lumen area and its attenuation may be subtle and may appear in different locations along the artery (i.e., a stenosis is expected to cause changes in the attenuation distal to the appearance in the lumen area), these two characteristics are first separately encoded. This is done using two non-shared convolutional layers with the LeakyReLU activation function applied in between the layers. Thereafter, the remaining characteristics are concatenated with the encoded features from the lumen area and its attenuation.

The information of all five extracted artery characteristics is merged by a common encoder, consisting of convolutional layers and a transformer layer, as follows: To increase the receptive field and reduce the dimensionality, average pooling with kernel size 4 is applied, followed by two convolutional layers with dilation 1 and 2, respectively. Each convolutional

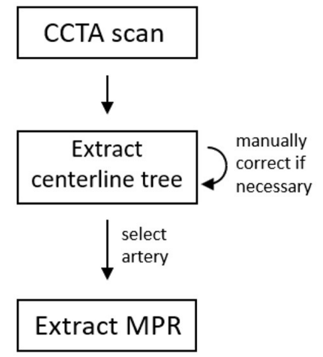


FIGURE 3
Pre-processing steps.

layer is followed by the LeakyReLU activation function, instance normalization and dropout. Subsequently, artery encodings are concatenated with the original lumen area and its attenuation, and fed to a transformer layer (42). Due to the global receptive field, the transformer layer connects all artery points with one another. This potentially enables modeling interaction between multiple lesions, and proximal and distal section of the artery. The network has two output heads that are each designed to perform a separate task: one performs regression of the FFR value and the other performs classification of the presence of a functionally significant stenosis in the artery. Inspired by the additive nature of sequential flow resistances, the regression head is designed to predict pressure drops along the artery. First, two layers of convolutions are applied, each followed by the LeakyReLU activation function, instance normalization and dropout. Thereafter, a third convolutional layer with a single output filter map is followed by a ReLU activation function to enforce positivity of the pressure drops. Finally, the predicted pressure drops are summed up along the artery using a sum pooling layer and the resulting overall FFR drop is transformed into the final FFR value by subtracting it from 1. The classification head predicts the presence of functionally significant stenosis (FFR$\leq$0.8). To explicitly relate proximal and distal sections, first, adaptive sum pooling with 5 output features is applied followed by 2 dense layers, each with LeakyReLU activation and dropout. At last, a dense layer with a single output filter map and sigmoid activation yields output probabilities for functionally significant stenosis.

For all convolutions throughout the network for stenosis assessment, a kernel size of 3 is employed in combination with zero-padding to prevent shrinkage of the features. Furthermore, for all convolutions as well as for the transformer, a relatively small number of 16 filter maps is utilized, to balance the required expressiveness and to prevent overfitting. For the same purpose, all dropout probabilities are set to 0.5.

During training, the regression head is supervised using the mean squared error with the reference FFR value. Since the invasive reference FFR is often not measured at the most distal location, predicted pressure drop contributions from anatomical locations distal to the measurement location are masked during training and testing. The measurement location is assumed to be 10 mm distal to the annotated lesion location, in line with measurement protocols from clinical practice. For the $\text{Test}_{\text{NoCath}}$ data set, as no measurement was taken, the most distal clinically relevant location (lumen area > 2 mm^2) was chosen as the measurement location. The classification task is supervised using the binary cross entropy loss function.

To combine strengths of the classification and the regression head, their outputs are merged into a single probability for the presence of a functionally significant stenosis in the artery. While the classification head directly predicts probabilities for the positive and negative class, the regressed FFR values are distributed around the threshold of 0.8. To allow their merging, the predicted FFR values are first transformed into pseudo-probabilities by linearly scaling a symmetric window around 0.8, using the following formula:

$$p_{pseudo} = \begin{cases} 0.5 - \frac{(FFR_{regress} - 0.8)}{0.4} & \text{for } FFR_{regress} \in [0.6, 1.0] \\ 1.0 & \text{for } FFR_{regress} \in [0.0, 0.6) \end{cases} \tag{1}$$

Figure 6 shows the transformation function, with x values corresponding to the predicted FFR (input) and y values to the pseudo probabilities (transformations).

To obtain the final prediction result, the pseudo-probabilities are averaged with the probabilities from the classification head.

We developed our method performing randomized 10 fold cross-validation (i.e., training of 10 networks on random 90% subsets of the development selection and testing on the remaining 10%). To increase robustness of the method and determine uncertainty of the prediction, during testing we ensemble the 10 networks by averaging the predicted probabilities and FFR values (43). For the prediction of the uncertainty, we calculate the standard deviation over the probabilities and the FFR values (44).

4. Evaluation

We evaluate the performance of our method by computing AUC, accuracy, sensitivity, and specificity using the invasively measured FFR as reference. This is done for the final prediction, obtained by merging the classification and regression results, and for the regressed FFR values and classification probabilities separately. For evaluation, the regressed FFR values and reference FFR values were dichotomized using the threshold of 0.8 for significant stenosis. To test for

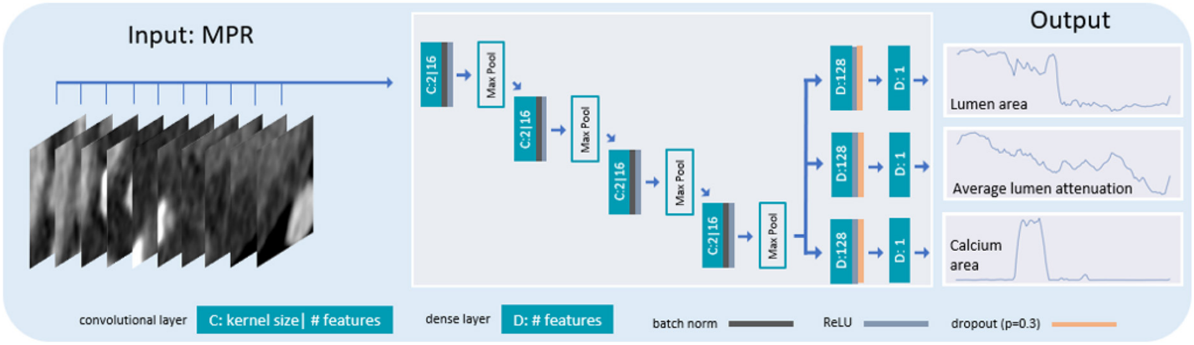


FIGURE 4
Architecture of the network for extracting artery characteristics. Stacks of 3 cross-sectional artery slices are fed to a 2D CNN with 4 pooling layers interleaved with convolutions. The network is trained to predict the lumen area, its average attenuation and the calcium area for the central slice of the 3 input slices.

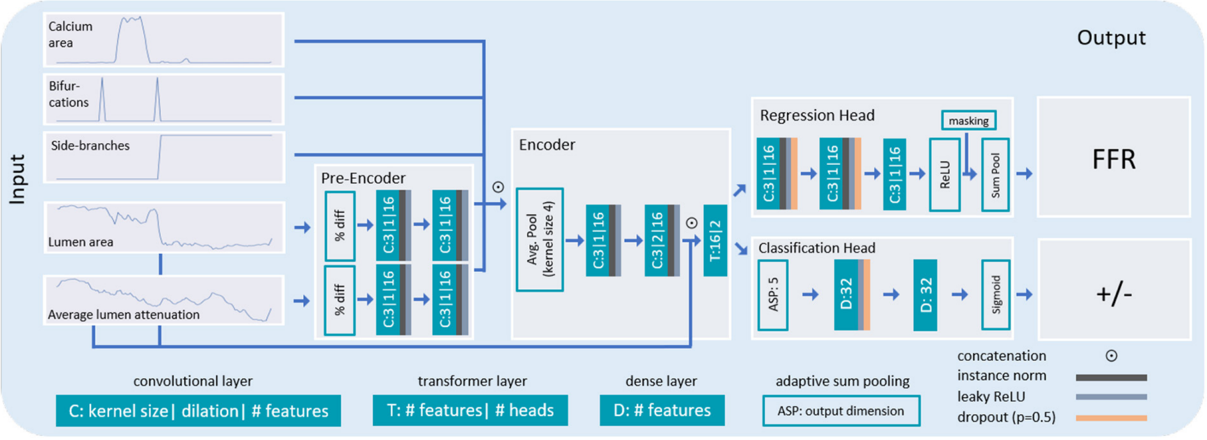


FIGURE 5
Architecture of the network used for stenosis assessment. The lumen area and its attenuation predicted by the characterization network are first pre-encoded and subsequently concatenated with the calcium area, and with additional characteristics indicating bifurcations and whether the analysis is performed in the main- or side-branch of the artery. The combined encodings are thereafter fed to the encoder. In the encoder, the features are first pooled and thereafter, convolutions and a transformer layer are applied. For final classification, two separate output heads are applied. In the regression head, two convolutional layers and the ReLU activation function are used. The resulting sequence is pooled along the artery dimension and subtracted from 1 to yield a single FFR value. In the classification head, the features are pooled to a fixed length of 5 (2.5 mm). Thereafter, two dense layers are used in combination with the sigmoid activation function to yield output probabilities for the presence of a functionally significant stenosis in the artery.

the statistical significance of the AUC differences between models, we performed permutation testing (45) with 1,000 iterations and report *p*-values. To obtain a patient-level prediction, the highest output value of all classified arteries in a patient is used to assign the predicted class to the patient. In the reference, patients were considered negative if none of the measured arteries had an FFR $\leq$ 0.8, and otherwise positive.

5. Experiments and results

5.1. Experimental settings

To account for possible overfitting during training of the network for artery characterization, the 56 annotated arteries were split into 42 arteries for training, 4 arteries for validation and 10 arteries for quantitative testing. The network was trained

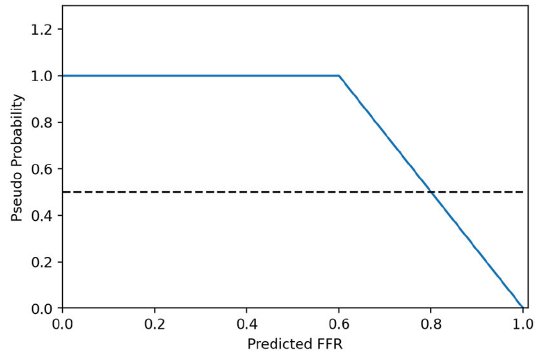


FIGURE 6
Transformation from predicted FFR values to pseudo probabilities.

TABLE 3 Performance of our method.

Algorithm	Selection	AUC	Accuracy	Sensitivity	Specificity
Merged	$\text{Test}_{\text{Cath}}$	0.78	0.79	0.84	0.61
	$\text{Test}_{\text{NoCath}}$	-	0.86	-	-
Classification	$\text{Test}_{\text{Cath}}$	0.68	0.55	0.53	0.61
	$\text{Test}_{\text{NoCath}}$	-	0.90	-	-
Regression	$\text{Test}_{\text{Cath}}$	0.83	0.89	0.95	0.67
	$\text{Test}_{\text{NoCath}}$	-	0.59	-	-

The table lists the obtained AUC, accuracy, sensitivity and specificity. Rows correspond to the classification, regression and merged outputs. To obtain binary predictions, for probabilities a threshold of 0.5 was used and for regressed FFR values a threshold of 0.8.

TABLE 4 Performance of our method on $\text{Test}_{\text{Cath}}$ per site.

Data set	Arteries	AUC	Accuracy	Sensitivity	Specificity
All	76	0.78	0.79	0.84	0.61
Site 1	58	0.84	0.84	0.85	0.78
Site 2	14	0.73	0.71	0.88	0.50
Site 3	4	(0.00)	(0.25)	(0.00)	(0.33)

Results demonstrate best performance on the data from Site 1. Only 4 arteries in $\text{Test}_{\text{Cath}}$ were acquired at Site 3 (Table 2). As this may not be sufficient to obtain a representative per-site performance, the respective numbers are presented in brackets.

utilizing the mean absolute error as loss function and the ADAMW (46) optimizer with a learning rate of 10^{-5} and a batch size of 512. Training was terminated after 800 epochs as convergence was reached. Based on preliminary experiments, the loss term of the lumen attenuation was scaled with a factor 0.1, such that all loss terms are within the same order of magnitude. After training, we applied the network to each cross-section of the MPR to obtain the lumen area, its average attenuation and the area of calcium along the length of the artery.

The network for stenosis assessment was trained for 150 epochs using the ADAMW (46) optimizer with a linearly scheduled cyclic learning rate. The cyclic learning rate varied between 5e-4 and 1e-5 over a period of 40 epochs. Because different artery lengths limit the network for stenosis assessment to process only a single artery at a time, the loss was accumulated over 8 training iterations before backpropagating, corresponding to an effective batch size of 8. The loss terms of the regression head and the classification head were weighted equally, as both terms are of similar magnitude and as both tasks are equally important.

5.2. Stenosis assessment

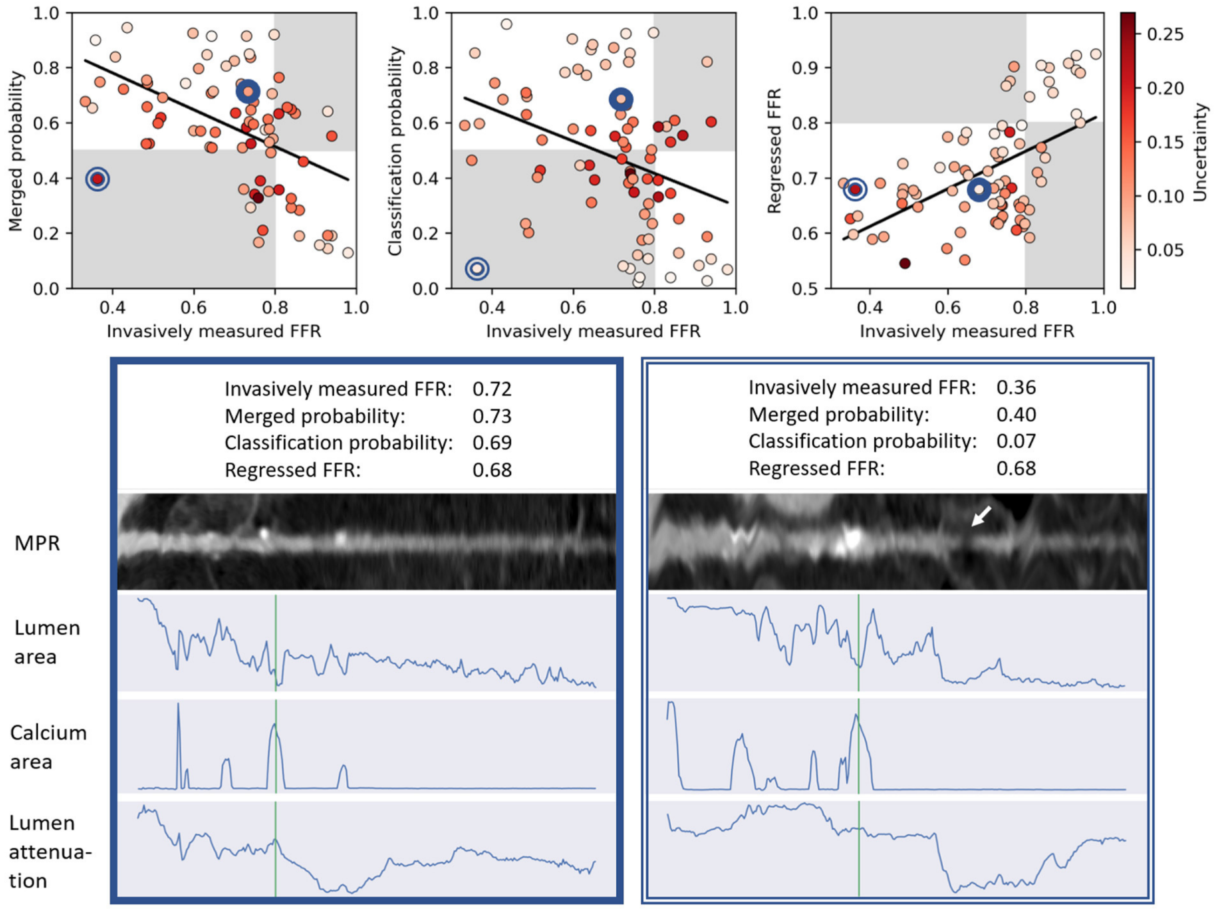
We evaluated the performance of our method on the two held out test sets, using the ensemble of 10 trained networks (Table 3). The method achieved an AUC of 0.78 on $\text{Test}_{\text{Cath}}$ for predicting the presence of functionally significant stenosis in an artery when merging the regression and the classification. In addition, we evaluated the FFR regression and stenosis classification separately. For the $\text{Test}_{\text{Cath}}$ data the results demonstrate that regression outperforms classification and the merged prediction, with an AUC of 0.83. On the patient level, our method achieved an AUC of 0.75 and an accuracy of 0.80.

To investigate the performance of our method on CCTA scans without or with low degree of anatomical coronary artery stenosis, we applied the method to the $\text{Test}_{\text{NoCath}}$ data set. Given that no scan contains functionally significant stenosis, we evaluated the performance in terms of accuracy. When merging classification and regression, the method achieved an accuracy of 0.86. The results demonstrate that for detection of arteries with little or no stenosis in the $\text{Test}_{\text{NoCath}}$ data set, stenosis classification outperforms the FFR regression.

To assess whether the method is robust to the differences in scanner types and acquisition parameters, we investigated the performance per acquisition site on the $\text{Test}_{\text{Cath}}$ data set (Table 4). The best performance was measured for Site 1. Note that the majority of training scans originated from this site (Table 2).

Figure 7 shows the invasively measured reference FFR versus the merged prediction, the classification probability and the regressed FFR. The method tends to be more uncertain in incorrectly classified or regressed arteries. Furthermore, Figure 7 depicts MPRs and predicted characteristics for two arteries.

To evaluate the added value of the uncertainty measure provided by our method, we simulated a semi-automatic setting in which cases with high uncertainty are referred for invasive FFR measurement. This was done by assigning the reference FFR to the 5, 10, or 20% of cases in the $\text{Test}_{\text{Cath}}$ data set with the highest uncertainty (Table 5). The results show that by referring



Invasively measured FFR:	0.72
Merged probability:	0.73
Classification probability:	0.69
Regressed FFR:	0.68

Invasively measured FFR:	0.36
Merged probability:	0.40
Classification probability:	0.07
Regressed FFR:	0.68

FIGURE 7
Top: Scatter plots relating the invasively measured FFR with the predictions for each artery from the Test$_{Cath}$ data set. The graph on the left-hand side corresponds to the merged output probability, in the graph in the middle the output probability from the classification head is shown and in the graph on the right-hand side the regressed FFR value is depicted. Points are colored in red according to their prediction uncertainty. Background colors indicate in which arteries the functional significance was assessed correctly (white) or incorrectly (gray). Whereas, for probabilities (left and middle), high values correspond to the positive class, for regression (right), low output values correspond to the positive class. Black lines show the linear fit to the data. **Bottom:** MPRs and predicted characteristics for two arteries (positions in scatter plots indicated by blue circles). The location of the annotated lesion is plotted in green. Whereas the merged probability assigned to the artery on the left corresponds to the correct class, for the artery on the right, output of the classification head was strongly negative (low probability for functionally significant stenosis), which when combined with the regressed FFR caused the merged probability to yield the incorrect class. Incorrect output of the classification head may be related to a visually minor step-and-shot artifact causing a low intensity section in the MPR on the right (indicated by arrow).

20% of the cases, a sensitivity of 0.92 with a specificity of 0.78 was reached.

5.3. Contribution of artery characteristics

To determine the specific importance of each regressed characteristic, we trained and evaluated models that each only get a single characteristic as input. Additionally, the tree characteristics (bifurcation and the side-branch) were used in each network. In Table 6, the obtained performances for the Test$_{Cath}$ data set are compared with the proposed method. The model with lumen as input performed best among the networks using only a single characteristic and the proposed method outperformed all tested models. Excluding the tree characteristics yielded a slight performance decrease.

5.4. Comparison with previous work

Table 7 compares the performance of our method with performances of previous methods determining presence of functionally significant coronary artery stenosis, as reported in the original works. However, note that these algorithms are not publicly available, and that all of the methods were trained and tested with different proprietary data sets. Hence, the

TABLE 5 Performance of our method on $\text{Test}_{\text{Cath}}$ when a percentage of cases with the highest uncertainty is excluded or corrected to the reference FFR.

		AUC	Accuracy	Sensitivity	Specificity
All arteries		0.78	0.79	0.84	0.61
Excluded	5 %	0.79	0.81	0.86	0.61
	10%	0.81	0.83	0.89	0.63
	20%	0.84	0.86	0.90	0.69
Corrected	5%	0.80	0.81	0.87	0.61
	10%	0.85	0.85	0.90	0.67
	20%	0.91	0.89	0.92	0.78

TABLE 6 Contribution of individual characteristics.

Characteristics	AUC	Accuracy	Sensitivity	Specificity	*p*
All	0.78	0.79	0.84	0.61	-
No lumen	0.72	0.71	0.76	0.56	0.039
No calcium	0.73	0.54	0.52	0.61	0.152
No attenuation	0.76	0.76	0.81	0.61	0.191
No tree	0.76	0.74	0.79	0.56	0.380

Rows correspond to the proposed approach, networks trained on artery characteristic [with the characteristics from the artery tree (bifurcation and side-branch)], and a network trained on all characteristics apart from the tree characteristics (no tree). Among the separate artery characteristics, the network trained on the lumen area performed best. Including all characteristics in the proposed approach lead to the best performance. Excluding the tree characteristics resulted in a slight decrease in performance. *p*-values indicate the statistical significance of AUC improvements of the model using all characteristics over the model using the respective characteristic.

differences in the reported performance can only be seen as an indication. For each method, we indicate whether it requires the segmentation of the coronary artery at test time. The comparison shows that methods that use the artery segmentation at test time attain higher performances. However, artery segmentation is a highly challenging task and results from potentially used automatic methods require manual correction. This manual correction is a time consuming process, leading to excessive analysis times. Our method outperforms the methods that like the proposed method do not use the artery segmentation at test time.

6. Discussion

We presented a deep learning method that assesses the presence of functionally significant stenosis in an artery from CCTA. The method first extracts relevant characteristics from the artery's MPR by regressing the lumen area, its attenuation and the amount of calcifications, and extracting characteristics of the artery within the coronary artery tree. Subsequently, using the extracted characteristics, regression of the FFR value and classification of the presence of a functionally significant stenosis in the artery are performed and thereafter merged to obtain the final result.

The primary target population consisted of patients with an intermediate or high anatomical degree of stenosis ($\text{Test}_{\text{Cath}}$), since these patients typically undergo invasive FFR measurement. Additionally, we investigated the performance of our method in patients with no or low degree of stenosis according to the clinically determined CAD-RADS score ($\text{Test}_{\text{NoCath}}$). In order to make analysis in a large set feasible, we restricted evaluation to the main coronary arteries.

Results demonstrate that regression performs better in the population with intermediate or high anatomical degree of stenosis ($\text{Test}_{\text{Cath}}$), while classification performs better in the population with low degree of anatomical stenosis or without stenosis ($\text{Test}_{\text{NoCath}}$). To combine the strengths of both approaches and obtain robust overall performance, in this work the outputs were merged. However, in a clinical setting, the classification or regression output could be used depending on the target population. The accuracy attained on this set was higher than on $\text{Test}_{\text{Cath}}$, demonstrating that arteries with FFR distributed around the threshold of 0.8, i.e., arteries from our primary target population, are more difficult to assess than arteries with little or no stenosis.

Literature shows that methods for predicting the presence of functionally significant stenosis from CCTA that require coronary artery segmentation achieve high performance (8, 10, 22, 24, 25, 47). However, since the performance is heavily dependent on the quality of the coronary artery segmentation, these approaches typically require manual correction of the segmentation, leading to extensive analysis times. Therefore, methods have been developed that omit the highly challenging segmentation task, leading to fast analysis. In a first investigation, Denzinger et al. (30) showed promising results for end-to-end prediction of the revascularization decision based on functional tests different from FFR in a predominantly negative population. Apart from this, methods that predict FFR without using the artery segmentation typically extract features in an unsupervised manner (31–33). These methods have not been shown to reach the same level of performance as the methods that exploit coronary artery segmentation. Hence, to incorporate information that has been shown to be important for FFR prediction (16, 20–22, 24, 26) while retaining fast analysis, we extract information directly from the MPR in a supervised manner. To do this, a limited number of artery segmentations is used to obtain reference characteristics for training a network to directly predict features characterizing the arteries at test time. During inference, our method does not require the artery segmentation and therefore, the method remains fast at inference.

While previous works used unsupervised feature extraction to describe the arteries, these features were not directly optimized to determine the FFR value (31, 32). As in previous RNN-based works (25, 26, 30), in this work extraction of

TABLE 7 Comparison of the performance on the $\text{Test}_{\text{Cath}}$ data set with previous work.

	Algorithm	Artery segmentation	Analysis level	Analysis time	Samples train	Samples test	AUC	Accuracy
Proposed	ML	no	Arteries	$<$ 5 min	438	76	0.78	0.79
Proposed	ML	no	Patients	$<$ 5 min		67	0.75	0.80
Denzinger et al. (30) ‡	ML	no	Arteries		345	*	0.88	0.87
Zreik et al. (31)	ML	no	Arteries		192	*	0.62	
Kumamaru et al. (33)	ML	no	Patients		131	*	0.69	0.71
Zreik et al. (32)	ML	no	Patients		126	*	0.74	0.70
Nørgaard et al. (8)	CFD	yes	Arteries	1–4 h	-	254	0.90	0.81
Itu et al. (21)	ML	yes	Arteries		simulations	125	0.90	0.83
Dey et al. (22)	ML	yes	Arteries		254	*	0.84	
Tesche et al. (10)	CFD	yes	Arteries	43 min	-	85	0.91	
Tesche et al. (10)	ML	yes	Arteries	41 min	Simulations	85	0.91	
Coenen et al. (11)	CFD	yes	Arteries	$>$30–60 min	Simulations	525	0.84	
Coenen et al. (11)	ML	yes	Arteries	$>$30–60 min	Simulations	525	0.84	
Ko et al. (47)	CFD	yes	Arteries		-	96	0.89	0.84
von Knebel Doeberitz et al. (23)	ML	yes	Arteries		Simulations	84	0.83	
von Knebel Doeberitz et al. (23)	ML + CFD	yes	Arteries		Simulations	84	0.93	
Wang et al. (25)	ML	yes	Arteries		Simulations	71	0.93	0.89
Gao et al. (26)	ML	yes	Arteries		Simulations	180	0.93	
Yang et al. (24)	ML	yes	Arteries		1,013	†	0.80	

*Cross-validation experiments.
†Bootstrap experiments.
‡Predicted FFR is compared to revascularization decision instead of invasive FFR.
Methods are categorized into using machine learning (ML) or computational fluid dynamics (CFD). Methods that use the artery segmentation at test time occasionally require manual interaction. Analysis times include the time needed for manual interaction.

features characterizing arteries and classification of the arteries are optimized together in an end-to-end fashion. However, unlike Wang et al. (25) and Gao et al. (26), we do not use pre-designed high level input features like the degree of stenosis or the lesion length. Instead, we use convolutions to locally encode the low level artery characteristics, enabling the model to learn high level features itself. Moreover, to model the interaction between proximal and distal artery segments, we include a transformer layer that enables learning global features. Furthermore, to regress the FFR value, sequential vascular resistance was modeled by adding up local pressure drops. Incorporating these inductive biases into the network enables targeted feature extraction for prediction, thereby reducing the amount of irrelevant parameters in the model. Together with a small number of descriptive characteristics per centerline point, this targeted model design mitigates the risk of potential overfitting and hence, enables end-to-end learning of high and low level features with limited training data. These features are learned using the predicted characteristics as input, which in some locations inhibit noise (see Supplementary materials). Therefore, our automatically learned features might be more robust to potential noise in the predicted characteristics than the pre-designed features used by Wang et al. (25) and Gao et al. (26). Nevertheless, training the characterization network with a larger data set of manually segmented lumen and calcium might improve the performance of our method.

To investigate the role of each characteristic, we trained additional models only on single artery characteristics. The results showed that the models using all characteristics but one reach reasonable performance and only omitting the lumen area lead to a statistically significant drop in performance ($p < 0.05$, Table 6). This is in line with previous research that underlines the importance of clinical indices derived from these characteristics for FFR prediction (14, 16). Including all characteristics in the proposed method yielded the highest performance, indicating that the extracted artery characteristics contain complementary information. Nevertheless, the proposed method is not limited to the used characteristics. Future work should investigate whether using additional characteristics, like the amount of non-calcified plaque, plaque composition, luminal diameter and artery remodeling (16, 22, 24), would further improve the performance. Furthermore, using unsupervised features (32) in addition to the targeted characteristics may be valuable as it may additionally enable extracting information that has not yet been discovered to be clinically relevant.

To identify possible causes of errors in the detection, arteries from $\text{Test}_{\text{Cath}}$ with the largest difference between the regressed and invasively measured FFR were inspected. We found that in these arteries errors in the extraction of their characteristics were made. They frequently corresponded to overestimation of the calcium area and accordingly, underestimation of the lumen area. This indicates that although the proposed method does not model lesions explicitly, it is sensitive to errors in the artery characterizations that resemble lesions. Therefore, to further improve the performance, future work could focus on improving the artery characterizations.

By employing multiple networks in an ensemble, the robustness of our method was increased and the uncertainty of the predictions was determined (43). The uncertainty measure may be valuable in clinical practice where the method could be employed in a semi-automatic setting. In particular, patients with arteries in which the method indicates high prediction uncertainty could be referred for invasive measurements.

Separate evaluation of the method on the data from each site showed that the best performance was attained for patients scanned at Site 1. This may be caused by the fact that the training set contained most (57%) of arteries from that site. Lower performance for underrepresented sites (Site 2 and Site 3), might have been caused by differences in scanner types and acquisition parameters. Furthermore, for patients from Site 2 and Site 3, the typical time interval between the CCTA acquisition and the FFR measurement was larger compared to Site 1, which may have introduced additional noise. Another reason for performance differences between sites may relate to differences in the protocol for measuring FFR. To only account for proximal measurement positions, pressure drop contributions distal to the estimated measurement position were masked. However, the measurement location may vary between the experts and this may have caused noise in the data which may have negatively impacted performance. Using a larger, more diverse data set will likely enable improved performance for the currently underrepresented sites.

Results in this work show that when the decision threshold is optimized for high sensitivity, our method enables sparing unnecessary FFR measurement in 44% of patients with intermediate degree of stenosis while detecting 95% of functionally significant stenoses (Supplementary materials). Alternatively, combining the proposed method with expert CCTA reading may improve the performance of non-invasive detection of significant stenosis from CCTA (48). While visual assessment of CCTA by an expert radiologist has been reported to have consistently high sensitivity for detection of obstructive CAD (5, 6), it suffers from limited specificity for indicating the functional significance of a stenosis. By specifically optimizing the decision threshold, the proposed method can potentially complement the high sensitivity of expert CCTA reading with high specificity. Future work could evaluate the clinical value of automatic stenosis assessment using the proposed method in combination with expert CCTA reading.

This study has several limitations. First, a relatively small number of scans with corresponding invasive FFR measurements was retrospectively included. While data was acquired in multiple hospitals, the hospitals were not represented equally in the data set. Future work should investigate potential improvements of our method when trained on a larger dataset, equally distributed across hospitals. To avoid biases in the test data, a large-scale (prospective) study in multiple centers is required to confirm the findings. Second, 13% of patients were excluded due to lacking image quality. This may have introduced a selection bias toward patients with preferable externalities, i.e., sinus rhythm and low body-mass-index, which may have caused exclusion of patients at higher risk of significant stenosis. Third, comparison of our method with previous work can only be seen as an indication, as each method was developed and tested on different data sets. At last, we tested our trained method on arteries with no or low degree of stenosis according to the clinically determined CAD-RADS score assuming an FFR > 0.8. However, it can not be fully excluded that despite the clinical stenosis assessment a small number of these arteries have FFR ≤ 0.8, e.g., due to diffuse CAD. Nevertheless, given the high sensitivity of visual assessment of CCTA for detection of CAD (5, 6), we expect this effect to be marginal.

7. Conclusion

We presented a deep learning approach for assessment of the functional significance of coronary artery stenosis from CCTA. Results demonstrate that the proposed approach outperforms previous works that do not require the artery segmentation as input. This indicates that the method may reduce the number of patients that unnecessarily undergo invasive measurements.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to n.hampe@amsterdamumc.nl.

Ethics statement

The studies involving human participants were reviewed and approved by Onze Lieve Vrouwziekenhuis, Moorselbaan 164, 9300 Aalst, Belgium for Site 1; University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, Netherlands for Site 2 and University Medical Center Amsterdam, Meibergdreef 9, 1105 AZ Amsterdam, Netherlands for Site 3. Patients/participants provided their written informed consent to participate in this study (Site 1) or informed consent was waived by the respective institutional review boards (Site 2, Site 3).

Author contributions

II and TL acquired the funding. NH, SV, and II designed the method, experiments, and drafted this manuscript. NH performed the experiments and analyzed the results. All authors critically revised this manuscript and approved the submitted version.

Funding

This work was supported by PIE Medical Imaging BV.

Conflict of interest

Author CC reports receiving institutional research grants from GE Healthcare, Siemens, Insight Lifetech, Coroventis Research, Medis Medical Imaging, Pie Medical Imaging, CathWorks, Boston Scientific, HeartFlow, Abbott Vascular, and consultancy fees from HeartFlow, Abbott Vascular, and Cryotherapeutics. Author II reports institutional research grants by Pie Medical Imaging, Dutch Technology Foundation with participation of Pie Medical Imaging and Philips Healthcare (DLMedIA P15-26). Author J-PA was employed by Pie Medical Imaging BV.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcvm.2022.964355/full#supplementary-material

References

1. Roth GA, Johnson C, Abajobir A, Abd-Allah F, Abera SF, Abyu G, et al. Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *J Am Coll Cardiol.* (2017) 70:1–25. doi: 10.1016/j.jacc.2017.04.052

2. Roth GA, Abate D, Abate KH, Abay SM, Abbafati. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet.* (2018) 392:1736–88. doi: 10.1016/S0140-6736(17)32152-9

3. Pijls NHJ, Fearon WF, Tonino PAL, Siebert U, Ikeno F, Bornschein B, et al. Fractional flow reserve versus angiography for guiding percutaneous coronary intervention in patients with multivessel coronary artery disease: 2-year follow-up of the FAME (Fractional flow reserve versus angiography for Multivessel Evaluation) study. *J Am Coll Cardiol.* (2010) 56:177–84. doi: 10.1016/j.jacc.2010.04.012

4. Pijls NHJ, Tanaka N, Fearon WF. Functional assessment of coronary stenoses: can we live without it? *Eur Heart J.* (2013) 34:1335–44. doi: 10.1093/eurheartj/ehs436

5. Meijboom WB, Van Mieghem CAG, van Pelt N, Weustink A, Pugliese F, Mollet NR, et al. Comprehensive assessment of coronary artery stenoses: computed tomography coronary angiography versus conventional coronary angiography and correlation with fractional flow reserve in patients with stable angina. *J Am Coll Cardiol.* (2008) 52:636–43. doi: 10.1016/j.jacc.2008.05.024

6. Ko BS, Cameron JD, Leung M, Meredith IT, Leong DP, Antonis PR, et al. Combined CT coronary angiography and stress myocardial perfusion imaging for hemodynamically significant stenoses in patients with suspected coronary artery disease: a comparison with fractional flow reserve. *JACC Cardiovasc Imaging.* (2012) 5:1097–111. doi: 10.1016/j.jcmg.2012.09.004

7. Taylor CA, Fonte TA, Min JK. Computational fluid dynamics applied to cardiac computed tomography for noninvasive quantification of fractional flow reserve: scientific basis. *J Am Coll Cardiol.* (2013) 61:2233–41. doi: 10.1016/j.jacc.2012.11.083

8. Nørgaard BL, Leipsic J, Gaur S, Seneviratne S, Ko BS, Ito H, et al. Diagnostic performance of noninvasive fractional flow reserve derived from coronary computed tomography angiography in suspected coronary artery disease: the NXT trial (Analysis of coronary blood flow using CT angiography: next steps). *J Am Coll Cardiol.* (2014) 63:1145–55. doi: 10.1016/j.jacc.2013.11.043

9. Tesche C, De Cecco CN, Albrecht MH, Duguay TM, Bayer RR, Litwin SE, et al. Coronary CT angiography-derived fractional flow reserve. *Radiology.* (2017) 285:17–33. doi: 10.1148/radiol.2017162641

10. Tesche C, De Cecco CN, Baumann S, Renker M, McLaurin TW, Duguay TM, et al. Coronary CT angiography-derived fractional flow reserve: machine learning algorithm versus computational fluid dynamics modeling. *Radiology.* (2018) 288:64–72. doi: 10.1148/radiol.2018171291

11. Coenen A, Kim YH, Kruk M, Tesche C, De Geer J, Kurata A, et al. Diagnostic accuracy of a machine-learning approach to coronary computed tomographic angiography-based fractional flow reserve: result from the MACHINE consortium. *Circ Cardiovasc Imaging.* (2018) 11:e007217. doi: 10.1161/CIRCIMAGING.117.007217

12. von Knebel Doeberitz PL, De Cecco CN, Schoepf UJ, Albrecht MH, van Assen M, De Santis D, et al. Impact of coronary computerized tomography angiography-derived plaque quantification and machine-learning computerized tomography fractional flow reserve on adverse cardiac outcome. *Am J Cardiol.* (2019) 124:1340–8. doi: 10.1016/j.amjcard.2019.07.061

13. Wong DTL, Ko BS, Cameron JD, Nerlekar N, Leung MCH, Malaiapan Y, et al. Transluminal attenuation gradient in coronary computed tomography angiography is a novel noninvasive approach to the identification of functionally significant coronary artery stenosis: a comparison with fractional flow reserve. *J Am Coll Cardiol.* (2013) 61:1271–9. doi: 10.1016/j.jacc.2012.12.029

14. Ko BS, Wong DTL, Nørgaard BL, Leong DP, Cameron JD, Gaur S, et al. Diagnostic performance of transluminal attenuation gradient and noninvasive fractional flow reserve derived from 320-detector Row CT angiography to diagnose hemodynamically significant coronary stenosis: an NXT substudy. *Radiology.* (2016) 279:75–83. doi: 10.1148/radiol.2015150383

15. Diaz-Zamudio M, Dey D, Schuhbaeck A, Nakazato R, Gransar H, Slomka PJ, et al. Automated quantitative plaque burden from coronary CT angiography noninvasively predicts hemodynamic significance by using fractional flow reserve in intermediate coronary lesions. *Radiology.* (2015) 276:408–15. doi: 10.1148/radiol.2015141648

16. Otaki Y, Han D, Klein E, Gransar H, Park RH, Tamarappoo B, et al. Value of semiquantitative assessment of high-risk plaque features on coronary CT angiography over stenosis in selection of studies for FFRct. *J Cardiovasc Comput Tomogr.* (2021) 16:27–33. doi: 10.1016/j.jcct.2021.06.004

17. Gould KL, Lipscomb K, Calvert C. Compensatory changes of the distal coronary vascular bed during progressive coronary constriction. *Circulation.* (1975) 51:1085–94. doi: 10.1161/01.CIR.51.6.1085

18. Dey D, Achenbach S, Schuhbaeck A, Pflederer T, Nakazato R, Slomka PJ, et al. Comparison of quantitative atherosclerotic plaque burden from coronary CT angiography in patients with first acute coronary syndrome and stable coronary artery disease. *J Cardiovasc Comput Tomogr.* (2014) 8:368–74. doi: 10.1016/j.jcct.2014.07.007

19. Hell MM, Dey D, Marwan M, Achenbach S, Schmid J, Schuhbaeck A. Non-invasive prediction of hemodynamically significant coronary artery stenoses by contrast density difference in coronary CT angiography. *Eur J Radiol.* (2015) 84:1502–8. doi: 10.1016/j.ejrad.2015.04.024

20. Ko BS, Wong DTL, Cameron JD, Leong DP, Soh S, Nerlekar N, et al. The ASLA score: a CT angiographic index to predict functionally significant coronary stenoses in lesions with intermediate severity–diagnostic accuracy. *Radiology.* (2015) 276:91–101. doi: 10.1148/radiol.15141231

21. Itu L, Rapaka S, Passerini T, Georgescu B, Schwemmer C, Schoebinger M, et al. A machine-learning approach for computation of fractional flow reserve from coronary computed tomography. *J Appl Physiol.* (2016) 121:42–52. doi: 10.1152/japplphysiol.00752.2015

22. Dey D, Gaur S, Ovrehus KA, Slomka PJ, Betancur J, Goeller M, et al. Integrated prediction of lesion-specific ischaemia from quantitative coronary CT angiography using machine learning: a multicentre study. *Eur Radiol.* (2018) 28:2655–64. doi: 10.1007/s00330-017-5223-z

23. von Knebel Doeberitz PL, De Cecco CN, Schoepf UJ, Duguay TM, Albrecht MH, van Assen M, et al. Coronary CT angiography-derived plaque quantification with artificial intelligence CT fractional flow reserve for the identification of lesion-specific ischemia. *Eur Radiol.* (2019) 29:2378–87. doi: 10.1007/s00330-018-5834-z

24. Yang S, Koo BK, Hoshino M, Lee JM, Murai T, Park J, et al. CT angiographic and plaque predictors of functionally significant coronary disease and outcome using machine learning. *JACC Cardiovascular imaging.* (2021) 14:629–41. doi: 10.1016/j.jcmg.2020.08.025

25. Wang ZQ, Zhou YJ, Zhao YX, Shi DM, Liu YY, Liu W, et al. Diagnostic accuracy of a deep learning approach to calculate FFR from coronary CT angiography. *J Geriatr Cardiol.* (2019) 16:42–8. doi: 10.11909/j.issn.1671-5411.2019.01.010

26. Gao Z, Wang X, Sun S, Wu D, Bai J, Yin Y, et al. Learning physical properties in complex visual scenes: an intelligent machine for perceiving blood flow dynamics from static CT angiography imaging. *Neural Networks.* (2020) 123:82–93. doi: 10.1016/j.neunet.2019.11.017

27. Ghanem AM, Hamimi AH, Matta JR, Carass A, Elgarf RM, Gharib AM, et al. Automatic coronary wall and atherosclerotic plaque segmentation from 3D coronary CT angiography. *Sci Rep.* (2019) 9:47. doi: 10.1038/s41598-018-37168-4

28. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal.* (2017) 42:60–88. doi: 10.1016/j.media.2017.07.005

29. Zreik M, Lessmann N, van Hamersvelt RW, Wolterink JM, Voskuil M, Viergever MA, et al. Deep learning analysis of the myocardium in coronary CT angiography for identification of patients with functionally significant coronary artery stenosis. *Med Image Anal.* (2018) 44:72–85. doi: 10.1016/j.media.2017.11.008

30. Denzinger F, Wels M, Ravikumar N, Breininger K, Reidelshöfer A, Eckert J, et al. Coronary artery plaque characterization from CCTA scans using deep learning and radiomics. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, et al., editors. *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019. Lecture Notes in Computer Science.* Cham: Springer International Publishing (2019). p. 593–601.

31. Zreik M, van Hamersvelt RW, Khalili N, Wolterink JM, Voskuil M, Viergever MA, et al. Deep learning analysis of coronary arteries in cardiac CT angiography for detection of patients requiring invasive coronary angiography. *IEEE Trans Med Imaging.* (2020) 39:1545–57. doi: 10.1109/TMI.2019.2953054

32. Zreik M, Hampe N, Leiner T, Khalili N, Wolterink JM, Voskuil M, et al. Combined analysis of coronary arteries and the left ventricular myocardium in cardiac CT angiography for detection of patients with functionally significant stenosis. In: *Medical Imaging 2021: Image Processing. Vol. 11596.* New Orleans, LA: International Society for Optics and Photonics (2021). p. 115961F.

33. Kumamaru KK, Fujimoto S, Otsuka Y, Kawasaki T, Kawaguchi Y, Kato E, et al. Diagnostic accuracy of 3D deep-learning-based fully automated estimation of patient-level minimum fractional flow reserve from coronary computed tomography angiography. *Eur Heart J Cardiovasc Imaging.* (2020) 21:437–45. doi: 10.1093/ehjci/jez160

34. Li G, Wang H, Zhang M, Tupin S, Qiao A, Liu Y, et al. Prediction of 3D Cardiovascular hemodynamics before and after coronary artery bypass surgery via deep learning. *Commun Biol.* (2021) 4:1–12. doi: 10.1038/s42003-020-01638-1

35. Charles RQ, Su H, Kaichun M, Guibas LJ. PointNet: deep learning on point sets for 3D classification and segmentation. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Honolulu, HI: IEEE (2017). p. 77–85.

36. van den Boogert TPW, Lopes RR, Lobe NHJ, Verwest TA, Stoker J, Henriques JP, et al. Patient-tailored contrast delivery protocols for computed tomography coronary angiography: lower contrast dose and better image quality. *J Thorac Imaging.* (2021) 36:353–59. doi: 10.1097/RTI.0000000000000593

37. Cury RC, Abbara S, Achenbach S, Agatston A, Berman DS, Budoff MJ, et al. CAD-RADS(TM) coronary artery disease - reporting and data system. An expert consensus document of the Society of Cardiovascular Computed Tomography (SCCT), the American College of Radiology (ACR) and the North American Society for Cardiovascular Imaging (NASCI). Endorsed by the American College of Cardiology. *J Cardiovasc Comput Tomogr.* (2016) 10:269–81. doi: 10.1016/j.jcct.2016.04.005

38. Newcombe RTF, Gosling RC, Rammohan V, Lawford PV, Hose DR, Gunn JP, et al. The relationship between coronary stenosis morphology and fractional flow reserve: a computational fluid dynamics modelling study. *Eur Heart J Digit Health.* (2021) 2:616–25. doi: 10.1093/ehjdh/ztab075

39. Wolterink JM, Leiner T, de Vos BD, van Hamersvelt RW, Viergever MA, Išgum I. Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks. *Med Image Anal.* (2016) 34:123–36. doi: 10.1016/j.media.2016.04.004

40. Wolterink JM, Leiner T, Išgum I. Graph Convolutional Networks for Coronary Artery Segmentation in Cardiac CT Angiography. In: Zhang D, Zhou L, Jie B, Liu M, editors. *Graph Learning in Medical Imaging. Lecture Notes in Computer Science.* Cham: Springer International Publishing (2019). p. 62–9.

41. Hampe N, Wolterink JM, Collet C, Planken RN, Išgum I. Graph attention networks for segment labeling in coronary artery trees. In: Landman BA, Išgum I, editors. *Medical Imaging 2021: Image Processing.* Online Only, United States: SPIE (2021). p. 50. Available online at: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11596/2581219/Graph-attention-networks-for-segment-labeling-in-coronary-artery-trees/10.1117/12.2581219.full. (accessed, 2020).

42. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: *Advances in Neural Information Processing Systems. Vol. 30.* Curran Associates Inc. (2017). Available online at: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html. (accessed, 2020).

43. Müller D, Soto-Rey I, Kramer F. An analysis on ensemble learning optimized medical image classification with deep convolutional neural networks. *IEEE Access.* (2022) 10:66467–80. doi: 10.1109/ACCESS.2022.3182399

44. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: *Advances in Neural Information Processing Systems. Vol. 30.* Curran Associates Inc. (2017). Available online at: https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html. (accessed, 2020).

45. Venkatraman ES, Begg CB. A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika.* (1996) 83:835–48. doi: 10.1093/biomet/83.4.835

46. Loshchilov I, Hutter F. Decoupled weight decay regularization. In: *International Conference on Learning Representations-ICLR 2019.* (2019).

47. Ko BS, Linde JJ, Ihdayhid AR, Norgaard BL, Kofoed KF, Sørgaard M, et al. Non-invasive CT-derived fractional flow reserve and static rest and stress CT myocardial perfusion imaging for detection of haemodynamically significant coronary stenosis. *Int J Cardiovasc Imaging.* (2019) 35:2103–12. doi: 10.1007/s10554-019-01658-x

48. van Hamersvelt RW, Zreik M, Voskuil M, Viergever MA, Išgum I, Leiner T. Deep learning analysis of left ventricular myocardium in CT angiographic intermediate-degree coronary stenosis improves the diagnostic accuracy for identification of functionally significant stenosis. *Eur Radiol.* (2019) 29:2350–9. doi: 10.1007/s00330-018-5822-3

frontiers | Frontiers in Cardiovascular Medicine

TYPE Original Research
PUBLISHED 15 December 2022
DOI 10.3389/fcvm.2022.1009445

Deep learning automates detection of wall motion abnormalities *via* measurement of longitudinal strain from ECG-gated CT images

Hui Li[1], Zhennong Chen[1], Andrew M. Kahn[2], Seth Kligerman[3], Hari K. Narayan[4] and Francisco J. Contijoch[1,3]*

[1]Department of Bioengineering, University of California, San Diego, La Jolla, CA, United States, [2]Department of Medicine, Division of Cardiovascular Medicine, University of California, San Diego, La Jolla, CA, United States, [3]Department of Radiology, University of California, San Diego, La Jolla, CA, United States, [4]Department of Pediatrics, University of California, San Diego, La Jolla, CA, United States

Check for updates

OPEN ACCESS

EDITED BY
Tim Leiner,
Mayo Clinic, United States

REVIEWED BY
Declan O'Regan,
Imperial College London,
United Kingdom
Fatma Taher,
Zayed University, United Arab Emirates

*CORRESPONDENCE
Francisco J. Contijoch
fcontijoch@ucsd.edu

SPECIALTY SECTION
This article was submitted to Cardiovascular Imaging, a section of the journal Frontiers in Cardiovascular Medicine

RECEIVED 01 August 2022
ACCEPTED 28 November 2022
PUBLISHED 15 December 2022

CITATION
Li H, Chen Z, Kahn AM, Kligerman S, Narayan HK and Contijoch FJ (2022) Deep learning automates detection of wall motion abnormalities *via* measurement of longitudinal strain from ECG-gated CT images. *Front. Cardiovasc. Med.* 9:1009445. doi: 10.3389/fcvm.2022.1009445

Introduction: 4D cardiac CT (cineCT) is increasingly used to evaluate cardiac dynamics. While echocardiography and CMR have demonstrated the utility of longitudinal strain (LS) measures, measuring LS from cineCT currently requires reformatting the 4D dataset into long-axis imaging planes and delineating the endocardial boundary across time. In this work, we demonstrate the ability of a recently published deep learning framework to automatically and accurately measure LS for detection of wall motion abnormalities (WMA).

Methods: One hundred clinical cineCT studies were evaluated by three experienced cardiac CT readers to identify whether each AHA segment had a WMA. Fifty cases were used for method development and an independent group of 50 were used for testing. A previously developed convolutional neural network was used to automatically segment the LV bloodpool and to define the 2, 3, and 4 CH long-axis imaging planes. LS was measured as the perimeter of the bloodpool for each long-axis plane. Two smoothing approaches were developed to avoid artifacts due to papillary muscle insertion and texture of the endocardial surface. The impact of the smoothing was evaluated by comparison of LS estimates to LV ejection fraction and the fractional area change of the corresponding view.

Results: The automated, DL approach successfully analyzed 48/50 patients in the training cohort and 47/50 in the testing cohort. The optimal LS cutoff for identification of WMA was −21.8, −15.4, and −16.6% for the 2-, 3-, and 4-CH views in the training cohort. This led to correct labeling of 85, 85, and 83% of 2-, 3-, and 4-CH views, respectively, in the testing cohort. Per-study accuracy was 83% (84% sensitivity and 82% specificity). Smoothing significantly improved agreement between LS and fractional area change (R^2: 2 CH = 0.38 vs. 0.89 vs. 0.92).

Conclusion: Automated LV blood pool segmentation and long-axis plane delineation *via* deep learning enables automatic LS assessment. LS values accurately identify regional wall motion abnormalities and may be used to complement standard visual assessments.

KEYWORDS

cardiac computed tomography, left ventricle, wall motion abnormality, longitudinal strain, image segmentation, deep learning

Introduction

Longitudinal strain (LS), measured using echocardiography (1) or cardiac magnetic resonance (2), has been proven useful in evaluating patients at risk of chemotherapy cardiotoxicity (3) and those with aortic stenosis (4, 5), cardiac amyloidosis (6) atrial fibrillation (7), and heart failure patients (8). In revascularized STEMI patients, CMR-based LS was superior and incremental to LVEF and scar size in the prediction of MACE (9).

LS can also be used as a quantitative metric to improve detection of wall motion abnormalities (WMA) (10, 11) and in the setting of infarction WMA have been shown to be independent predictors of adverse events (12, 13). Further, in patients without overt cardiovascular disease, presence of a WMA leads to a 2.4–3.4 higher risk of cardiovascular morbidity and mortality, independent of established risk factors (14).

Cardiac computed tomography (CT) is increasingly used to evaluate both coronary artery anatomy (15, 16) and cardiac function (17). Recent work has shown that ECG-gated CT can detect regional wall motion abnormalities (18–21) and that findings agree with echocardiography (22, 23) and CMR (18, 24). However, quantitative evaluation of cardiac function on 4D CT data can require significant computational processing such as 3D segmentation or measurement of wall thickening.

While several automated methods have been developed for the evaluation of cardiac chamber size and global function (25–28), automated estimation of LS from 4DCT is not currently available as it requires the combination of manual/semi-automated reformatting of the 4D dataset into long-axis imaging planes as well as delineation of the endocardial boundary across frames (29).

Recently, a deep learning framework has been shown to automatically and accurately identify the long-axis planes within a 4D CT dataset and, using the same architecture, segment the LA and LV blood pools (30). Specifically, long-axis views generated *via* the DL method were in close agreement with user-defined planes and >94% of views were diagnostically accurate. By segmenting both the LV and LA blood pools, this creates the opportunity to evaluate LS by measuring the LV endocardial perimeter (after removal of the mitral valve plane).

In this study, we evaluate the ability of this recently developed deep learning algorithm to be adapted to obtain automated LS estimates from each long-axis view. To test the clinical utility of our approach, we evaluated whether automatic LS can be used to detect WMA in a set of 100 clinical cases which were visually analyzed by three trained experts for the presence of WMA. We created two cohorts ($n = 50$ training and $n = 50$ testing cases). We used the training cohort to determine the optimal LS threshold for detecting a WMA and report accuracy in the independent testing cohort.

Methods

Study population

This study was approved by our system's institutional review board with waiver of informed consent. Five hundred and five ECG-gated contrast enhanced cardiac CT studies were acquired between April 2018 and December 2020 which had (1) full R-wave to R-wave (RR) coverage and (2) an imaging report including the explicit mention of cardiac function as normal or abnormal (either globally or regionally) (Table 1). All CT scans were performed on the same wide-detector CT scanner with 256 detector rows and 16 cm z-axis coverage (Revolution scanner, GE Healthcare, Chicago IL).

Visual inspection by (author ZC) resulted in 97 studies being excluded due to poor image quality, lead artifacts which impacted the LV blood pool, or failure to visualize the entire LV.

Imaging reports were used to attempt to balance the study cohort. Two hundred and forty six studies were reported to have "normal" function in the report while 162 were classified as having "abnormal" function. To balance between patients with normal and abnormal function, the studies with normal function acquired at the end of the review period (acquired between August and December 2020, $n = 66$ studies total) were excluded. From the remaining $n = 180$ studies with normal function and $n = 162$ studies with abnormal function, 100 studies were randomly selected. As described below, studies selected were then visually inspected by three experts for the determination of normal/abnormal used in our study. Therefore, this step was aimed at arriving at a relatively balanced distribution of normal and abnormal studies without introducing bias into the selection process. The process is shown as a flowchart in Figure 1.

All studies had functional phases reconstructed at 10% RR intervals using the vendor default cardiac function image

TABLE 1 Patient cohort information.

	Entire dataset	Training cohort	Validation cohort
Cohort size, *n*	100	50	50
Age, years	59 ± 14	59 ± 15	59 ± 13
Male, %	61	58	64
Median LVEF, %	62.4 (IQR: 41.7–69.3)	62.1 (IQR: 38.9–69.6)	63.8 (IQR: 45.1–68.5)
Abnormal segments	27% (432/1,600)	27% (219/800)	27% (213/800)
Normal studies, *n*	54	28	26
Study indication, *n*			
Coronary disease	50	21	29
Pulm. vein ablation	33	19	14
Heart failure	9	4	5
Aortic stenosis	5	4	1
Cardi-oncology	3	2	1

LVEF, left ventricular ejection fraction.

reconstruction method. Images were reconstructed on a 512 × 512 pixel matrix in the axial plane over a field of view of 240 ± 20 mm with 0.625 mm slice thickness.

Expert identification of wall motion abnormalities

The CT studies were independently evaluated for WMA by three cardiovascular imagers with 15 years (A.K), 14 years (SK), and 5 years (HKN) of experience interpreting cardiac studies. For each study, wall motion at 16 AHA segment locations (not including the apical segment) was labeled, in a blinded fashion, as either (1) normal, (2) hypokinetic, (3) akinetic or dyskinetic. This was performed using movie reformats of the 4D CT dataset along standard 2D short- and long-axis views. This led to 1,600 segments being labeled. Given the limited number of hypo- and dyskinetic segments and the interobserver variability, we combined hypokinetic, akinetic and dyskinetic labels into a single "abnormal" class and only performed per-imaging plane and per-study comparison. A long-axis view was considered abnormal if it contained one or more AHA segments that were labeled abnormal. Given that three long axis videos were made per patient, this resulted in 300 long-axis videos (150 in the training and 150 in the testing cohort), each with a normal or abnormal designation. A CT study was classified as abnormal if it had one or more abnormal LAX video. For comparison to our DL-based approach, the three expert scores were combined such that a segment was labeled abnormal if there was agreement by two or more readers.

Automated estimation of longitudinal strain along each long-axis plane

As described by Chen at al. (30), automated blood pool and long-axis views were generated by using a modified U-net architecture. Briefly, the algorithm was first trained to perform blood pool segmentation of the left atrium and ventricle. Then, an output was added after the last max-pooling layer in the downsampling path. This was used to regress the translation vector (to define the spatial position of the long-axis view) and direction vectors (to define the orientation of the view) for each of the long-axis views. The code to perform this segmentation and slice planning is available here: https://github.com/ucsd-fcrl/DL_CT_Seg-Plane_Prediction_Final_v_ZC.

The bloodpool segmentation at each of the long-axis views was evaluated and the left atrial segmentation was used to identify portions of the left ventricle bloodpool which correspond to the mitral valve. Based on this designation, the length of the LV endocardial boundary was calculated. This methodology has been previously been used with echocardiographic imaging (31, 32) and prior work in CT has measured global LS using epicardial contours (33). The process is shown in Figure 2. We expect our approach will more closely match speckle tracking echocardiography (as GLS is measured close to the endocardial boundary) rather than tagged CMR (where evaluation focuses primarily on mid-myocardial deformation) (34). Further, by measuring LS using an automated approach, our method aims to eliminate a significant source of variation (manual contouring by operators) (34).

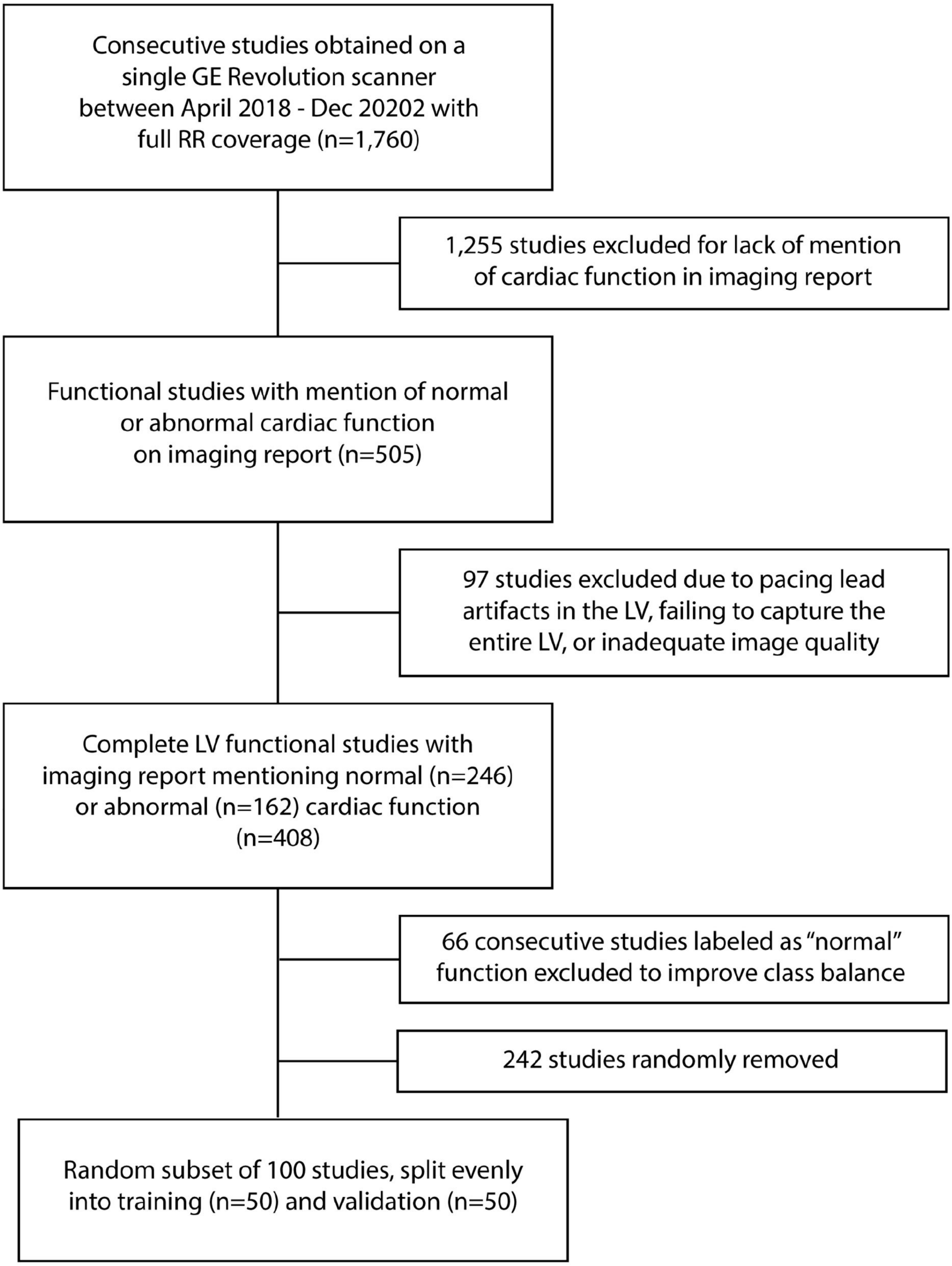

Consecutive studies obtained on a single GE Revolution scanner between April 2018 - Dec 20202 with full RR coverage (n=1,760)
1,255 studies excluded for lack of mention of cardiac function in imaging report
Functional studies with mention of normal or abnormal cardiac function on imaging report (n=505)
97 studies excluded due to pacing lead artifacts in the LV, failing to capture the entire LV, or inadequate image quality
Complete LV functional studies with imaging report mentioning normal (n=246) or abnormal (n=162) cardiac function (n=408)
66 consecutive studies labeled as "normal" function excluded to improve class balance
242 studies randomly removed
Random subset of 100 studies, split evenly into training (n=50) and validation (n=50)

FIGURE 1
Flowchart of patient inclusion/exclusions.

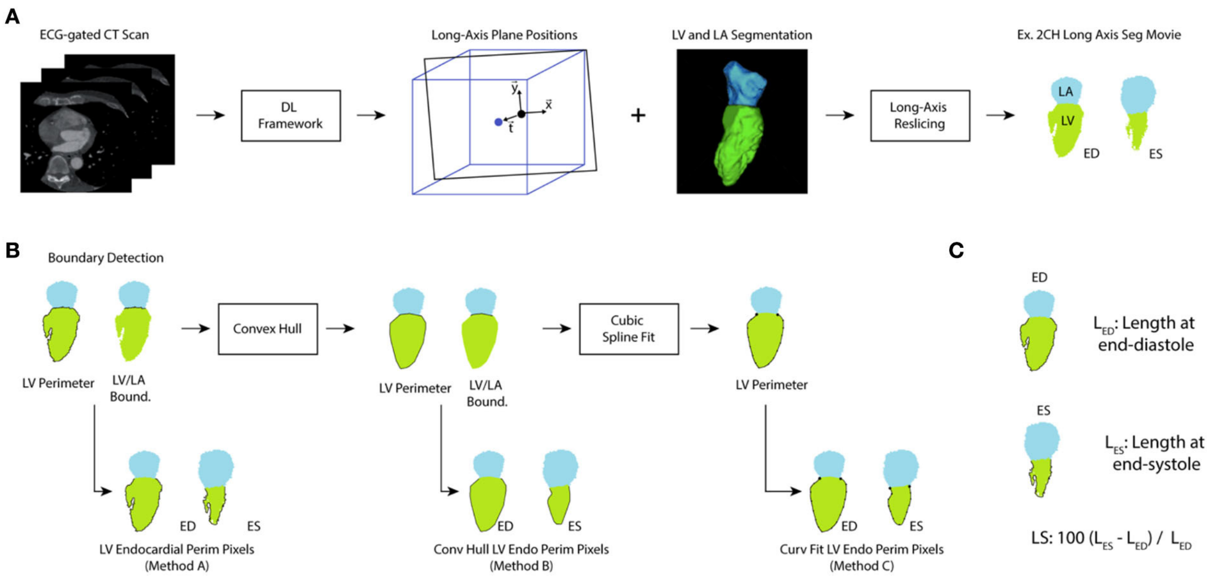

A
ECG-gated CT Scan
DL Framework
Long-Axis Plane Positions
LV and LA Segmentation
Long-Axis Reslicing
Ex. 2CH Long Axis Seg Movie
LA
LV
ED
ES
B
Boundary Detection
LV Perimeter
LV/LA Bound.
Convex Hull
Cubic Spline Fit
LV Endocardial Perim Pixels (Method A)
Conv Hull LV Endo Perim Pixels (Method B)
Curv Fit LV Endo Perim Pixels (Method C)
C
L_{ED}: Length at end-diastole
L_{ES}: Length at end-systole
LS: 100 (L_{ES} - L_{ED}) / L_{ED}

FIGURE 2
Processing of ECG-gated CT for evaluation of LS. **(A)** ECG-gated volumes are analyzed using a deep-learning (DL) framework that provides the location of the 2-chamber, 3-chamber, and 4-chamber long axis planes and delineates the LV and LA blood pools. From this information, long-axis slices of the segmentations were created throughout the cardiac cycle. **(B)** The perimeter of the left ventricle and the LV/LA boundary pixels were identified and used to extract the LV perimeter. Method A did not perform any additional processing of the perimeter. However, a convex hull was applied to correct for papillary muscle artifacts (leading to Method B). Further, a cubic splint was fit to the result of the convex hull to correct for variations in texture (Method C). **(C)** For each long-axis view and each analysis method, the length of the perimeter was measured at end-diastole (the timeframe with largest LV volume) and end-systole (the timeframe with smallest LV volume) and used to calculate LS.

Papillary muscle artifacts and correction approaches

Measuring LS directly from the segmentation was susceptible to artifacts due to the papillary muscles. An example is shown in Figure 3A. Two smoothing approaches were implemented and evaluated, First, the concave areas created by the papillary muscles were "filled in" by using the binary "close" function with a disk of 10 pixels and then fitting a convex hull to the perimeter of the endocardial bloodpool for each frame (35). An example result of this approach is shown in Figure 3.

However, there are limitations with this approach. First, the perimeter measured depends on the "texture" of the surface. This may lead to overestimation of the perimeter. Second, use of the convex hull fills the area of the papillary muscle insertion with a straight line that may underestimate the perimeter. To address these limitations, we fit a "natural" spline curve (36) to the perimeter obtained after closing and filling *via* the convex hull. Fitting was performed after downsampling the curve by a user-defined factor of 5. The result of the three methods, in the same patient as above, is shown in Figure 3. The code used to generate the different LS measures is available here: https://github.com/ucsd-fcrl/DL_CT_GLS_Final.

For all three methods, LS was calculated as the change in length over time. The unsmooth LS result as well as LS after convex hull and convex hull + curve fitting refinement were evaluated by comparing the LS estimate to the LV ejection fraction and the fractional area change (FAC) of the corresponding view.

Determination of LS cutoffs in training cohort and evaluation in testing cohort

We varied the threshold used to determine whether a LS value (for a particular view) accurately detected the presence of a WMA, as determined by our three experts. Using the training cohort ($n = 50$), we identified the thresholds which optimized performance for each LAX view and identified the single threshold that had peak performance when applied to all LAX views. Optimal performance was based on the threshold corresponding to the upper left most point on the receiver operating characteristic (ROC) curve.

The accuracy, sensitivity, and specificity of these thresholds were then evaluated in an independent cohort of $n = 50$ patients.

Statistical evaluation

Normally distributed values are expressed as mean ± standard deviation while non-normal values are reported

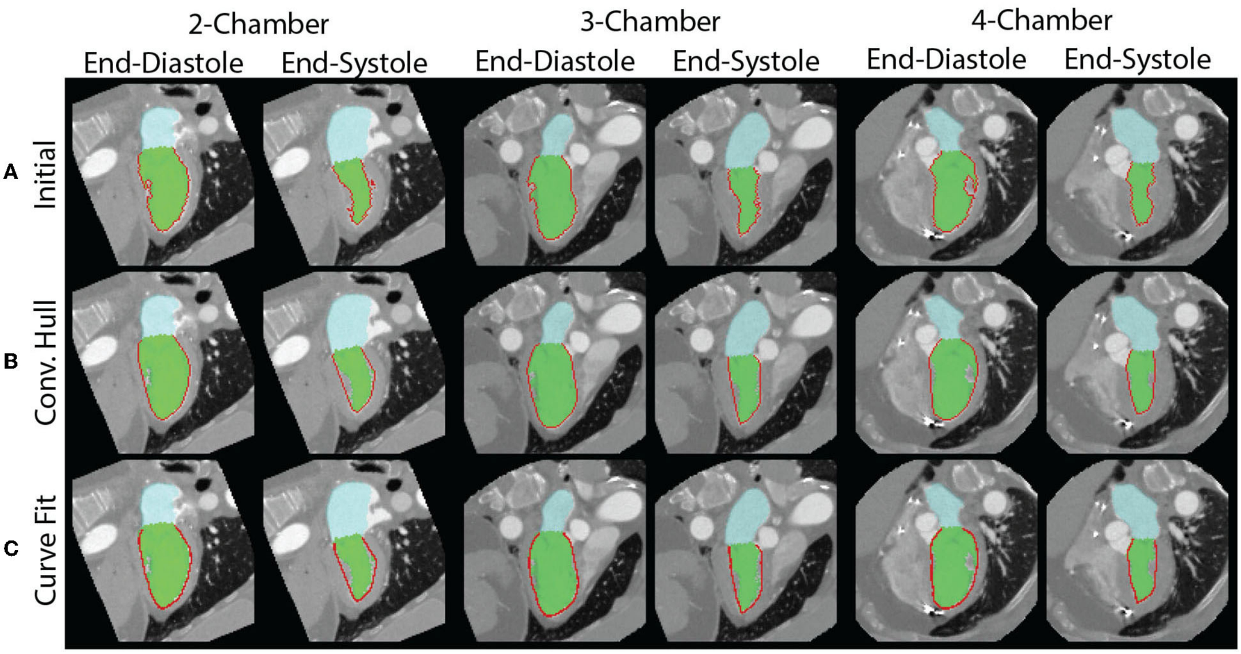


FIGURE 3
Measurement of endocardial perimeter based on the blood pool segmentation is susceptible to artifacts created by the papillary muscles. **(A)** The papillary muscles create indentations which impact the measurement of the perimeter. The end-diastolic (left) and end-systolic (right) perimeters for each of the views are shown. They are all overestimated. **(B)** By modeling the blood pool as a convex hull, we can correct for the errors from indentations created by papillary muscles. However, the perimeter measurement remains affected by the perimeter's texture. **(C)** Fitting of a curve to the perimeter avoids issues related to the surface texture.

using the median and interquartile range (IQR). Two-tailed categorical z-test was used to compare data proportions (e.g., proportions of abnormal videos) in the training and a testing cohort. To compare R^2 values between fractional area change (FAC) and LS for different smoothing methods in dependent samples, the Fisher's r-to-z transformation was utilized to determine statistical significance. Statistical significance was set at $P \leq 0.05$.

The ability of LS to detect WMA was compared against the expert labeled ground truth label and was reported *via* confusion matrix and Cohen's kappa value. Both per-long axis video and per-study comparisons were performed. Readers reviewed long-axis and short-axis movies of the cardiac cycle and labeled each AHA segment. A video was labeled as abnormal if it had one or more abnormal AHA segments present. A study was defined as abnormal if it had one or more long-axis videos labeled as abnormal. Interobserver agreement in terms of labeling wall motion as normal or abnormal between three experts was measured using Fleiss's Kappa (37) since there were more than two observers.

Anonymized long-axis images, calculated perimeters, and corresponding expert annotations will be made available upon request.

Results

Sixty-one subjects were men and 49 were women with a mean age of 59 ± 14. Studies were obtained for evaluation of coronary disease ($n = 50$), pre-ablation assessment of pulmonary vein anatomy ($n = 33$), assessment prior to left ventricular assist device placement ($n = 9$), preoperative assessment for transcatheter aortic valve replacement ($n = 5$), and evaluation of cardiac function after chemotherapy ($n = 3$). The LV blood pool had a median intensity of 530 HU (IQR: 435–663). Out of the 1,600 segments evaluated, 27% (432/1,600) were labeled abnormal by experts. This led to 39.3% (118/300) abnormal long-axis videos and 46 studies with at least one abnormal AHA segment. There were no significant differences (all P-values > 0.05) between the training and testing cohorts in terms of the percentages of sex, abnormal videos, abnormal CT studies.

Median LV ejection fraction (EF) for the training and validation cohorts were 62.1 and 63.8%, respectively. In the training cohort, normal studies had an EF of 69.0% (interquartile range of 65.1–73.0%) while abnormal studies had an EF of 38.1% (IQR: 28.3–48.6%). In the validation cohort, normal studies had an EF of 67.8%

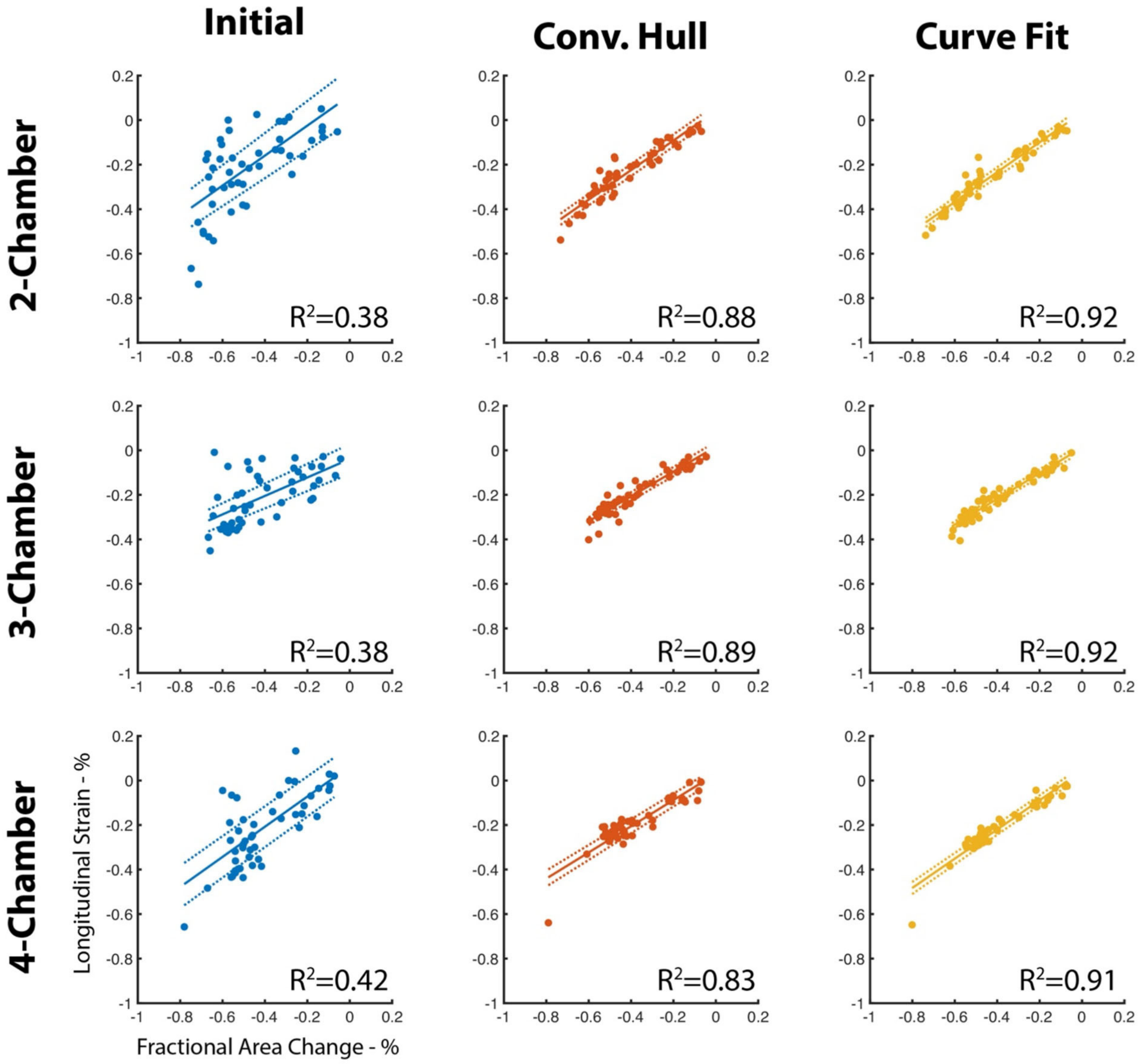


FIGURE 4
Agreement between LS and FAC increases with use of the convex hull and perimeter curve fitting. The perimeter measured using our deep learning method is susceptible to artifacts due to the insertion points of the papillary muscles and by the texture of the endocardial surface. Use of a convex hull to "fill" in the papillary insertions and curve fitting of the surface improves agreement (R^2) with fractional area change of the corresponding long-axis view. Dotted lines represent the 95% confidence interval of the linear fit.

(IQR: 63.6–74.2%) and abnormal studies had an EF of 49.0% (IQR: 26.0–56.0%).

Automated, DL approach successfully analyzed 48/50 patients in the training cohort and 47/50 in the testing cohort. The five failures occurred due to incorrect prediction of long-axis planes. In two of these five cases, the patients had a metal prosthetic mitral valve.

84.6% (1,354/1,600) of segments were labeled identically by all three reviewers. The interobserver agreement amongst the three observers in terms of classifying a segmental wall motion into normal vs. abnormal, measured *via* Fleiss's Kappa, was 0.746, which indicates strong agreement. Fleiss's Kappa for agreement in classifying a LAX video was 0.800 (0.791, 0.811, and 0.797 for the 2, 3, and 4 CH views, respectively) and the value for classifying a patient was 0.786.

Correction for papillary muscle artifacts

The papillary muscle artifacts and the rough endocardial surface led to poor agreement between the fractional area change and longitudinal strain (LS) when LS is measured

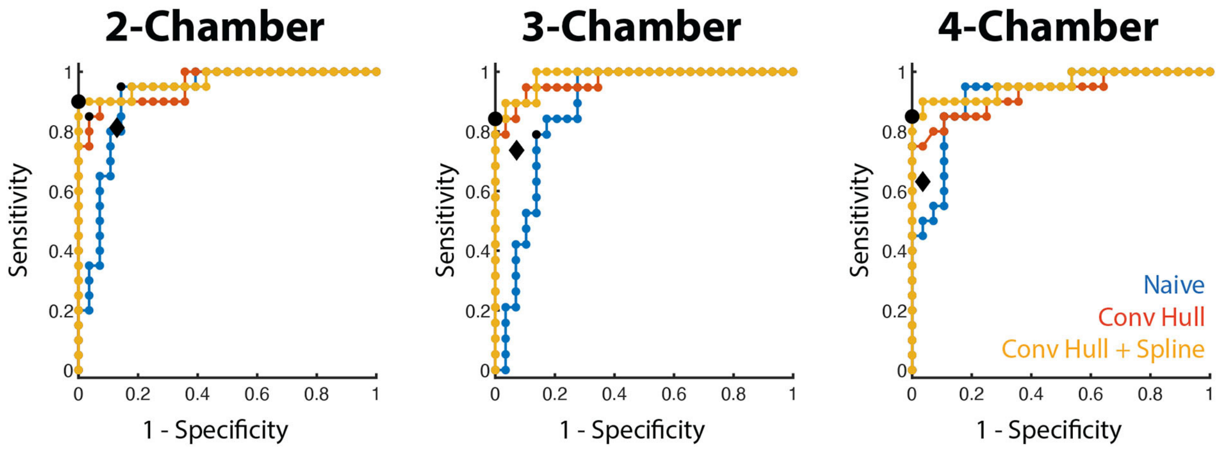


FIGURE 5
WMA classification accuracy using LS in the training cohort. Receiver operating characteristic curves for the three long-axis views are shown for the three LS methods (blue: naive, red: convex hull, orange: convex hull + curve fitting). The optimal operating point for the convex hull with curve fitting is depicted by a black dot. The operating point of the convex hull with curve fitting in the testing cohort is shown by the black diamond.

without use of the convex hull or surface smoothing (Figure 4). Specifically, the R^2 between fractional area change (FAC) and LS is between 0.38 and 0.42 depending on the long-axis view. When the convex hull is used to fill in the voids created by papillary muscles, R^2 increases (0.83–0.89, Figure 4). Curve fitting of the endocardial surface leads to a further increase in R^2 (0.91–0.92, Figure 4). The increase in R^2 was statistically significant ($p < 0.05$) for all views.

Determination of LS cutoffs and classification performance in training cohort

For all long-axis views, the area under the ROC curve using the convex hull and curve fitting was high (0.957–0.984, Figure 5) and the optimal threshold corresponded to a 100% specificity performance, accuracy >91.7% and sensitivity between 84.2 and 90.0% There was a small range of LS thresholds amongst LAX views with a higher cutoff identified for the 2 CH view (−0.218) relative to the 3 and 4 CH views (−0.154 and −0.166, respectively). Per-patient performance (95.8% accuracy, 90.0% sensitivity, 100% specificity) was comparable to the values obtained for each long-axis view.

We also evaluated the ability of a single threshold to classify WMA across all long-axis views. When pooled, LS thresholding had an area under the ROC of 0.965 and the use of −0.170 as the cutoff led to 92.4% accuracy, 83.0% sensitivity, and 100% specificity. This led to 95.8% accuracy, 90.0% sensitivity, and 100% specificity when classifying patients. Complete values are shown in Table 2.

Per-study and per-video classification performance in testing cohort

Using the convex hull and curve fitting approach, we then applied the thresholds identified in the training cohort to the testing population. The accuracy and specificity remained high (>83.0 and >87.1%, respectively) when each view was evaluated independently. Sensitivity ranged between 63.2% (4 CH view) and 81.3% (2 CH view). This led to an overall accuracy in classifying LAX views of 84.4% with a specificity of 92.0%. The use of a single threshold had similar performance (85.1% accuracy, 94.3% specificity). In both the individual and single threshold case, the per-patient accuracy was 83.0% in the testing cohort. Complete values are shown in Table 3.

Discussion

We demonstrate how deep learning (DL) segmentation of the left atrial and left ventricular bloodpools can be combined with automated prediction of the long-axis imaging planes to automatically calculate longitudinal strain along each long-axis view and detect wall motion abnormalities. In this study, we applied the previously trained DL tool to our CT studies without retraining or refinement and developed steps to extract LS from the resulting data. To the best of our knowledge, this is the first study to automatically quantify LS along long-axis views from ECG-gated cardiac CT angiograms. To demonstrate the clinical

TABLE 2 Use of training cohort for identification of LS cutoffs for WMA detection using the curve fitting approach.

		Thresh	AUC	Acc	Sens	Spec	PPV
Individual threshold	2 CH	−0.218	0.970 (0.914–1)	93.8 (89.9–100)	90.0 (76.9–100)	100	100
	3 CH	−0.154	0.984 (0.942–1)	91.7 (83.9–99.5)	84.2 (67.8–100)	100	100
	4 CH	−0.166	0.957 (0.892–1)	91.7 (83.9–99.5)	85.0 (69.4–100)	100	100
	Per-LAX view			92.4 (88.0–96.7)	81.4 (71.4–91.3)	100	100
	Per-patient			95.8 (90.2–100)	90.0 (76.9–100)	100	100
Single threshold	Per-LAX view	−0.170	0.965 (0.930–0.999)	92.4 (88.0–96.7)	83.1 (73.5–92.3)	100	100
	Per-patient			95.8 (90.2–100)	90.0 (76.9–100)	100	100

Thresh, optimal threshold identified for classification; AUC, area under the receiver operating characteristic curve; Sens, sensitivity; Spec, specificity; PPV, positive predictive value; 2 CH, two-chamber view; 3 CH, three-chamber view; 4 CH, four-chamber view; LAX, long-axis view. 95% confidence interval values are given in the parenthesis.

TABLE 3 Performance of LS in the testing cohort using the curve-fitting approach.

		Thresh	Acc	Sens	Spec	PPV
Individual threshold	2 CH	−0.218	85.1 (74.9–95.3)	81.3 (62.1–100)	87.1 (75.3–98.9)	76.5 (56.3–96.6)
	3 CH	−0.154	85.1 (74.9–95.3)	73.7 (53.5–93.5)	92.9 (83.3–100)	87.5 (71.3–100)
	4 CH	−0.166	83.0 (72.2–93.7)	63.2 (41.5–84.9)	96.4 (89.6–100)	92.3 (77.8–100_
	Per-LAX view		84.4 (78.4–90.4)	72.2 (60.3–84.2)	92.0 (86.2–97.7)	84.8 (74.4–95.2)
	Per-patient		83.0 (72.2–93.7)	84.2 (67.8–100)	82.1 (68.0–96.3)	76.2 (58.0–94.4)
Single threshold	Per-LAX view	−0.170	85.1 (79.2–91.0)	70.4 (58.2–82.6)	94.3 (89.4–99.1)	88.4 (78.8–98.0)
	Per-patient		83.0 (72.2–93.7)	79.0 (60.6–97.3)	85.7 (72.8–98.7)	79.0 (60.6–97.3)

Thresh, optimal threshold identified for classification; AUC, area under the receiver operating characteristic curve; Sens, sensitivity; Spec, specificity; PPV, positive predictive value; 2 CH, two-chamber view; 3 CH, three-chamber view; 4 CH, four-chamber view; LAX, long-axis view. 95% confidence interval values are given in the parenthesis.

utility, we evaluated the ability of automated LS to detect WMA. When applied to the testing cohort, the LS identified WMA with accuracy > 83.0% and specificity > 92.9%.

A single LS threshold value of −17.0% had similar performance during the training phase as unique thresholds for each long-axis view and higher performance in the testing cohort. This LS cutoff is similar to those previously reported in other populations and with other imaging methods. In a meta-analysis of chemotherapy-induced cardiotoxicity, Oikonomou et al. reviewed studies which had high-risk cutoff values of −21.0 to −13.8% (3). Similarly, Kearney et al. found LS in controls to be −21 ± 2% while patients with AS had LS between −18 and −15% depending on the AS severity (4) and Zhu et al. found mortality in AS patients was higher in those with LS > −15.2% (5). Recently, Chen et al. reported another automated method to detect wall motion abnormalities using ECG-gated CT which relies on a volume rendering approach (38). Our results are slightly lower than the per-patient accuracy (93.5%), sensitivity (91.9%), and specificity (94.7%) reported in this prior work. This is likely due to the fact that

LS provides a single metric of performance which may mask subtle abnormalities.

This method could add to the clinical interpretation of cardiac CT angiograms by serving as an aid for expert readers. It is also likely that providing the LS score for each view is of value. For example, reporting the LS score along with the relevant cutoff would enable the expert to gain a sense of both the prediction of the algorithm as well as the confidence of the prediction. Also, it is possible that a high sensitivity threshold provides more clinically useful predictions, especially if applied to patients in a screening type of setting. However, this utility is left for future studies. Full R-R ECG-gated imaging has higher dose than obtaining only a single phase. This can be partially mediated by dose modulation. Twenty-five percentage of the studies evaluated in this study had mA reduction of >50% during the cardiac cycle without an impact on clinical interpretability.

While the development of the deep learning segmentation required specialized graphics hardware, the use of the DL and the subsequent LS processing can be easily incorporated into a clinical pipeline and can be readily performed on conventional computers. Further, there are additional metrics that can be readily obtained from this tool, such as the mitral annular plane systolic excursion (MAPSE). However, the extraction and utility of such metrics is left for future studies.

As mentioned, 3D methods to measure endocardial displacement using ECG-gated CT have been previously described (18–21). Solving for endocardial displacement is computationally intensive and delineating the endocardial surface throughout the chamber can be time-intensive. However, recent work aims to avoid these limitations (39). Therefore, our streamlined, automated approach could serve as an initial check to determine whether more extensive assessment is needed.

Our study had several limitations. First, our single site/scanner study only evaluated studies which had global function reported on radiology reports. These factors could introduce biases and motivate a dedicated study to validate our findings in an external, broader cohort across multiple vendors. However, detailed evaluation of wall motion abnormalities in a standardized, AHA segment fashion is not readily available. Second, the DL segmentation failed to produce accurate segmentations and/or long-axis imaging planes in 5/100 patients ($n = 2$ in the training cohort and $n = 3$ in the testing cohort). The 95% success rate is likely sufficient for clinical use, especially given that the result of the DL blood pool segmentation and long-axis planes can be displayed to the reader for review. Our study excluded studies with low image quality, lead artifacts, and incomplete coverage of the LV as the DL method developed by Chen et al. relied on these exclusion criteria (25). Therefore, future work is needed to determine the failure rate in a larger, more diverse, dataset. Further, our approach identifies WMA using LS since the DL segmentation only provides endocardial boundary information. If epicardial segmentations were available, then other metrics such as regional wall thickening could be measured. As a retrospective study, paired echocardiography and MRI data were not available. Future work should directly compare LS measured with CT to these more-conventional methods. Lastly, LS is correlated with other metrics of function such as fractional area change (FAC) and ejection fraction (EF). Our study was not designed nor powered to identify whether LS is a better independent predictor of WMA than these other metrics but others have documented the utility of LS (7, 9).

In conclusion, longitudinal strain (LS), typically measured with MRI or echocardiography, has been previously shown to be diagnostic and prognostic of several patient populations. We leverage a recently developed deep learning approach to automate LS estimation in ECG-gated CT angiograms (cineCT) and demonstrate that LS can be used to detect wall motion abnormalities.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by UCSD Institutional Review Board. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

HL, ZC, and FC performed the data analysis while AK, SK, and HN performed the visual evaluation of the data. All authors contributed to drafting the revising of the manuscript.

Funding

FC was supported by National Institutes of Health grant HL 143113.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Committee Members, Cheitlin MD, Armstrong WF, Aurigemma GP, Beller GA, Bierman FZ, et al. ACC/AHA/ASE 2003 guideline update for the clinical application of echocardiography: summary article: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (ACC/AHA/ASE Committee to Update the 1997 Guidelines for the Clinical Application of Echocardiography). *Circulation.* (2003) 108:1146–62. doi: 10.1161/01.CIR.0000073597.57414.A9

2. Writing Committee Members, Hundley WG, Bluemke DA, Finn JP, Flamm SD, Fogel MA, et al. ACCF/ACR/AHA/NASCI/SCMR 2010 expert consensus document on cardiovascular magnetic resonance: A Report of the American College of Cardiology Foundation Task Force on Expert Consensus Documents. *Circulation.* (2010) 121:2462–508. doi: 10.1161/CIR.0b013e3181d44a8f

3. Oikonomou EK, Kokkinidis DG, Kampaktsis PN, Amir EA, Marwick TH, Gupta D, et al. Assessment of prognostic value of left ventricular global longitudinal strain for early prediction of chemotherapy-induced cardiotoxicity: a systematic review and meta-analysis. *JAMA Cardiol.* (2019) 4:1007. doi: 10.1001/jamacardio.2019.2952

4. Kearney LG, Lu K, Ord M, Patel SK, Profitis K, Matalanis G, et al. Global longitudinal strain is a strong independent predictor of all-cause mortality in patients with aortic stenosis. *Eur Heart J.* (2012) 13:827–33. doi: 10.1093/ehjci/jes115

5. Zhu D, Ito S, Miranda WR, Nkomo VT, Pislaru SV, Villarraga HR, et al. Left ventricular global longitudinal strain is associated with long-term outcomes in moderate aortic stenosis. *Circ Cardiovasc Imaging.* (2020) 13:e009958. doi: 10.1161/CIRCIMAGING.119.009958

6. Phelan D, Collier P, Thavendiranathan P, Popović ZB, Hanna M, Plana JC, et al. Relative apical sparing of longitudinal strain using two-dimensional speckle-tracking echocardiography is both sensitive and specific for the diagnosis of cardiac amyloidosis. *Heart.* (2012) 98:1442–8. doi: 10.1136/heartjnl-2012-302353

7. Russo C, Jin Z, Sera F, Lee ES, Homma S, Rundek T, et al. Left ventricular systolic dysfunction by longitudinal strain is an independent predictor of incident atrial fibrillation: a community-based cohort study. *Circ Cardiovasc Imaging.* (2015) 8:e003520. doi: 10.1161/CIRCIMAGING.115.003520

8. Sengeløv M, Jørgensen PG, Jensen JS, Bruun NE, Olsen FJ, Fritz-Hansen T, et al. Global longitudinal strain is a superior predictor of all-cause mortality in heart failure with reduced ejection fraction. *JACC Cardiovasc Imaging.* (2015) 8:1351–9. doi: 10.1016/j.jcmg.2015.07.013

9. Reindl M, Tiller C, Holzknecht M, Lechner I, Beck A, Plappert D, et al. Prognostic implications of global longitudinal strain by feature-tracking cardiac magnetic resonance in ST-elevation myocardial infarction. *Circ Cardiovasc Imaging.* (2019) 12:e009404. doi: 10.1161/CIRCIMAGING.119.009404

10. Mele D, Pasanisi G, Heimdal A, Cittanti C, Guardigli G, Levine RA, et al. Improved recognition of dysfunctioning myocardial segments by longitudinal strain rate versus velocity in patients with myocardial infarction. *J Am Soc Echocardiogr.* (2004) 17:313–21. doi: 10.1016/j.echo.2003.12.018

11. Cimino S, Canali E, Petronilli V, Cicogna F, De Luca L, Francone M, et al. Global and regional longitudinal strain assessed by two-dimensional speckle tracking echocardiography identifies early myocardial dysfunction and transmural extent of myocardial scar in patients with acute ST elevation myocardial infarction and relatively preserved LV function. *Eur Heart J.* (2013) 14:805–11. doi: 10.1093/ehjci/jes295

12. Carluccio E, Tommasi S, Bentivoglio M, Buccolieri M, Prosciutti L, Corea L. Usefulness of the severity and extent of wall motion abnormalities as prognostic markers of an adverse outcome after a first myocardial infarction treated with thrombolytic therapy. *Am J Cardiol.* (2000) 85:411–5. doi: 10.1016/S0002-9149(99)00764-X

13. Møller JE, Hillis GS, Oh JK, Reeder GS, Gersh BJ, Pellikka PA. Wall motion score index and ejection fraction for risk stratification after acute myocardial infarction. *Am Heart J.* (2006) 151:419–25. doi: 10.1016/j.ahj.2005.03.042

14. Cicala S, de Simone G, Roman MJ, Best LG, Lee ET, Wang W, et al. Prevalence and prognostic significance of wall-motion abnormalities in adults without clinically recognized cardiovascular disease: the strong heart study. *Circulation.* (2007) 116:143–50. doi: 10.1161/CIRCULATIONAHA.106.652149

15. Douglas PS, Hoffmann U, Patel MR, Mark DB, Al-Khalidi HR, Cavanaugh B, et al. Outcomes of anatomical versus functional testing for coronary artery disease. *N Engl J Med.* (2015) 372:1291–300. doi: 10.1056/NEJMoa1415516

16. The SCOT-HEART. Investigators. Coronary CT angiography and 5-year risk of myocardial infarction. *N Engl J Med.* (2018) 379:924–33. doi: 10.1056/NEJMoa1805971

17. Cardiac Computed Tomography Writing Group, Taylor AJ, Cerqueira M, Hodgson JMcB, Mark D, Min J, et al. ACCF/SCCT/ACR/AHA/ASE/ASNC/NASCI/SCAI/SCMR 2010 appropriate use criteria for cardiac computed tomography: A Report of the American College of Cardiology Foundation Appropriate Use Criteria Task Force, the Society of Cardiovascular Computed Tomography, the American College of Radiology, the American Heart Association, the American Society of Echocardiography, the American Society of Nuclear Cardiology, the North American Society for Cardiovascular Imaging, the Society for Cardiovascular Angiography and Interventions, and the Society for Cardiovascular Magnetic Resonance. *Circulation.* (2010) 122:e525–55. doi: 10.1161/CIR.0b013e3181fcae66

18. Pourmorteza A, Chen MY, van der Pals J, Arai AE, McVeigh ER. Correlation of CT-based regional cardiac function (SQUEEZ) with myocardial strain calculated from tagged MRI: an experimental study. *Int J Cardiovasc Imaging.* (2016) 32:817–23. doi: 10.1007/s10554-015-0831-7

19. Pourmorteza A, Schuleri KH, Herzka DA, Lardo AC, McVeigh ER. A new method for cardiac computed tomography regional function assessment: stretch quantifier for endocardial engraved zones (SQUEEZ). *Circ Cardiovasc Imaging.* (2012) 5:243–50. doi: 10.1161/CIRCIMAGING.111.970061

20. Pourmorteza A, Keller N, Chen R, Lardo A, Halperin H, Chen MY, et al. Precision of regional wall motion estimates from ultra-low-dose cardiac CT using SQUEEZ. *Int J Cardiovasc Imaging.* (2018) 34:1277–86. doi: 10.1007/s10554-018-1332-2

21. Contijoch FJ, Groves DW, Chen Z, Chen MY, McVeigh ER. A novel method for evaluating regional RV function in the adult congenital heart with low-dose CT and SQUEEZ processing. *Int J Cardiol.* (2017) 249:461–6. doi: 10.1016/j.ijcard.2017.08.040

22. Tavakoli V, Sahba N. Cardiac motion and strain detection using 4D CT images: comparison with tagged MRI, and echocardiography. *Int J Cardiovasc Imaging.* (2014) 30:175–84. doi: 10.1007/s10554-013-0305-8

23. Buss SJ, Schulz F, Mereles D, Hosch W, Galuschky C, Schummers G, et al. Quantitative analysis of left ventricular strain using cardiac computed tomography. *Eur J Radiol.* (2014) 83:e123–30. doi: 10.1016/j.ejrad.2013.11.026

24. Kaniewska M, Schuetz GM, Willun S, Schlattmann P, Dewey M. Noninvasive evaluation of global and regional left ventricular function using computed tomography and magnetic resonance imaging: a meta-analysis. *Eur Radiol.* (2017) 27:1640–59. doi: 10.1007/s00330-016-4513-1

25. Chen C, Qin C, Qiu H, Tarroni G, Duan J, Bai W, et al. Deep learning for cardiac image segmentation: a review. *Front Cardiovasc Med.* (2020) 7:25. doi: 10.3389/fcvm.2020.00025

26. Litjens G, Ciompi F, Wolterink JM, de Vos BD, Leiner T, Teuwen J, et al. State-of-the-art deep learning in cardiovascular image analysis. *JACC Cardiovasc Imaging.* (2019) 12:1549–65. doi: 10.1016/j.jcmg.2019.06.009

27. Bernard O, Lalande A, Zotti C, Cervenansky F, Yang X, Heng PA, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Trans Med Imaging.* (2018) 37:2514–25. doi: 10.1109/TMI.2018.2837502

28. Vigneault DM, Xie W, Ho CY, Bluemke DA, Noble JA. *Ω*-Net (Omega-Net): fully automatic, multi-view cardiac MR detection, orientation, and

segmentation with deep neural networks. *Med Image Anal.* (2018) 48:95–106. doi: 10.1016/j.media.2018.05.008

29. Cerqueira MD, Weissman NJ, Dilsizian V, Jacobs AK, Kaul S, Laskey WK, et al. Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart. *J Cardiovasc Magn Reson.* (2002) 4:203–10. doi: 10.1081/JCMR-120003946

30. Chen Z, Rigolli M, Vigneault DM, Kligerman S, Hahn L, Narezkina A, et al. Automated cardiac volume assessment and cardiac long- and short-axis imaging plane prediction from electrocardiogram-gated computed tomography volumes enabled by deep learning. *Eur Heart J Digital Health.* (2021) 2:311–22. doi: 10.1093/ehjdh/ztab033

31. Shah AM, Claggett B, Sweitzer NK, Shah SJ, Anand IS, Liu L, et al. Prognostic importance of impaired systolic function in heart failure with preserved ejection fraction and the impact of spironolactone. *Circulation.* (2015) 132:402–14. doi: 10.1161/CIRCULATIONAHA.115.015884

32. Fujikura K, Peltzer B, Tiwari N, Shim HG, Dinhofer AB, Shitole SG, et al. Reduced global longitudinal strain is associated with increased risk of cardiovascular events or death after kidney transplant. *Int J Cardiol.* (2018) 272:323–8. doi: 10.1016/j.ijcard.2018.07.088

33. Marwan M, Ammon F, Bittner D, Röther J, Mekkhala N, Hell M, et al. CT-derived left ventricular global strain in aortic valve stenosis patients: a comparative analysis pre and post transcatheter aortic valve implantation. *J Cardiovasc Comput Tomogr.* (2018) 12:240–4. doi: 10.1016/j.jcct.2018.01.010

34. Amzulescu MS, De Craene M, Langet H, Pasquet A, Vancraeynest D, Pouleur AC, et al. Myocardial strain imaging: review of general principles, validation, and sources of discrepancies. *Eur Heart J Cardiovasc Imaging.* (2019) 20:605–19. doi: 10.1093/ehjci/jez041

35. Mansell DS, Frank EG, Kelly NS, Agostinho-Hernandez B, Fletcher J, Bruno VD, et al. Comparison of the within-reader and inter-vendor agreement of left ventricular circumferential strains and volume indices derived from cardiovascular magnetic resonance imaging. *PLoS ONE.* (2020) 15:e0242908. doi: 10.1371/journal.pone.0242908

36. Lee ETY. Choosing nodes in parametric curve interpolation. *Comp Aided Design.* (1989) 21:363–70. doi: 10.1016/0010-4485(89)90003-1

37. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull.* (1971) 76:378–82. doi: 10.1037/h0031619

38. Chen Z, Contijoch F, Colvert GM, Manohar A, Kahn AM, Narayan HK, et al. Detection of left ventricular wall motion abnormalities from volume rendering of 4DCT cardiac angiograms using deep learning. *Front Cardiovasc Med.* (2022) 9:919751. doi: 10.3389/fcvm.2022.919751

39. Razeghi O, Heinrich M, Fastl TE, Corrado C, Karim R, De Vecchi A, et al. Hyperparameter optimisation and validation of registration algorithms for measuring regional ventricular deformation using retrospective gated computed tomography images. *Sci Rep.* (2021) 11:5718. doi: 10.1038/s41598-021-84935-x

frontiers | Frontiers in Cardiovascular Medicine

TYPE Original Research
PUBLISHED 22 December 2022
DOI 10.3389/fcvm.2022.983868

Check for updates

OPEN ACCESS

EDITED BY
Steffen Erhard Petersen,
Queen Mary University of London,
United Kingdom

REVIEWED BY
Wufeng Xue,
Shenzhen University, China
Edward Ferdian,
The University of Auckland,
New Zealand

*CORRESPONDENCE
Marcel Beetz
marcel.beetz@eng.ox.ac.uk

SPECIALTY SECTION
This article was submitted to
Cardiovascular Imaging,
a section of the journal
Frontiers in Cardiovascular Medicine

RECEIVED 01 July 2022
ACCEPTED 24 October 2022
PUBLISHED 22 December 2022

CITATION
Beetz M, Corral Acero J, Banerjee A, Eitel I, Zacur E, Lange T, Stiermaier T, Evertz R, Backhaus SJ, Thiele H, Bueno-Orovio A, Lamata P, Schuster A and Grau V (2022) Interpretable cardiac anatomy modeling using variational mesh autoencoders. *Front. Cardiovasc. Med.* 9:983868. doi: 10.3389/fcvm.2022.983868

Interpretable cardiac anatomy modeling using variational mesh autoencoders

Marcel Beetz[1]*, Jorge Corral Acero[1], Abhirup Banerjee[1,2], Ingo Eitel[3,4,5], Ernesto Zacur[1], Torben Lange[6,7], Thomas Stiermaier[3,4,5], Ruben Evertz[6,7], Sören J. Backhaus[6,7], Holger Thiele[8,9], Alfonso Bueno-Orovio[10], Pablo Lamata[11], Andreas Schuster[6,7] and Vicente Grau[1]

[1]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, United Kingdom, [2]Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, United Kingdom, [3]University Heart Center Lübeck, Medical Clinic II, Cardiology, Angiology, and Intensive Care Medicine, Lübeck, Germany, [4]University Hospital Schleswig-Holstein, Lübeck, Germany, [5]German Centre for Cardiovascular Research, Partner Site Lübeck, Lübeck, Germany, [6]Department of Cardiology and Pneumology, University Medical Center Göttingen, Georg-August University, Göttingen, Germany, [7]German Centre for Cardiovascular Research, Partner Site Göttingen, Göttingen, Germany, [8]Department of Internal Medicine/Cardiology, Heart Center Leipzig at University of Leipzig, Leipzig, Germany, [9]Leipzig Heart Institute, Leipzig, Germany, [10]Department of Computer Science, University of Oxford, Oxford, United Kingdom, [11]Department of Biomedical Engineering, King's College London, London, United Kingdom

Cardiac anatomy and function vary considerably across the human population with important implications for clinical diagnosis and treatment planning. Consequently, many computer-based approaches have been developed to capture this variability for a wide range of applications, including explainable cardiac disease detection and prediction, dimensionality reduction, cardiac shape analysis, and the generation of virtual heart populations. In this work, we propose a variational mesh autoencoder (mesh VAE) as a novel geometric deep learning approach to model such population-wide variations in cardiac shapes. It embeds multi-scale graph convolutions and mesh pooling layers in a hierarchical VAE framework to enable direct processing of surface mesh representations of the cardiac anatomy in an efficient manner. The proposed mesh VAE achieves low reconstruction errors on a dataset of 3D cardiac meshes from over 1,000 patients with acute myocardial infarction, with mean surface distances between input and reconstructed meshes below the underlying image resolution. We also find that it outperforms a voxelgrid-based deep learning benchmark in terms of both mean surface distance and Hausdorff distance while requiring considerably less memory. Furthermore, we explore the quality and interpretability of the mesh VAE's latent space and showcase its ability to improve the prediction of major adverse cardiac events over a clinical benchmark. Finally, we investigate the method's ability to generate realistic virtual populations of cardiac anatomies and find good alignment between the synthesized and gold standard mesh populations in terms of multiple clinical metrics.

KEYWORDS

mesh VAE, 3D ventricular shape analysis, virtual anatomy generation, clinical outcome prediction, acute myocardial infarction, major adverse cardiac events, graph neural networks, geometric deep learning

1. Introduction

Inter-subject variability in human cardiac anatomy and function plays a decisive role in the accurate diagnosis and treatment of many cardiovascular diseases, including myocardial infarction, heart failure, and reinfarction (1–3). Therefore, it is a key objective of computational models of the heart to be able to capture this variability across a population in order to obtain realistic representations of cardiac morphology and physiology. Such models not only enable a more accurate definition of shape normality within a given subpopulation but also improve the ability to detect abnormalities while retaining interpretability of the diagnosis (4, 5). Cardiac magnetic resonance imaging (MRI) is considered the gold standard imaging modality for the non-invasive assessment of cardiac anatomy and function in clinical practice (6). Accordingly, it has been extensively used as the basis to investigate anatomical shape variability in previous literature. While many works have focused on quantifying anatomical information based on the 2D slices of the underlying cine MRI acquisition (4, 7–10), this provides only an approximation of the heart's true 3D shape and therefore neglects more localized shape variability which is crucial for the detection and diagnosis of various cardiac diseases (2, 11, 12). Consequently, other works have conducted cardiac shape analysis directly on 3D representations of the heart which have either been reconstructed from 2D slices (13–18) or been acquired using a 3D MRI acquisition protocol (19). In order to study the anatomical variability in the obtained 3D heart shapes, principal component analysis (PCA) has been widely used in previous literature, as it allows to easily and quickly identify the most important modes of shape variation within the population (5). This low-dimensional representation of cardiac shape information can then be used for a variety of follow-up tasks, such as to investigate the association of shape and cardiovascular risk factors (20, 21), determine the probability of future major adverse cardiac events (MACE) (2), study the connection between shape and simulated cardiac function (22, 23), generate virtual population cohorts for *in silico* trials (24), or predict myocardial infarction (MI) (12). While the low-dimensional scores are obtained by PCA in an unsupervised manner, supervised methods, such as linear discriminant analysis (LDA) and information maximizing component analysis, have also been proposed to directly take into account information about the task-specific objective in the cardiac anatomy modeling (25).

More recently, deep learning approaches based predominantly on the variational autoencoder (VAE) (26) framework have been increasingly used to capture population-wide anatomical variability for a variety of tasks (5). Similar to PCA, the VAE allows for the representation of 3D shape information in a low-dimensional space with individual components corresponding to different aspects of inter-subject variability. However, in contrast to standard PCA or LDA approaches, VAEs are capable of modeling considerably more complex relations, primarily due to their deep learning-based architecture with a high number of trainable parameters and the presence of non-linear functions. The autoencoder structure with a low-dimensional latent space representation also allows for the straightforward integration with other common image-based tasks while maintaining a good degree of interpretability. Such tasks include the detection of coronary artery disease (27) and hypertrophic cardiomyopathy (28), image segmentation with shape priors (29–32), multi-task segmentation and regression (33), image-to-image synthesis (33), and survival prediction (34). However, the aforementioned approaches mostly rely on representing cardiac shapes as fixed-size 3D voxelgrids and use standard grid-based deep learning operations. This is not only inefficient in terms of memory and time requirements but also complicates effective feature learning when processing anatomical surface data. In order to overcome these issues, geometric deep learning techniques (35) have been introduced to enable accurate learning directly on non-Euclidean data, such as point clouds or graphs. This enables the anatomical surface information of the heart to be represented and processed in a highly efficient manner targeted to the data-type at hand and hence, has seen various applications in cardiac image analysis. For example, point cloud-based deep learning approaches have been proposed for the generation of virtual cardiac anatomies (36), classification of cardiac disease (37), modeling of 3D deformation of the heart (38), surface reconstruction of cardiac anatomy (13, 39), combined reconstruction and segmentation of the left ventricular (LV) wall (40), and the joint modeling of cardiac anatomy and electrocardiogram data (41, 42). Similarly, graph neural networks have been investigated for the simulation of cardiac mechanics (43), reconstruction of cardiac meshes (44), prediction of cardiac depolarization times (45), and the estimation of wall shear stress in 3D artery models (46).

Following these advancements, we propose in this work a variational mesh autoencoder (mesh VAE) as a novel approach to cardiac anatomy modeling. The mesh VAE is specifically designed to work directly on 3D mesh representations of the heart and thus overcomes the limitations of voxelgrid-based approaches. This enables the efficient processing of high-resolution 3D cardiac anatomy data and provides a more accurate modeling of 3D shape variability. The mesh VAE combines graph convolution and mesh sampling layers in a hierarchical setup to allow effective multi-scale feature learning of non-linear relationships. At the same time, the VAE framework ensures a high degree of interpretability with a disentangled, low-dimensional latent space. The architecture is also highly adaptable and can be used in combination with different imaging modalities, disease types, and application domains in a similar way as grid-based autoencoders. In summary, we make the following contributions in this work:

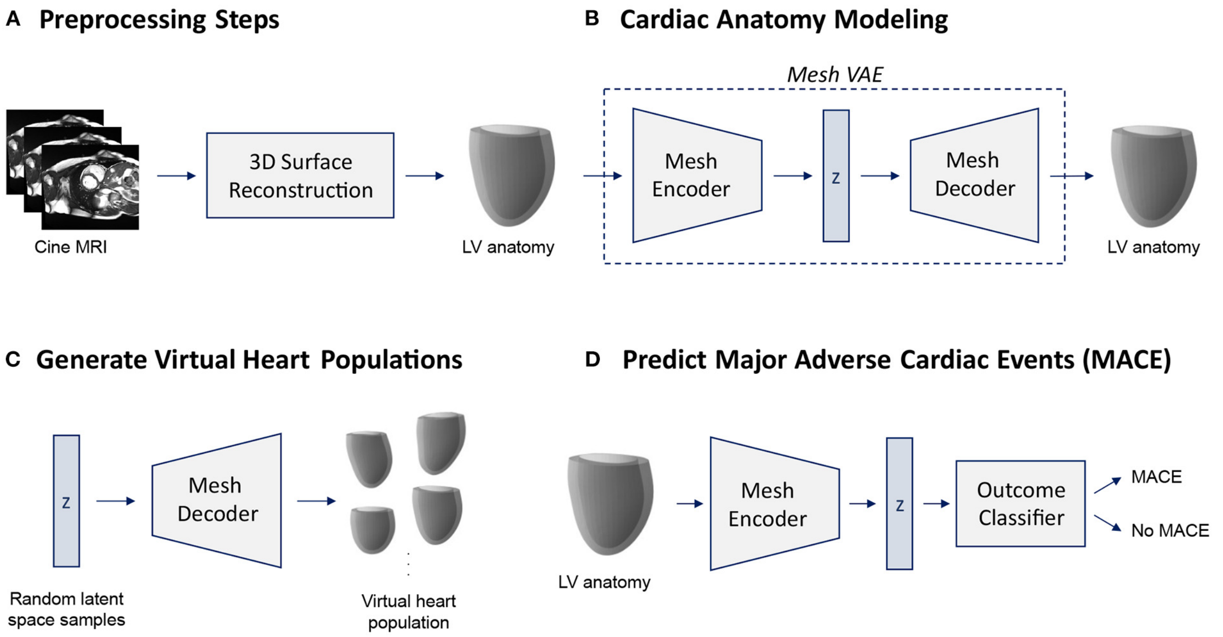


FIGURE 1
Overview of the proposed cardiac shape analysis pipeline **(A,B)** and the two possible applications investigated in this work **(C,D)**. First, we pass raw cine MRI acquisitions through a multi-step 3D surface reconstruction pipeline (Section 2.2) to reconstruct 3D surface meshes of the left ventricular (LV) anatomy as a preprocessing step **(A)**. It consists of an image segmentation step using two cascaded U-Nets followed by a personalized template mesh fitting step based on variational warping techniques (Section 2.2). Next, as the core part of the pipeline, we propose a variational mesh autoencoder with an interpretable latent space *z* to efficiently capture population-wide variability in 3D cardiac shapes in a single geometric deep learning model **(B)** (Section 2.3). The decoder of the pre-trained mesh VAE can then be used to generate virtual population cohorts of 3D heart meshes **(C)** (Section 3.5). The pre-trained encoder converts input meshes into low-dimensional latent space representations which can serve as inputs to a MACE outcome classifier **(D)** (Section 3.6).

- We develop a novel variational mesh autoencoder for 3D cardiac anatomy modeling directly on 3D surface meshes.
- We successfully embed the mesh VAE into a multi-step cardiac anatomy modeling pipeline to enable clinical applicability.
- We evaluate the mesh VAE's ability to reconstruct high-resolution anatomy meshes on a multi-domain cine MRI dataset of myocardial infarction patients at both the end-diastolic (ED) and end-systolic (ES) phases of the cardiac cycle.
- We conduct a comparative analysis of the mesh VAE and a voxelgrid-based VAE benchmark in terms of both their reconstruction capabilities and technical specifications and demonstrate the advantages of the mesh VAE for the processing of anatomical surface data.
- We investigate the latent space of the mesh VAE as an efficient low-dimensional encoding of high-dimensional cardiac surface anatomy information in terms of its interpretability, disentanglement, association with generated output meshes, and accurate representation of inter-subject shape variability.
- We analyze the suitability of the mesh VAE for the generation of realistic virtual population cohorts of cardiac anatomy meshes.
- We explore the utility of the mesh VAE's latent space representations to capture pathology-specific shape biomarkers and predict MACE events in post-MI patients.
- We provide a pertinent literature review and a detailed discussion of the results including the proposed method's limitations and possible future use cases.

2. Materials and methods

In this section, we give an overview of the proposed cardiac shape modeling pipeline (Section 2.1), describe the dataset and preprocessing steps (Section 2.2) used for method development, and explain the architecture (Section 2.3), loss function (Section 2.4), and training procedure (Section 2.5) of the proposed mesh VAE network.

2.1. Overview

In this work, we introduce a novel variational mesh autoencoder embedded into a multi-step pipeline to enable efficient non-linear 3D shape analysis of the human heart (Figure 1).

In the first part of our pipeline, we apply several preprocessing steps to prepare the raw images of our dataset for 3D shape modeling with the mesh VAE (Figure 1A). The input consists of the short-axis (SAX) slices of a standard cine MRI acquisition which we first segment using a two-step cascaded U-Net (47) approach (Section 2.2). Next, we fit a template mesh to the resulting SAX contours in a numerical optimization procedure in order to obtain 3D mesh representations of the cardiac anatomy (Section 2.2). We then use these 3D surface meshes to train and evaluate the proposed mesh VAE to capture cardiac shape variability across the population (Figure 1B). The mesh VAE constitutes the core part of the shape modeling pipeline with its architecture specifically tailored to process complex 3D surfaces of the cardiac anatomy (Section 2.3). This enables a variety of different clinical and research-related use cases. As two possible sample applications, we investigate both the generation of virtual populations of cardiac anatomy meshes using the mesh VAE's decoder (Figure 1C) and the binary prediction of MACE outcomes based on 3D cardiac shape information as encoded in the mesh VAE's latent space representations (Figure 1D). In the following sections, we describe each part of the pipeline in greater detail with a particular emphasis on the proposed mesh VAE.

2.2. Dataset and preprocessing steps

Our dataset consists of 1,021 post-MI patients for which cine MR images were acquired a median of 3 days after the infarction event in a multi-center study. It is based on both the TATORT-NSTEMI trials (Thrombus Aspiration in Thrombus Containing Culprit Lesions in Non-ST-Elevation Myocardial Infarction; NCT01612312) and the AIDA-STEMI trials (Abciximab Intracoronary vs. Intravenously Drug Application in ST-Elevation Myocardial Infarction; NCT00712101) and hence includes both Non-ST-Elevation Myocardial Infarction (NSTEMI) and ST-Elevation Myocardial Infarction (STEMI) patients (48–50). Electrocardiography-gated balanced steady-state free precession sequences were used for all acquisitions. The pixel resolution varied across the acquired images with a mean value of 1.36 mm (range: [1.16, 2.08] mm) and standard deviation (SD) of 0.21 mm. Each patient was followed up for 12 months post-MI with MACE (reinfarction, new congestive heart failure, or all-cause death) defined as the clinical endpoint. Overall, 74 patients experienced a MACE outcome. Further details regarding the study population and image acquisition can be found in (2, 48–50).

We first apply a multi-step preprocessing pipeline to reconstruct 3D surface mesh representations of the left ventricular anatomy from the raw cine MRI acquisitions. The first step of this pipeline consists of the segmentation of left ventricular (LV) myocardium on the cine cardiac MRI using two cascaded U-Nets with enhanced preprocessing (51, 52). The first U-Net locates the LV to crop and orient the images accordingly, while the second U-Net performs the fine segmentation. This architecture addresses both canonical orientation for regional metrics quantification and label imbalance for segmentation performance improvement. Next, two personalized 3D LV meshes at the ED and ES phases are built from the segmentation contours for each patient. The reconstruction of these 3D meshes uses a solution based on smooth cubic Hermite interpolation, where, in brief, an idealized LV template mesh is fitted to the 3D myocardium segmentation mask by combining image registration and mesh projection techniques (17, 53, 54). The Hermite template mesh is an idealized LV (truncated ellipsoid of 6 longitudinal × 12 circumferential × 1 radial elements). Since the same template is used for all the patients, homologous points are directly obtained. Further details on the pipeline can be found in (2). The resulting 3D surface meshes of the left ventricular anatomy are then used as inputs for training and evaluating the mesh VAE. We split the mesh dataset into 70% training, 5% validation, and 25% test datasets while maintaining the same class imbalance between MACE and no MACE cases in each subset. Finally, we apply standardization (i.e., subtracting the mean and dividing by the SD) to each mesh before inputting it into the network.

2.3. Variational mesh autoencoder architecture

As the core part of our shape modeling pipeline, we propose a novel mesh VAE architecture, specifically designed based on recent advances in mesh-based deep learning (55, 56) to efficiently process triangular mesh data of the 3D cardiac anatomy (Figure 2).

The overall architecture consists of an encoder and a decoder connected by an interpretable 16-dimensional latent space with the ability to capture high-dimensional cardiac shape information in a low-dimensional representation. The building blocks of the network follow recent advances in mesh-based deep learning to enable effective learning of non-linear relationships directly on triangular mesh data. The main feature extraction is accomplished by graph convolution blocks which are composed of spectral graph convolutional layers (56) followed by rectified linear unit (ReLU) activation functions. Multiple mesh downsampling operations (55) are positioned between successive graph convolution blocks along the encoder to allow for stepwise decreases in mesh resolution and multi-scale hierarchical learning. The decoder follows a symmetric design to the encoder with mesh upsampling operations interspersed between graph convolution blocks and the same mesh resolutions, number of levels and feature maps as the encoder. This enables the decoder to reconstruct high-resolution anatomical meshes from the latent space in a gradual

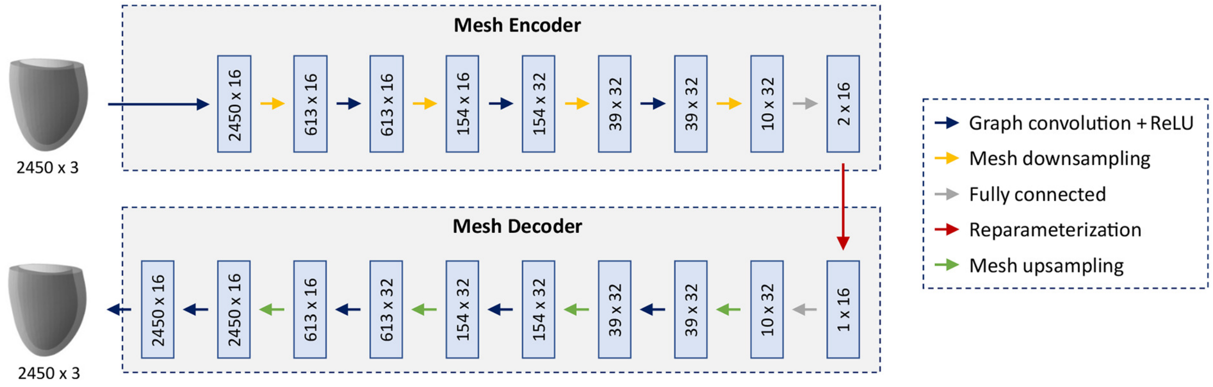


FIGURE 2
Architecture of the mesh VAE. The input and output are cardiac anatomy models represented as 3D triangular meshes. All meshes in the dataset share the same connectivity and consist of 2,450 vertices with associated x,y,z coordinates. Multi-scale feature learning directly on mesh data is enabled by alternating graph convolution and sampling operations which are arranged in a hierarchical setup. The mesh encoder and decoder, connected by a 16-dimensional latent space, follow a symmetric design with the same number of levels and the same mesh resolution per level.

multi-scale process akin to the stepwise encoding operation of the encoder. Fully connected layers are introduced before and after the latent space to connect the tensors representing downsampled mesh information and the latent space in an effective way. All spectral graph convolutions use the Chebyshev polynomial approximation (56, 57) of order 5 for efficient calculation. The mesh sampling operation uses quadric error minimization to identify the vertices that are removed in the downsampling step, then saves their location in barycentric coordinates, before using these coordinate values to reinsert them in the upsampling step (55). All input anatomies are represented as 3D triangular surface meshes with 2,450 vertices and identical vertex connectivity across the dataset.

2.4. Loss function

The loss function of the proposed mesh VAE is based on the β-VAE framework (58) and consists of the sum of a reconstruction loss term and a Kullback-Leiber (KL) divergence term, weighted by a parameter β.

$$L_{total} = L_{reconstruction} + \beta * L_{KL}. \quad (1)$$

The weighting parameter β is used to control the importance of each of the two loss terms during training. Similar to previous approaches for cardiac shape analysis using point cloud deep learning (36), we follow a monotonic annealing schedule (59) for β and set it to small values (starting at 0.0001) at the beginning of training before gradually increasing it until 0.001 at the end of training. This allows the network to put more emphasis on the accurate mesh reconstruction task first and then step-by-step focus more on also achieving a high latent space quality. The Kullback-Leibler divergence term in the total loss function used in this work is defined as follows:

$$L_{KL} = D_{KL}\left[Q(z|X)\|P(z)\right]. \quad (2)$$

Here, X refers to the input mesh, z to the latent space of the mesh VAE, and $Q(z|X)$ to the posterior distribution of the VAE's latent space. $P(z)$ is the prior distribution of the VAE's latent space for which we choose a multivariate standard Gaussian distribution in this work. This helps the VAE to achieve a smooth and disentangled latent space, thus improving the representation of cardiac shape variability.

We select the mean squared error (MSE) between the coordinate values of the corresponding vertices n in the input mesh x and the reconstructed mesh y as our reconstruction loss term. This encourages the VAE to put more emphasis on larger vertex distances between input and ground truth meshes which facilitates the task of capturing the full extent of cardiac shape variability across the population.

$$L_{reconstruction} = \frac{1}{N}\sum_{n=1}^{N}(x_n - y_n)^2 \quad (3)$$

2.5. Network training and implementation

We train the mesh VAE for 250 epochs with a learning rate of 0.001 and a batch size of 8 using the Adam optimizer (60) on a CPU. The reparameterization trick (26) is applied during training. All general deep learning code in this work was based on the PyTorch framework (61), while the PyTorch Geometric library (62) was used for graph-specific deep learning operations. The machine learning classifiers for the experiments in Section 3.6 were implemented using the scikit-learn library (63).

3. Experiments and results

We evaluate the proposed mesh VAE in a variety of different settings using various evaluation metrics (Section 3.1) to showcase its versatility and demonstrate its usefulness in multiple applications related to cardiac shape analysis. These include an assessment of its ability to accurately reconstruct 3D cardiac mesh inputs (Section 3.2), a comparative analysis with a voxelgrid-based deep learning benchmark (Sections 3.2, 3.3), and an investigation of its latent space quality (Section 3.4). In addition, we evaluate its ability to generate realistic virtual population cohorts of cardiac anatomies (Section 3.5) and the utility of its latent space to predict MACE events (Section 3.6) as two possible sample applications of the mesh VAE.

3.1. Evaluation metrics

We utilize multiple metrics in this work to enable a thorough evaluation of the mesh VAE in a variety of settings and tasks.

We select the mean surface distance (MSD) (Equation 4) between two triangular meshes X and Y as our first metric to quantify the averaged difference between two anatomical surfaces. This allows the assessment of the general alignment between our method's predictions and the corresponding gold standard.

$$MSD(X, Y) = \frac{1}{2}\left(\frac{1}{|X|}\sum_{x \in X} d(x, Y) + \frac{1}{|Y|}\sum_{y \in Y} d(y, X)\right) \tag{4}$$

In addition to the average distance between two meshes, we also want to obtain the maximum difference between the two. This allows us to see whether larger deviations are present in smaller regions on the mesh surfaces which is important for a more localized cardiac shape analysis. We choose the Hausdorff distance (HD) between input meshes X and Y for this purpose.

$$HD(X, Y) = \max\left\{\sup_{x \in X}\inf_{y \in Y} d(x, y), \sup_{y \in Y}\inf_{x \in X} d(x, y)\right\} \tag{5}$$

While MSD and HD provide a geometric quantification of mesh alignment, we also aim to provide an assessment in terms of commonly used clinical metrics to facilitate the proposed method's application in clinical practice. To this end, we choose the LV endocardial volume and the LV myocardial mass to validate the anatomical realism of the virtual meshes generated by our method across a given population. Here, we define the LV mass as the difference between LV epicardial and LV endocardial volumes multiplied by a constant ($\rho = 1.05\ g/mL$) to represent the average density of the myocardial tissue.

An advantage of any variational autoencoder architecture is the existence of a latent space that aims to provide an accurate low-dimensional and disentangled representation of the high-dimensional population distribution. In order to improve interpretability and performance on multiple follow-up tasks, each dimension of the latent space ideally encodes a different aspect of the inter-subject anatomical variability. To this end, we quantify the contribution of each latent space dimension u using the activity metric (64).

$$Activity_u = Cov_X(\mathbb{E}_{u \sim q(u|X)}[u]) \tag{6}$$

Here, X refers to the input meshes, $\mathbb{E}$ to the expected value, Cov to the covariance, and q to the posterior probability distribution of the latent space component u. Intuitively, a higher activity score indicates that a larger amount of population-wide shape variability is captured by the given latent space dimension.

Finally, in order to assess the utility of the mesh VAE's latent space representation for the binary MACE prediction task, we select the area under the receiver operating characteristic (AUROC) curve as our metric due to the class imbalance in our dataset.

3.2. Mesh reconstruction

In our first experiments, we want to assess whether the proposed mesh VAE is capable of accurately encoding and reconstructing the complex high-dimensional anatomical meshes at both the ED and ES phases of the cardiac cycle. To this end, we first train separate mesh VAE models for each of the two cardiac phases and evaluate their reconstruction performance qualitatively by comparing the output meshes with the corresponding input meshes of the unseen test dataset. The obtained results for five sample cases are depicted for both ED and ES phases in Figures 3A,B, respectively.

We observe that the shapes of the input and reconstructed meshes closely resemble each other on both a local and global level and for both the ED and ES phases. The relationship between epicardial and endocardial surfaces remains consistent between input and predicted meshes, while the most noticeable differences appear in regions with remaining slice misalignment artifacts and at the base of the left ventricular anatomy. We also notice a slight smoothing effect of the mesh VAE outputs compared to the respective input meshes, especially in localized regions affected by surface reconstruction artifacts.

In order to quantify the mesh VAE's encoding and reconstruction ability, we calculate both the mean surface distance (Equation 4) and the Hausdorff distance (Equation 5) between the predicted meshes and gold standard meshes of the test dataset (Table 1). This enables an assessment of both the average and worst-case performance of how accurately the mesh

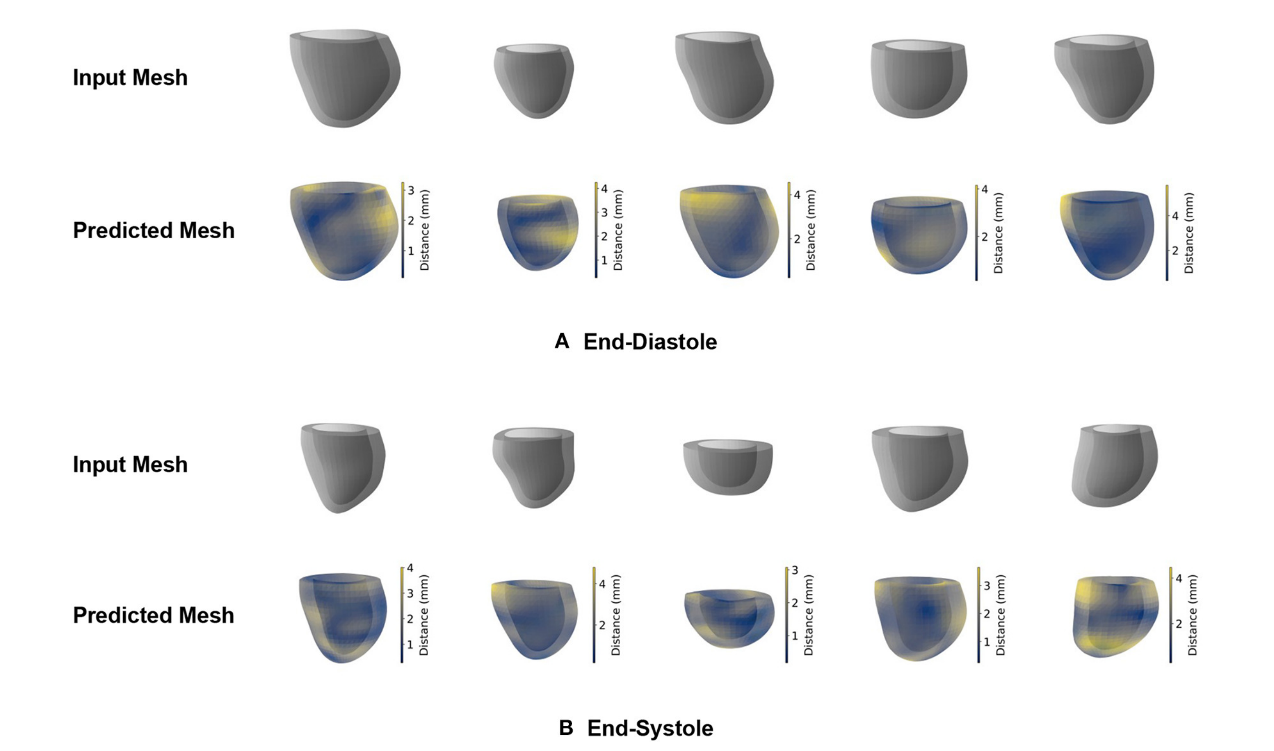


FIGURE 3
Qualitative reconstruction results of the mesh VAE for five sample cases. Results are presented separately for ED **(A)** and ES **(B)** phases. Predicted meshes are color-coded based on the vertex-wise distances to their corresponding input meshes.

TABLE 1 Reconstruction results of the mesh VAE and the 3D VAE on the test dataset.

Phase	Method	Hausdorff distance (mm)	Surface distance (mm)
ED	3D VAE	6.43 (±2.57)	1.26 (±0.48)
	Mesh VAE	4.73 (±1.46)	0.96 (±0.23)
ES	3D VAE	6.64 (±2.81)	1.53 (±0.66)
	Mesh VAE	4.43 (±1.23)	0.99 (±0.20)

Values represent mean (± standard deviation) in all cases.

VAE can process unseen 3D shape information. As a benchmark method for comparative validation, we choose a 3D VAE which is designed for 3D voxelgrid data and has previously been used to process 3D cardiac surface information (28, 65). We train and evaluate the 3D VAE on the same training and test datasets as our mesh VAE and report the results in Table 1. In order to apply the 3D VAE to our mesh dataset, we first voxelize each 3D triangular mesh and then place it in the center of a 128×128×128 voxelgrid with a voxel size of 1.5 × 1.5 × 1.5 mm. The resulting voxelgrids serve as the input and gold standard data for the 3D VAE. The architecture, loss function, and training procedure of the 3D VAE are chosen to be as close as possible to the mesh VAE's design in order to enable a fair comparison. The graph convolutions and mesh pooling layers are replaced by standard convolution and max pooling operations, respectively.

We find that the mesh VAE obtains mean surface distance values considerably below the pixel resolution of the underlying image acquisition (1.36 mm) for both ED and ES phases. It also achieves significantly lower distance scores than the 3D VAE for both HD and MSD metrics and for both ED and ES phases. For both evaluated methods and phases, the Hausdorff distance values are substantially larger than the MSD scores indicating that certain small localized regions exhibit larger differences between reconstructed and gold standard meshes than the global average.

3.3. Technical comparison

In addition to assessing the mesh VAE in terms of its reconstruction performance, we also evaluate its memory footprint in comparison to the 3D VAE. To this end, we calculate both the size of each data instance used as an input to the respective networks and the number of trainable network parameters in each approach (Table 2).

In terms of both metrics, the mesh VAE shows considerably better scores than the 3D VAE. It requires only about 25 times fewer trainable network parameters and processes

TABLE 2 Technical comparison of the mesh VAE and the 3D VAE.

Method	Data type	Data instance size	Network parameters
3D VAE	Voxelgrid	$\sim 2.1 \times 10^6$ ($128 \times 128 \times 128$)	$\sim 1.1 \times 10^6$
Mesh VAE	Mesh	$\sim 7.4 \times 10^3$ (2450×3)*	$\sim 4.4 \times 10^4$

*Vertex connectivity is the same for each mesh in the dataset.

approximately 285 times smaller input data while still representing the underlying cardiac anatomy with higher fidelity. This also allows us to train the mesh VAE on a standard CPU as opposed to the GPU required for the 3D VAE. While this makes a direct comparison of the run time difficult, we would expect it to lead to a considerably faster execution of the Mesh VAE.

3.4. Latent space analysis

After having shown the mesh VAE's ability to accurately model 3D cardiac shapes with high efficiency, we want to further investigate its latent space as a key architectural component to successfully represent inter-subject anatomy changes. As indicated by the two terms in the VAE's loss function, the objective of the latent space is similarly two-fold. On the one hand, it aims to provide a suitable low-dimensional encoding of high-dimensional input meshes that allows for accurate reconstruction. On the other hand, it is also tasked to represent important aspects of population-wide shape variability in a disentangled and interpretable way by approximating a multivariate Gaussian distribution. While the experiments in Section 3.2 show the adequacy of the mesh VAE's latent space for the reconstruction task, we want to focus on its role in modeling variability in this section.

To this end, we first pass all meshes of our dataset through the pretrained encoder of the mesh VAE to obtain the latent space representation of each case. We then use these representations to calculate the activity of each latent space component. Intuitively, the activity scores give an indication as to how much of the overall shape variability across the population is captured by a given latent space dimension. We follow these steps separately for the mesh VAEs trained on ED and ES data, respectively, and report the results in Figure 4. Hereby, the activity of each latent space dimension is presented as a percentage of the total activity, and latent space dimension are arranged in decreasing order of their respective activity percentages.

We observe that while the majority of the 16 latent space components capture more than 8% of the overall population variability each, a few components model close to 0%. Among the significantly contributing components, differences between the most and least active dimensions are relatively small at approximately 3%. Results are mostly consistent between the ED and ES phases with the ED phase showing one additional significantly contributing component and consequently slightly smaller activity scores for each of them.

After quantifying the variability in the low-dimensional latent space, we want to investigate the effect of changes in the individual latent space components on 3D shape variability. This allows us to identify if different latent space dimensions are responsible for modeling different aspects of the 3D anatomical variability. To this end, we first determine the mean latent space encoding across the population and then vary individual latent space components while keeping the other components fixed at the population mean value. We next pass these latent space representations through the decoder of the pre-trained mesh VAE in order to visualize the effect of the change in the particular latent space component on the 3D anatomy. Each individual latent space component is varied by 3 standard deviations from its mean value in both the positive and negative directions to analyze shape changes in both sides of the unimodal probability distribution. Based on our findings in Figure 4, we depict the resulting meshes corresponding to changes in the four most active latent space components, the least active one, and the one with the largest activity difference between ED and ES phases in Figure 5.

We observe that variations in each of the four most active latent space components result in gradual and easily identifiable changes in 3D anatomical shapes for both the ED and ES phases. These include changes in overall heart size, the pointedness of the apex, the basal plane tilt, and the longitudinal curvature and elongation of the ventricle. Variations along the least active latent space component do not cause any easily noticeable changes to the overall shape in either the ED or ES phase. We also find clear 3D shape changes when varying the last component with significant activity scores for the ED phase (component 11), while the same component for ES phase represents the first component with low activity scores and does not produce any easily visible 3D shape changes.

3.5. Generation of virtual cardiac mesh populations

As a first possible sample application of our mesh VAE, we next want to analyze whether it is able to generate new as well as realistic 3D cardiac meshes altogether. Such virtual population cohorts have a variety of use cases, such as data augmentation for disease classification or electrophysiological computer simulations as part of *in silico* trials. Hereby, the synthesized meshes should be as indistinguishable as possible from the real ones on both an individual and a population level. In order to evaluate the mesh VAE's performance in this task,

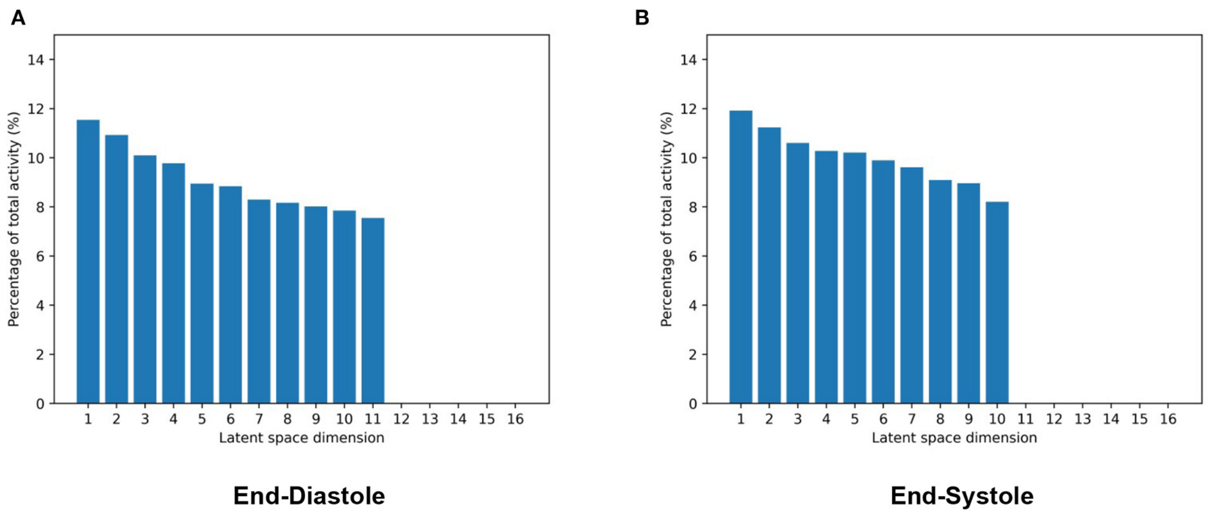


FIGURE 4
Activity values of each latent space dimension as a percentage of total activity in mesh VAEs trained separately on ED **(A)** and ES **(B)** data. Latent space dimensions are presented in decreasing order of their activity percentage values.

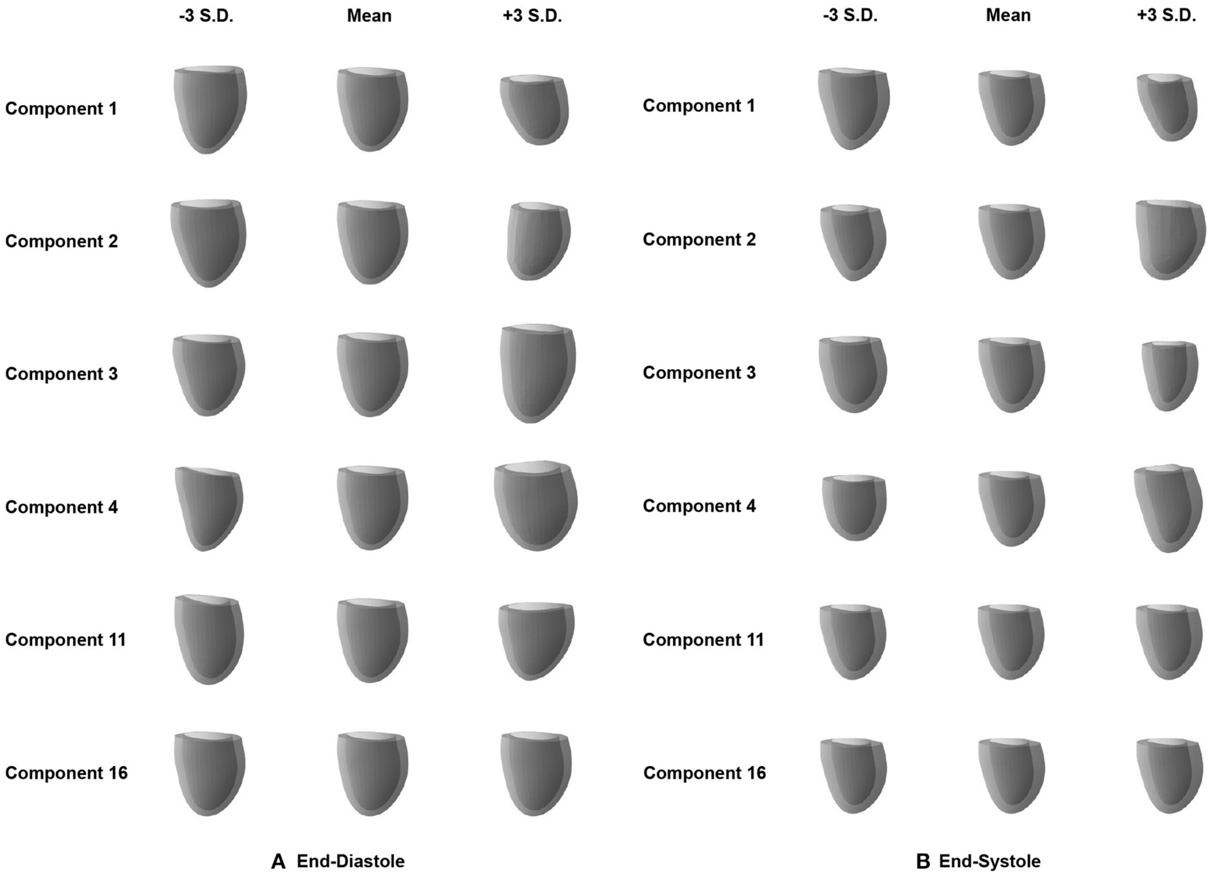


FIGURE 5
Effect of changes in the individual latent space components (rows) on the 3D mesh shapes reconstructed by the decoder of the mesh VAE for ED **(A)** and ES **(B)** phases. Results are shown for the four most active latent space components (1–4), the least active one (16), and the one with the largest activity difference between ED and ES phases (11).

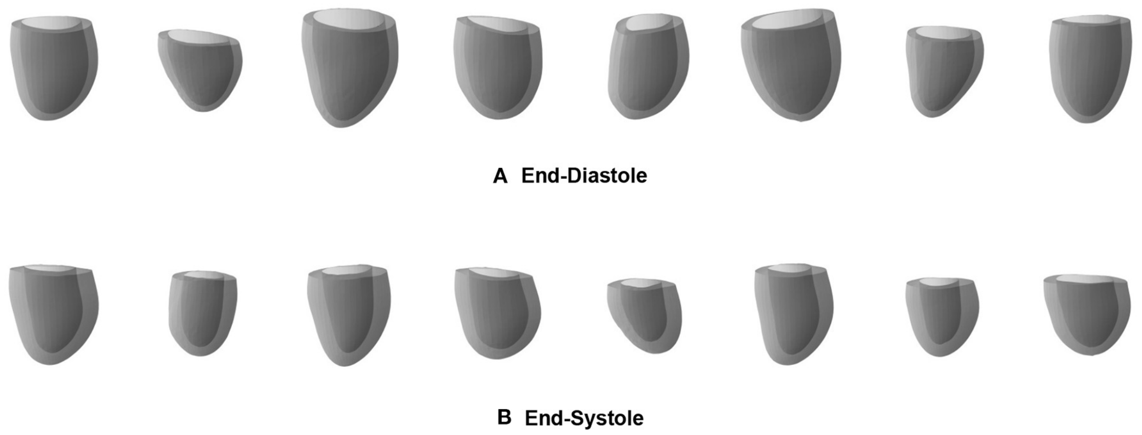


FIGURE 6
Sample meshes generated by separate pretrained ED **(A)** and ES **(B)** mesh VAEs.

we draw random samples from the multivariate latent space distribution and pass each of them through the trained decoder of the VAE to obtain the corresponding virtual output meshes. We perform this procedure separately for the mesh VAE decoder trained on ED and ES and depict 8 sample results for each phase in Figure 6.

We observe that the generated meshes exhibit a degree of shape variation close to that of the true population for both the ED and ES phases, while still maintaining realistic anatomical shapes on an individual level. Typical shape changes regarding for example the overall heart size, mid-cavity diameter, or basal plane tilt, are successfully represented in the virtual population. In addition, the heart meshes at ES phase generally show a thicker myocardium and a smaller overall volume than the ED population, which is again reflective of the true population. Since we use separate networks for the ED and ES phases, no per-case correspondence between individual generated meshes of the two phases is enforced.

In order to quantify the realism of the generated heart population, we first randomly sample 1,000 latent space vectors and pass them through the mesh VAE's decoder to obtain a large virtual population of 3D cardiac meshes. We then calculate the widely used clinical metrics LV volume and LV mass for each mesh in the generated population and report the resulting population mean and standard deviation values in Table 3. For a comparative analysis, we also provide the same scores for the meshes in the unseen test dataset, which we assume as our gold standard in this work. We apply this procedure for both ED and ES phases using the respective pre-trained networks and report the results separately for each phase in Table 3.

We find similar mean and standard deviation values between the synthesized mesh population and the gold standard mesh population for both evaluation metrics and cardiac phases. The average difference in population means across all scores is 2.5%, with slightly larger deviations for the ED phase compared to the ES phase.

TABLE 3 Clinical metrics of generated and gold standard mesh populations.

Phase	Clinical metric	Gold standard	Mesh VAE
ED	LV volume (ml)	156 (±42)	152 (±40)
	LV mass (g)	123 (±28)	120 (±26)
ES	LV volume (ml)	81 (±32)	79 (±28)
	LV mass (g)	128 (±32)	125 (±31)

Values represent mean (± standard deviation) in all cases.

3.6. MACE prediction

In addition to its utility for generating virtual mesh populations, we also want to investigate whether the mesh VAE can capture pathology-specific information that is useful for cardiac disease detection and diagnosis. In this work, we focus on MACE as a possible sample outcome and want to first study whether there exist differences in the mesh VAE's latent space representations of post-MI subjects with and without an associated MACE. We therefore pass all MACE cases through the encoder of the pre-trained mesh VAE, average the resulting latent space representations, and then feed the resulting mean vector through the pre-trained decoder to obtain the average mesh representation of all MACE cases in the population. We repeat the same process for all cases without MACE and depict

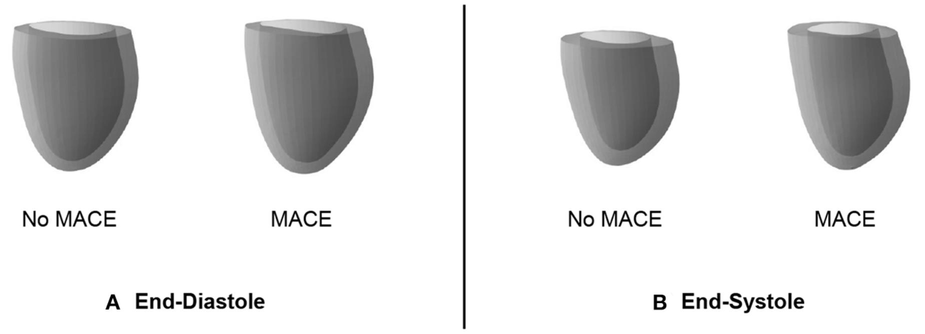


FIGURE 7
Cardiac anatomy meshes reconstructed by decoders of pre-trained mesh VAEs from averaged latent space representations of patients with and without subsequent MACE. Results are shown for ED **(A)** and ES **(B)** data.

the obtained averaged meshes in Figure 7A for ED data and in Figure 7B for ES data.

We observe that the differences between averaged MACE and no MACE anatomies are small for ED data, but easily visible for ES data. This indicates that ED shape alone is considerably less predictive of incident MACE outcomes than ES shape, which is in line with clinical guidelines and previous research work (2).

Since the observed differences in 3D ES shapes between averaged MACE and no MACE cases were obtained using the same pre-trained decoder for mesh reconstruction, they should be caused by corresponding differences in the mesh VAE's latent space representations. As a next step, we want to investigate whether these low dimensional representations of 3D cardiac anatomies are not only suitable to represent subpopulation-specific average shapes, but also to predict future MACE outcomes for individual patients. To this end, we employ a logistic regression classifier to predict the binary outcome MACE vs. no MACE based on per-patient latent space encodings obtained from the mesh VAE encoder. As a clinical benchmark, we select the ES volume as a widely used metric in clinical practice and use it as the input to the another logistic regression model with the same settings. We choose the AUROC as a comparative metric for binary prediction performance due to the high class imbalance between MACE and no MACE cases in the dataset. In order to maintain the same class imbalance in the respective train and test sets and to improve the robustness of our analysis, we conduct stratified 10-fold cross validation experiments with both classifiers and report the averaged results in Table 4.

We find that the mesh VAE's latent space representations achieve an about 7% higher AUROC score than the ES volume values for the task of binary MACE prediction. We note, however, that the primary objective of this experiment is only to showcase the utility of the mesh VAE for a possible clinical application and not to present a method optimized for MACE prediction specifically, which we leave for future work.

TABLE 4 Results of binary MACE classification.

Metric	ES volume	Mesh VAE
AUROC	0.627 (±0.042)	0.671 (±0.038)

Values represent mean (± standard deviation).

4. Discussion

In this work, we have presented a novel geometric deep learning approach specifically designed for cardiac mesh processing as part of a multi-step cardiac shape analysis pipeline and demonstrated its versatility in multiple applications.

4.1. Mesh reconstruction

In our experiments, we find that the mesh VAE is able to accurately encode and decode complex 3D cardiac anatomy shapes with high degrees of realism, by attaining average surface distances between predicted and ground truth anatomies in the test dataset smaller than the underlying image resolution. This demonstrates that it is not only capable of capturing anatomical surface information in individual cases, but also correctly represents population-wide variability in cardiac shapes. These results are consistent across both the ED and ES mesh datasets and show that the mesh VAE is suitable for processing cardiac anatomy data at various phases of the cardiac cycle, albeit with separate networks for each phase. As these represent the two extreme ends of the cardiac cycle, we hypothesize that an application to intermediate frames is equally feasible. We observe this good alignment between predicted and input meshes not only on a global but also on a local surface level. This indicates that inter-subject shape variation is also successfully captured on a smaller, more localized scale which promises to aid in the discovery of new image-based biomarkers of cardiac abnormalities that go beyond the purely volume-based metrics widely used in current clinical practice. The largest localized prediction errors occur at regions with remaining slice misalignment artifacts and near the base of the left ventricle. We believe this to be at least partially a consequence of the limitations of the 3D surface reconstruction step (Section 4.6) rather than an issue of the mesh VAE itself. In fact, the observed slight smoothing effect of the mesh VAE typically occurs in localized regions that are still affected by reconstruction artifacts. This hints at potentially favorable effects of this small smoothing behavior, since its implicit slight misalignment corrections often result in more anatomically plausible 3D meshes without significant loss of true localized shape details. Multiple ways that would likely reduce the smoothing effect can be easily integrated into our mesh VAE framework, such as increasing the weight value of the reconstruction loss term during training or using a reconstruction loss that puts a disproportionately higher penalty on larger vertex-wise reconstruction errors. However, such measures could also lead to other unwanted effects, such as a reduced quality of the latent space distribution or an increased number of unnatural local deformations in the output meshes that mimic errors in the original 3D surface reconstruction process. In general, we note that the anatomically accurate reconstruction results have been achieved on a challenging dataset of pathological subjects acquired from multiple studies, in contrast to more homogeneous datasets of healthy subjects, such as the UK Biobank study (66). This further demonstrates the robustness of our mesh VAE.

4.2. Latent space quality

The ability of the mesh VAE to successfully model 3D cardiac shape variability across the population is further corroborated by the analysis of its latent space. We find that variations in latent space components are associated with realistic changes in reconstructed 3D shapes and that individual components are responsible for encoding different aspects of the population-wide shape variability. Examples of such easily visible effects include changes to the overall heart size, mid-ventricular diameter, and basal plane tilt, which are all similar to previous findings in cardiac shape modeling (19). This high degree of disentanglement enables an improved understanding of the key components of cardiac shape variations and higher levels of interpretability for the multiple clinical applications of the mesh VAE. When comparing the contribution of individual latent space components to overall shape variability, we find that some components are responsible for a larger percentage of the total variation than others. This is similar to the results of other widely used shape analysis techniques, such as PCA, where different principal components account for different proportions of the overall variance. Contrary to PCA, however, we find that the differences in activity percentage scores decrease only very slowly for the majority of components before a sharp drop after the 11th and 10th most contributing components for ED and ES data, respectively. All following components play almost no part in explaining the population variance. This is in contrast to PCA where the percentage of explained variance by each component typically decreases sharply for the first few most contributing components with very little change between less contributing components (2, 19–21). We hypothesize that this is due to the non-linearities in the mesh VAE's architecture which enable the modeling of richer and more condensed relationships between high-dimensional input data and low-dimensional latent space representations compared to purely linear approaches, such as PCA. This results in a different way of encoding shape variability with more equal activity scores for each contributing component. When varying along the latent space components with close to zero activity and observing the effect on the reconstructed 3D anatomies, we indeed find very little change, especially on a global level (Figure 4). However, similar to PCA, such components might still encode meaningful information about smaller, more localized shape variations. This induces a certain amount of risk when removing seemingly non-contributing components post-training, as otherwise important variability might be inadvertently removed. When experimenting with different latent space sizes in our mesh VAE, we find that differences in reconstruction accuracy are minimal between larger and smaller latent space dimensionalities. We also observe that 5–50% of the latent space dimensions have minor contributions to the overall variance, regardless of the choice of latent space size. Hence, we reason that the more condensed encoding of the same amount of shape information is a property of the overall mesh VAE architecture itself rather than solely a consequence of the latent space size. This also shows that changing the latent space size has little effect on removing potentially redundant latent space dimensions as the network adjusts its encoding accordingly. Furthermore, we also notice that training the same network with the same parameter settings can result in varying numbers of significantly contributing latent space dimensions. This indicates that the various sources of randomness involved in training deep learning networks (e.g., trainable parameter initialization, order of cases seen during training) affect the way the mesh VAE encodes shape information in the latent space. As such, we conclude that our choice of 16 as the latent space size reflects a reasonable trade-off between having too many modes and a representation that is too condensed, both of which would negatively affect interpretability. This also means that our choice is not a fixed optimal value but rather that it should be chosen with the particular dataset and downstream application in mind. When comparing the results for ED and ES meshes, we find very minor differences with only one additional contributing latent space component for ED and similar levels of disentanglement. We therefore conclude that the mesh VAE can successfully capture 3D shape variability at different phases of the cardiac cycle.

4.3. Generation of virtual cardiac mesh populations

In addition to accurately capturing 3D anatomical patterns of existing subjects, we also find that the mesh VAE is capable of generating realistic virtual populations of 3D heart meshes. Hereby, we observe a high degree of realism in both the individually generated meshes and in the amount of variability present in the overall virtual population, which closely mimics the true underlying population. We have shown this both qualitatively and quantitatively by calculating commonly used clinical metrics on a population level. We also note that we find not only a good alignment between virtual generated population and gold standard population in terms of their mean values but also in terms of their standard deviation scores, indicating that the variability across the population is well-captured in the virtual population. The mesh VAE achieves these positive results for both ED and ES phases which shows its versatility to different datasets and suggests the possibility of a feasible extension to other phases of the cardiac cycle. For example, the generated ES anatomies typically exhibit a thicker myocardium and an overall smaller size than their ED counterparts which is reflective of real cardiac morphology. The high degree of realism in the generated meshes is also visible on both a global and local level. This is particularly important for virtual population cohorts used in computer simulations of cardiac electrophysiology which model conduction patterns granularly for each face in a mesh. We also note that all generated virtual meshes retain the same vertex connectivity as a result of the chosen network architecture which is another beneficial property for many follow-up tasks. In our experiments, we find that sampling from a latent space distribution based on encoder predictions of the training dataset leads to better mesh generation results than sampling from a multivariate standard Gaussian distribution. We attribute this to the trade-off between the reconstruction and latent space terms in the loss function which cause the latent space to only approximate the idealized prior distribution in order to retain a high reconstruction accuracy.

4.4. MACE prediction

As a compact low-dimensional representation of high-dimensional cardiac anatomies, the latent space should also be able to capture subpopulation-specific differences based solely on information in the input shapes. In our experiments, we observe such differences for the MACE vs. no MACE subpopulations in the ES data and to a lesser extent in the ED data. In both cases, we hypothesize this to be a reflection of the information contained in the 3D shapes instead of a potential inadequacy of the latent space itself. This is corroborated by findings in previous work and clinical practice where metrics based on ES heart shapes are considerably more predictive than corresponding ED-based scores (2). We then show how these latent space differences for ES data can be successfully used to predict MACE outcomes and outperform a common clinical benchmark. We presume that the classifier's access to a condensed representation of the full 3D shape information as opposed to a single value to coarsely approximate said 3D shape is the key reason for this result. This allows the classifier to take into account finer and more localized patterns without getting overwhelmed by too much information, as the complex 3D shape has already been sensibly encoded by the mesh VAE's encoder in a non-linear way. We note, however, that the objective of this experiment was not to achieve the best possible classification performance but rather to generally showcase the utility of the mesh VAE's latent space representation for this task. Hence, we achieve good results without any specific classification loss term during network training but relying only on general encodings of shape variability. Such a multi-task learning approach would likely have improved the separability of the latent space to further differentiate between MACE and no MACE cases, while maintaining a high degree of interpretability. This high degree of extensibility is a key advantage to the presented mesh VAE approach which we aim to explore further in future work.

4.5. Network architecture and training

In general, the positive results obtained by the mesh VAE in the previously discussed experiments demonstrate that both its architectural design, loss function, and training procedure were adequately chosen for effective cardiac anatomy modeling with 3D surface mesh data. The graph convolutional layers combined with the mesh downsampling and upsampling operations enable multi-scale feature learning in a hierarchical setup that successfully considers both global and local aspects of cardiac shape variability, which is important to its many possible clinical and research applications. While this is in principle similar to conventional convolution and pooling operations on voxelgrid data, we find that these achieve higher reconstruction errors than a geometric deep learning architecture that is specifically designed to process triangular surface mesh data. In addition, the mesh VAE achieves this outperformance in terms of accuracy while using only about 4% of the number of trainable parameters. This significantly reduces the training time and memory requirements of the algorithm and allowed us to train and evaluate our deep learning models on a standard CPU as opposed to a GPU which is typically required for the 3D voxelgrid VAE. At the same time, the mesh VAE allows for anatomical shapes to be represented as triangular surface meshes which reduces the required data storage costs considerably compared to voxelgrids despite not losing any anatomical information. Furthermore, meshes allow for a continuous encoding of vertex coordinates as opposed to the discretization needed to store similar data in grid-based formats. This sets an upper bound on the possible data resolution due to limited available memory which in turns affects the quality of the anatomical representation.

The choice of VAE framework also allows for straightforward ways to include other metadata, such as patient characteristics or acquisition conditions, into the network as conditional inputs in addition to the anatomical shape information (36, 67). These can, for example, be included as per-vertex features in combination with the coordinate values, or concatenated to the latent space vector or intermediate layers of the network (36, 67, 68). Such an extension enables subpopulation-specific cardiac anatomy modeling while still using a single network and dataset. This is in contrast to PCA, which would need to be applied separately for each subpopulation and would hence only be able to use smaller partitions of the original dataset without making use of synergies across the subpopulations. In order to then achieve the same performance, more data would likely be necessary, whose acquisition is particularly costly in case of medical images.

Regarding the mesh VAE's training procedure, we find that setting the weighting parameter β in the loss function to a suitable value for the given dataset is important to find the right balance between reconstruction and latent space quality and enable effective cardiac anatomy modeling. In our experiments, prioritizing reconstruction quality and using a monotonic annealing schedule resulted in the best overall performance which is in line with previous applications of the β-VAE framework to 3D shape modeling (36, 55). In addition, the mesh VAE is also highly flexible and can work with a variety of different input modalities and reconstruction pipelines, as long as vertex correspondence between meshes in the dataset is ensured. Its architecture also allows for the easy integration of other network components (e.g., as a separate encoder or decoder branch) and multiple different objectives (e.g., cardiac disease classification) in a multi-task learning setting without loss of interpretability. This is akin to grid-based VAEs and therefore creates the possibility of further improvements in similar use cases (27–34).

4.6. Limitations

The proposed shape modeling approach also comes with some limitations. All meshes in the input dataset need to exhibit vertex-to-vertex correspondence between each other. As this needs to be established in the preprocessing steps, it limits the method's flexibility and increases its complexity. While this is a common requirement for most shape modeling approaches, including PCA, it is in contrast to voxelgrid-based (28) and point cloud-based shape modeling approaches (36), where such point correspondence is not strictly needed. As a deep learning approach, the mesh VAE requires 2–3 h of CPU training time in our current setup before the population-wide shape variability is accurately captured and follow-up tasks can be performed. This is contrasted with traditional machine learning approaches for the same purpose, such as PCA, which are typically faster in determining their respective data transformation parameters. However, as mentioned in Section 4.5, the mesh VAE still compares favorably in terms of memory footprint and training time to other deep learning approaches based on voxelgrid or point cloud processing.

Furthermore, we have only investigated shape variability in the left ventricle and at the ED and ES phases in this work. In addition, we have trained separate models for ED and ES data which likely results in limited per-subject correspondence between ED and ES meshes in the generated populations. However, we believe that the presented approach can be extended to other cardiac chambers and to other phases of the cardiac cycle, including a combined multi-temporal modeling setup, which we plan to explore in future work. This could be achieved by introducing conditional inputs into various parts of the current architecture that control the cardiac phases to be modeled. Alternatively, separate phase-specific encoder-decoder blocks could be used with a shared latent space to capture multiple cardiac phases at once. This would then likely enable the shape analysis and virtual population generation of paired ED and ES heart meshes. We also note that errors introduced in the preprocessing steps (e.g., MRI segmentation, 3D surface reconstruction) of our shape modeling pipeline affect the results of both the mesh VAE and its follow-up tasks presented in this work. In particular, the reconstruction step does not take into account the information of long-axis slices of the cine MRI acquisition, leading to possible inaccuracies in the basal and apical areas of the 3D heart mesh. While the 3D surface reconstruction step explicitly tries to correct for slice misalignment due to respiratory motion during image acquisition, there are likely still some smaller errors present in the resulting meshes. However, we find that the mesh VAE can successfully process such cases and is often even able to remove unnatural curvatures of the anatomical surface in its reconstructed outputs.

We have also only evaluated the mesh VAE on post-MI subjects and not on a purely healthy cohort. Specifically regarding the MACE classification experiment, we did not consider any patient metadata that would likely help to further improve the results (e.g., sex, age). However, we note that the objective of this work was not to achieve the best possible performance in a single one of the presented tasks but rather to showcase the versatility and applicability of the mesh VAE as a novel approach to 3D cardiac anatomy modeling.

5. Conclusion

To conclude, we have presented the mesh VAE as a novel approach to 3D cardiac anatomy modeling that can be directly applied to surface meshes of the heart in an efficient manner. We have demonstrated its ability to accurately capture complex 3D cardiac shapes at both ends of the cardiac cycle while using low-dimensional and easily interpretable latent space representations. The mesh VAE also compares favorably to voxelgrid-based deep learning approaches in terms of both accuracy and memory requirements. Furthermore, we have shown its utility for two exemplary applications, namely the generation of realistic virtual population cohorts of 3D cardiac anatomies and the prediction of MACE outcomes in post-MI patients.

Data availability statement

The datasets presented in this article are not publicly available. Generated virtual data can be made available upon reasonable request. Requests to access the datasets should be directed to VG, vicente.grau@eng.ox.ac.uk.

Ethics statement

The studies involving human participants were reviewed and approved by TATORT-NSTEMI trial: Ethical Committee at the University of Leipzig and all Local Ethical Committees of the participating sites; AIDA-STEMI trial: Ethics Committee of the National Regulatory Authorities and the participating centers. The patients/participants provided their written informed consent to participate in this study.

Author contributions

VG and MB conceptualized and designed the study. AS, TL, TS, SB, HT, and IE acquired the input dataset. JC, AB, EZ, PL, AB-O, and MB implemented and applied the preprocessing steps. MB developed the deep learning methods, conducted the experiments, and created a first draft of the manuscript. VG supervised the work. All authors revised and approved the final version of the manuscript.

Funding

The work of MB was supported by the Stiftung der Deutschen Wirtschaft (Foundation of German Business). The work of JC was supported by the EU's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie (g.a. 764738) and the EPSRC Impact Acceleration Account (D4D00010 DF48.01), funded by UK Research and Innovation. AB is a Royal Society University Research Fellow and supported by the Royal Society (Grant No. URF\R1\221314). The work of AB and VG was supported by the British Heart Foundation (BHF) Project under Grant HSR01230. The work of VG was supported by the CompBioMed 2 Centre of Excellence in Computational Biomedicine (European Commission Horizon 2020 research and innovation programme, grant agreement No. 823712). AB-O holds a BHF Intermediate Basic Science Research Fellowship (FS/17/22/32644). The work was also supported by the German Center for Cardiovascular Research, the British Heart Foundation (PG/16/75/32383), and the Wellcome Trust (209450/Z/17).

Acknowledgments

The authors acknowledge the use of the facilities and services of the Institute of Biomedical Engineering (IBME), Department of Engineering Science, University of Oxford.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Armstrong AC, Gidding S, Gjesdal O, Wu C, Bluemke DA, Lima JA. LV mass assessed by echocardiography and CMR, cardiovascular outcomes, and medical practice. *JACC Cardiovasc Imaging.* (2012) 5:837–48. doi: 10.1016/j.jcmg.2012.06.003

2. Corral Acero J, Schuster A, Zacur E, Lange T, Stiermaier T, Backhaus SJ, et al. Understanding and improving risk assessment after myocardial infarction using automated left ventricular shape analysis. *JACC Cardiovasc Imaging.* (2022) 15:1563–74. doi: 10.1016/j.jcmg.2021.11.027

3. Ponikowski P, Voors A, Anker S, Bueno H, Cleland J, Coats A, et al. 2016 ESC Guidelines for the diagnosis treatment of acute chronic heart failure: the Task Force for the diagnosis treatment of acute chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. *Eur J Heart Fail*. (2016) 18:891–975. doi: 10.1002/ejhf.592

4. Bai W, Suzuki H, Huang J, Francis C, Wang S, Tarroni G, et al. A population-based phenome-wide association study of cardiac and aortic structure and function. *Nat Med*. (2020) 26:1654–62. doi: 10.1038/s41591-020-1009-y

5. Gilbert K, Mauger C, Young AA, Suinesiaputra A. Artificial intelligence in cardiac imaging with statistical atlases of cardiac anatomy. *Front Cardiovasc Med*. (2020) 7:102. doi: 10.3389/fcvm.2020.00102

6. Stokes MB, Roberts-Thomson R. The role of cardiac imaging in clinical practice. *Australian Prescriber*. (2017) 40:151–5. doi: 10.18773/austprescr.2017.045

7. Kawel-Boehm N, Hetzel SJ, Ambale-Venkatesh B, Captur G, Francois CJ, Jerosch-Herold M, et al. Reference ranges ("normal values") for cardiovascular magnetic resonance (CMR) in adults and children: 2020 update. *J Cardiovasc Magnet Reson*. (2020) 22:1–63. doi: 10.1186/s12968-020-00683-3

8. Kawel-Boehm N, Maceira A, Valsangiacomo-Buechel ER, Vogel-Claussen J, Turkbey EB, Williams R, et al. Normal values for cardiovascular magnetic resonance in adults and children. *J Cardiovasc Magnet Reson*. (2015) 17:1–33. doi: 10.1186/s12968-015-0111-7

9. Petersen SE, Aung N, Sanghvi MM, Zemrak F, Fung K, et al. Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort. *J Cardiovasc Magnet Reson*. (2017) 19:18. doi: 10.1186/s12968-017-0327-9

10. Prakken NH, Velthuis BK, Teske AJ, Mosterd A, Mali WP, Cramer MJ. Cardiac MRI reference values for athletes and nonathletes corrected for body surface area, training hours/week and sex. *Eur J Prev Cardiol*. (2010) 17:198–203. doi: 10.1097/HJR.0b013e3283347fdb

11. Di Folco M, Moceri P, Clarysse P, Duchateau N. Characterizing interactions between cardiac shape and deformation by non-linear manifold learning. *Med Image Anal*. (2022) 75:102278. doi: 10.1016/j.media.2021.102278

12. Suinesiaputra A, Ablin P, Alba X, Alessandrini M, Allen J, Bai W, et al. Statistical shape modeling of the left ventricle: myocardial infarct classification challenge. *IEEE J Biomed Health Inform*. (2017) 22:503–15. doi: 10.1109/JBHI.2017.2652449

13. Beetz M, Banerjee A, Grau V. Biventricular surface reconstruction from cine MRI contours using point completion networks. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. Nice. (2021). p. 105–9. doi: 10.1109/ISBI48211.2021.9434040

14. Banerjee A, Camps J, Zacur E, Andrews CM, Rudy Y, Choudhury RP, et al. A completely automated pipeline for 3D reconstruction of human heart from 2D cine magnetic resonance slices. *Philos Trans R Soc A Math Phys Eng Sci*. (2021) 379:20200257. doi: 10.1098/rsta.2020.0257

15. Banerjee A, Zacur E, Choudhury RP, Grau V. Optimised misalignment correction from cine MR slices using statistical shape model. In: *Annual Conference on Medical Image Understanding and Analysis*. Oxford: Springer (2021). p. 201–9. doi: 10.1007/978-3-030-80432-9_16

16. Banerjee A, Zacur E, Choudhury RP, Grau V. Automated 3D whole-heart mesh reconstruction from 2D cine MR slices using statistical shape model. In: *44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. Glasgow. (2022) p. 1702–6. doi: 10.1109/EMBC48229.2022.9871327

17. Lamata P, Sinclair M, Kerfoot E, Lee A, Crozier A, Blazevic B, et al. An automatic service for the personalization of ventricular cardiac meshes. *J R Soc Interface*. (2014) 11:20131023. doi: 10.1098/rsif.2013.1023

18. Villard B, Grau V, Zacur E. Surface mesh reconstruction from cardiac MRI contours. *J Imaging*. (2018) 4:16. doi: 10.3390/jimaging4010016

19. Bai W, Shi W, de Marvao A, Dawes TJ, O'Regan DP, Cook SA, et al. 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*. (2015) 26:133–45. doi: 10.1016/j.media.2015.08.009

20. Gilbert K, Bai W, Mauger C, Medrano-Gracia P, Suinesiaputra A, Lee AM, et al. Independent left ventricular morphometric atlases show consistent relationships with cardiovascular risk factors: a UK Biobank Study. *Sci Rep*. (2019) 9:1–9. doi: 10.1038/s41598-018-37916-6

21. Mauger C, Gilbert K, Lee AM, Sanghvi MM, Aung N, Fung K, et al. Right ventricular shape and function: cardiovascular magnetic resonance reference morphology and biventricular risk factor morphometrics in UK Biobank. *J Cardiovasc Magnet Reson*. (2019) 21:1–13. doi: 10.1186/s12968-019-0551-6

22. Corral Acero J, Margara F, Marciniak M, Rodero C, Loncaric F, Feng Y, et al. The "Digital Twin" to enable the vision of precision cardiology. *Eur Heart J*. (2020) 41:4556–64. doi: 10.1093/eurheartj/ehaa159

23. Rodero C, Strocchi M, Marciniak M, Longobardi S, Whitaker J, O'Neill MD, et al. Linking statistical shape models and simulated function in the healthy adult human heart. *PLoS Comput Biol*. (2021) 17:e1008851. doi: 10.1371/journal.pcbi.1008851

24. Romero P, Lozano M, Martinez-Gil F, Serra D, Sebastián R, Lamata P, et al. Clinically-driven virtual patient cohorts generation: an application to aorta. *Front Physiol*. (2021) 12:713118. doi: 10.3389/fphys.2021.713118

25. Zhang X, Ambale-Venkatesh B, Bluemke DA, Cowan BR, Finn JP, Kadish AH, et al. Information maximizing component analysis of left ventricular remodeling due to myocardial infarction. *J Transl Med*. (2015) 13:1–9. doi: 10.1186/s12967-015-0709-4

26. Kingma DP, Welling M. Auto-encoding variational Bayes. *arXiv preprint arXiv:13126114*. (2013). doi: 10.48550/arXiv.1312.6114

27. Clough JR, Oksuz I, Puyol-Antón E, Ruijsink B, King AP, Schnabel JA. Global and local interpretability for cardiac MRI classification. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Shenzhen: Springer (2019). p. 656–64. doi: 10.1007/978-3-030-32251-9_72

28. Biffi C, Cerrolaza JJ, Tarroni G, Bai W, de Marvao A, Oktay O, et al. Explainable anatomical shape analysis through deep hierarchical generative models. *IEEE Trans Med Imaging*. (2020) 39:2088–99. doi: 10.1109/TMI.2020.2964499

29. Chen C, Biffi C, Tarroni G, Petersen S, Bai W, Rueckert D. Learning shape priors for robust cardiac MR segmentation from multi-view images. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Shenzhen: Springer (2019). p. 523–31. doi: 10.1007/978-3-030-32245-8_58

30. Duan J, Bello G, Schlemper J, Bai W, Dawes TJ, Biffi C, et al. Automatic 3D bi-ventricular segmentation of cardiac images by a shape-refined multi-task deep learning approach. *IEEE Trans Med Imaging*. (2019) 38:2151–64. doi: 10.1109/TMI.2019.2894322

31. Oktay O, Ferrante E, Kamnitsas K, Heinrich M, Bai W, Caballero J, et al. Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. *IEEE Trans Med Imaging*. (2017) 37:384–95. doi: 10.1109/TMI.2017.2743464

32. Zotti C, Luo Z, Lalande A, Jodoin PM. Convolutional neural network with shape prior applied to cardiac MRI segmentation. *IEEE J Biomed Health Informatics*. (2018) 23:1119–28. doi: 10.1109/JBHI.2018.2865450

33. Chartsias A, Joyce T, Papanastasiou G, Semple S, Williams M, Newby DE, et al. Disentangled representation learning in cardiac image analysis. *Med Image Anal*. (2019) 58:101535. doi: 10.1016/j.media.2019.101535

34. Bello GA, Dawes TJ, Duan J, Biffi C, De Marvao A, Howard LS, et al. Deep-learning cardiac motion analysis for human survival prediction. *Nat Mach Intell*. (2019) 1:95–4. doi: 10.1038/s42256-019-0019-2

35. Bronstein MM, Bruna J, LeCun Y, Szlam A, Vandergheynst P. Geometric deep learning: going beyond Euclidean data. *IEEE Signal Process Mag*. (2017) 34:18–42. doi: 10.1109/MSP.2017.2693418

36. Beetz M, Banerjee A, Grau V. Generating subpopulation-specific biventricular anatomy models using conditional point cloud variational autoencoders. In: *International Workshop on Statistical Atlases and Computational Models of the Heart*. Strasbourg: Springer (2021). p. 75–83. doi: 10.1007/978-3-030-93722-5_9

37. Chang Y, Jung C. Automatic cardiac MRI segmentation and permutation-invariant pathology classification using deep neural networks and point clouds. *Neurocomputing*. (2020) 418:270–9. doi: 10.1016/j.neucom.2020.08.030

38. Beetz M, Ossenberg-Engels J, Banerjee A, Grau V. Predicting 3D cardiac deformations with point cloud autoencoders. In: *International Workshop on Statistical Atlases and Computational Models of the Heart*. Strasbourg: Springer (2021). p. 219–28. doi: 10.1007/978-3-030-93722-5_24

39. Xiong Z, Stiles MK, Yao Y, Shi R, Nalar A, Hawson J, et al. Automatic 3D surface reconstruction of the left atrium from clinically mapped point clouds using convolutional neural networks. *Front Physiol*. (2022) 13:880260. doi: 10.3389/fphys.2022.880260

40. Ye M, Huang Q, Yang D, Wu P, Yi J, Axel L, et al. PC-U Net: learning to jointly reconstruct and segment the cardiac walls in 3D from CT data. *arXiv preprint arXiv:200808194*. (2020). doi: 10.1007/978-3-030-68107-4_12

41. Beetz M, Banerjee A, Grau V. Multi-domain variational autoencoders for combined modeling of MRI-based biventricular anatomy and ECG-based cardiac electrophysiology. *Front Physiol*. (2022) 13:886723. doi: 10.3389/fphys.2022.886723

42. Beetz M, Banerjee A, Sang Y, Grau V. Combined generation of electrocardiogram and cardiac anatomy models using multi-modal variational autoencoders. In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. Kolkata. (2022). p. 1–4. doi: 10.1109/ISBI52829.2022.9761590

43. Dalton D, Lazarus A, Rabbani A, Gao H, Husmeier D. Graph neural network emulation of cardiac mechanics. In: *3rd International Conference on Statistics: Theory and Applications (ICSTA'21)*. Prague. (2021). p. 1–8. doi: 10.11159/icsta21.127

44. Kong F, Wilson N, Shadden S. A deep-learning approach for direct whole-heart mesh reconstruction. *Med Image Anal*. (2021) 74:102222. doi: 10.1016/j.media.2021.102222

45. Meister F, Passerini T, Audigier C, Lluch É, Mihalef V, Ashikaga H, et al. Graph convolutional regression of cardiac depolarization from sparse endocardial maps. In: *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer (2020). p. 23–34. doi: 10.1007/978-3-030-68107-4_3

46. Suk J, Haan Pd, Lippe P, Brune C, Wolterink JM. Mesh convolutional neural networks for wall shear stress estimation in 3D artery models. In: *International Workshop on Statistical Atlases and Computational Models of the Heart*. Strasbourg: Springer (2021). p. 93–102. doi: 10.1007/978-3-030-93722-5_11

47. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Munich: Springer (2015). p. 234–41. doi: 10.1007/978-3-319-24574-4_28

48. Eitel I, Wöhrle J, Suenkel H, Meissner J, Kerber S, Lauer B, et al. Intracoronary compared with intravenous bolus abciximab application during primary percutaneous coronary intervention in ST-segment elevation myocardial infarction: cardiac magnetic resonance substudy of the AIDA STEMI trial. *J Am Coll Cardiol*. (2013) 61:1447–54. doi: 10.1016/j.jacc.2013.01.048

49. Eitel I, Stiermaier T, Lange T, Rommel KP, Koschalka A, Kowallick JT, et al. Cardiac magnetic resonance myocardial feature tracking for optimized prediction of cardiovascular events following myocardial infarction. *JACC Cardiovasc Imaging*. (2018) 11:1433–44. doi: 10.1016/j.jcmg.2017.11.034

50. Thiele H, de Waha S, Zeymer U, Desch S, Scheller B, Lauer B, et al. Effect of aspiration thrombectomy on microvascular obstruction in NSTEMI patients: the TATORT-NSTEMI trial. *J Am Coll Cardiol*. (2014) 64:1117–24. doi: 10.1016/j.jacc.2014.05.064

51. Corral Acero J, Xu H, Zacur E, Schneider JE, Lamata P, Bueno-Orovio A, et al. Left ventricle quantification with cardiac MRI: deep learning meets statistical models of deformation. In: *International Workshop on Statistical Atlases and Computational Models of the Heart*. Shenzhen: Springer (2019). p. 384–94. doi: 10.1007/978-3-030-39074-7_40

52. Corral Acero J, Zacur E, Xu H, Ariga R, Bueno-Orovio A, Lamata P, et al. SMOD-data augmentation based on statistical models of deformation to enhance segmentation in 2D cine cardiac MRI. In: *International Conference on Functional Imaging and Modeling of the Heart*. Bordeaux: Springer (2019). p. 361–9. doi: 10.1007/978-3-030-21949-9_39

53. Lamata P, Niederer S, Nordsletten D, Barber DC, Roy I, Hose DR, et al. An accurate, fast and robust method to generate patient-specific cubic Hermite meshes. *Med Image Anal*. (2011) 15:801–13. doi: 10.1016/j.media.2011.06.010

54. Lamata P, Niederer S, Barber D, Norsletten D, Lee J, Hose R, et al. Personalization of cubic hermite meshes for efficient biomechanical simulations. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Beijing: Springer (2010). p. 380–7. doi: 10.1007/978-3-642-15745-5_47

55. Ranjan A, Bolkart T, Sanyal S, Black MJ. Generating 3D faces using convolutional mesh autoencoders. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Munich. (2018). p. 704–20. doi: 10.1007/978-3-030-01219-9_43

56. Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. In: Lee DD, Sugiyama M, von Luxburg U, Guyon I, Garnett R, editors. *Advances in Neural Information Processing Systems 29*. Barcelona: Curran Associates, Inc. (2016). p. 3837–3845.

57. Hammond DK, Vandergheynst P, Gribonval R. Wavelets on graphs via spectral graph theory. *Appl Comput Harmonic Anal*. (2011) 30:129–50. doi: 10.1016/j.acha.2010.04.005

58. Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, et al. beta-VAE: learning basic visual concepts with a constrained variational framework. In: *5th International Conference on Learning Representations (ICLR)*. Toulon. (2017). p. 1–13.

59. Bowman SR, Vilnis L, Vinyals O, Dai AM, Jozefowicz R, Bengio S. Generating sentences from a continuous space. *arXiv preprint arXiv:151106349*. (2015). doi: 10.18653/v1/K16-1002

60. Kingma DP, Ba J. Adam: a method for stochastic optimization. *arXiv preprint arXiv:14126980*. (2014). doi: 10.48550/arXiv.1412.6980

61. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, editors. *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc. (2019). p. 8024–35. doi: 10.48550/arXiv.1912.01703

62. Fey M, Lenssen JE. Fast graph representation learning with PyTorch geometric. In: *ICLR Workshop on Representation Learning on Graphs and Manifolds*. New Orleans. (2019).

63. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. (2011) 12:2825–30. doi: 10.5555/1953048.2078195

64. Burda Y, Grosse R, Salakhutdinov R. Importance weighted autoencoders. *arXiv preprint arXiv:150900519*. (2015). doi: 10.48550/arXiv.1509.00519

65. Xu H, Zacur E, Schneider JE, Grau V. Ventricle surface reconstruction from cardiac MR slices using deep learning. In: *International Conference on Functional Imaging and Modeling of the Heart*. Bordeaux: Springer (2019). p. 342–51. doi: 10.1007/978-3-030-21949-9_37

66. Petersen SE, Matthews PM, Bamberg F, Bluemke DA, Francis JM, Friedrich MG, et al. Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of UK Biobank-rationale, challenges and approaches. *J Cardiovasc Magnet Reson*. (2013) 15:1–10. doi: 10.1186/1532-429X-15-46

67. Sohn K, Lee H, Yan X. Learning structured output representation using deep conditional generative models. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, editors. *Advances in Neural Information Processing Systems 28.* Montreal, QC: Curran Associates, Inc. (2015). p. 3483–3491.

68. Ma Q, Tang S, Pujades S, Pons-Moll G, Ranjan A, Black MJ. Dressing 3D humans using a conditional Mesh-VAE-GAN. *arXiv preprint arXiv:190713615.* (2019). doi: 10.48550/arXiv.1907.13615

frontiers | Frontiers in Cardiovascular Medicine

TYPE Original Research
PUBLISHED 10 January 2023
DOI 10.3389/fcvm.2022.1016703

MITEA: A dataset for machine learning segmentation of the left ventricle in 3D echocardiography using subject-specific labels from cardiac magnetic resonance imaging

Debbie Zhao[1]*, Edward Ferdian[2], Gonzalo D. Maso Talou[1], Gina M. Quill[1], Kathleen Gilbert[1], Vicky Y. Wang[1], Thiranja P. Babarenda Gamage[1], João Pedrosa[3], Jan D'hooge[4], Timothy M. Sutton[5], Boris S. Lowe[6], Malcolm E. Legget[7], Peter N. Ruygrok[6,7], Robert N. Doughty[6,7], Oscar Camara[8], Alistair A. Young[2,9] and Martyn P. Nash[1,10]*

[1]Auckland Bioengineering Institute, University of Auckland, Auckland, New Zealand, [2]Department of Anatomy and Medical Imaging, University of Auckland, Auckland, New Zealand, [3]Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Porto, Portugal, [4]Department of Cardiovascular Sciences, KU Leuven, Leuven, Belgium, [5]Counties Manukau Health Cardiology, Middlemore Hospital, Auckland, New Zealand, [6]Green Lane Cardiovascular Service, Auckland City Hospital, Auckland, New Zealand, [7]Department of Medicine, University of Auckland, Auckland, New Zealand, [8]Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain, [9]Department of Biomedical Engineering, King's College London, London, United Kingdom, [10]Department of Engineering Science, University of Auckland, Auckland, New Zealand

Segmentation of the left ventricle (LV) in echocardiography is an important task for the quantification of volume and mass in heart disease. Continuing advances in echocardiography have extended imaging capabilities into the 3D domain, subsequently overcoming the geometric assumptions associated with conventional 2D acquisitions. Nevertheless, the analysis of 3D echocardiography (3DE) poses several challenges associated with limited spatial resolution, poor contrast-to-noise ratio, complex noise characteristics, and image anisotropy. To develop automated methods for 3DE analysis, a sufficiently large, labeled dataset is typically required. However, ground truth segmentations have historically been difficult to obtain due to the high inter-observer variability associated with manual analysis. We address this lack of expert consensus by registering labels derived from higher-resolution subject-specific cardiac magnetic resonance (CMR) images, producing 536 annotated 3DE images from 143 human subjects (10 of which were excluded). This heterogeneous population consists of healthy controls and patients with

Check for updates

OPEN ACCESS

EDITED BY
Ahmed S. Fahmy,
Harvard Medical School, United States

REVIEWED BY
Manuel Morales,
Beth Israel Deaconess Hospital,
United States
Arkadiusz Sitek,
Massachusetts General Hospital
and Harvard Medical School,
United States

*CORRESPONDENCE
Debbie Zhao
✉ debbie.zhao@auckland.ac.nz
Martyn P. Nash
✉ martyn.nash@auckland.ac.nz

SPECIALTY SECTION
This article was submitted to Cardiovascular Imaging, a section of the journal Frontiers in Cardiovascular Medicine

RECEIVED 11 August 2022
ACCEPTED 06 December 2022
PUBLISHED 10 January 2023

CITATION
Zhao D, Ferdian E, Maso Talou GD, Quill GM, Gilbert K, Wang VY, Babarenda Gamage TP, Pedrosa J, D'hooge J, Sutton TM, Lowe BS, Legget ME, Ruygrok PN, Doughty RN, Camara O, Young AA and Nash MP (2023) MITEA: A dataset for machine learning segmentation of the left ventricle in 3D echocardiography using subject-specific labels from cardiac magnetic resonance imaging. *Front. Cardiovasc. Med.* 9:1016703. doi: 10.3389/fcvm.2022.1016703

cardiac disease, across a range of demographics. To demonstrate the utility of such a dataset, a state-of-the-art, self-configuring deep learning network for semantic segmentation was employed for automated 3DE analysis. Using the proposed dataset for training, the network produced measurement biases of -9 ± 16 ml, -1 ± 10 ml, -2 ± 5 %, and 5 ± 23 g, for end-diastolic volume, end-systolic volume, ejection fraction, and mass, respectively, outperforming an expert human observer in terms of accuracy as well as scan-rescan reproducibility. As part of the Cardiac Atlas Project, we present here a large, publicly available 3DE dataset with ground truth labels that leverage the higher resolution and contrast of CMR, to provide a new benchmark for automated 3DE analysis. Such an approach not only reduces the effect of observer-specific bias present in manual 3DE annotations, but also enables the development of analysis techniques which exhibit better agreement with CMR compared to conventional methods. This represents an important step for enabling more efficient and accurate diagnostic and prognostic information to be obtained from echocardiography.

KEYWORDS

3D echocardiography (3DE), machine learning (ML), segmentation (image processing), left ventricle (LV), multimodal imaging, cardiac magnetic resonance (CMR) imaging, domain adaptation, Cardiac Atlas Project

1. Introduction

Machine learning (ML) has shown considerable promise for automated analysis and interpretation in the domain of cardiovascular imaging (1, 2). Already, its application to cardiac magnetic resonance (CMR) imaging has exhibited excellent results with high accuracy and reproducibility by leveraging several large cohort databases such as the UK Biobank (3–5). Although CMR offers higher spatial resolution and tissue contrast for the assessment of cardiac mass and volume, transthoracic echocardiography remains at the frontline of cardiac imaging as the most widely used and readily accessible modality for screening, diagnosis, and management of cardiovascular disease. Technological advances in ultrasonography have enabled three-dimensional echocardiography (3DE), consequently removing the dependency on accurate plane positioning and geometric assumptions required for standard two-dimensional echocardiography (2DE). As a result, several studies have shown that 3DE-derived measurements are generally superior to 2DE in terms of chamber quantification accuracy (6, 7), reproducibility (8), and prognostic power (9). Despite these advantages, 3DE has not yet been universally integrated into clinical practice for the assessment of cardiac function due to limitations in image quality, and increased costs associated with acquisition and long analysis times compared with 2DE.

In comparison to other cardiac imaging modalities, analysis of 3DE is particularly challenging owing to the limited spatial resolution, low contrast-to-noise ratio (CNR), complex noise characteristics (speckle in combination with common artifacts), and image anisotropy. Several factors can influence the image quality of 3DE including, but not limited to, sonographer experience, vendor-specific processing, acquisition settings, and patient body habitus. Discrepancies in the delineation of important cardiac structures, such as the left ventricle (LV), compared to those from a reference modality such as CMR, have been shown to be observer- and software-dependent, as well as exhibit regional variability in terms of the magnitude of differences in geometry (10). In particular, acoustic shadowing and signal dropout further compromise local image quality, leading to greater inter- and intra-observer variability in manual annotations at these locations. To address this, statistical shape priors (or atlases) can be used to provide suitable estimates in regions where image information is corrupted or missing (11–14). However, these approaches are ultimately limited by the generalizability of such templates and may be ill-suited in cases of atypical anatomy.

The primary challenge associated with the development of automated methods for 3DE analysis is the prerequisite of a sufficiently large training dataset. Historically, reference annotations have been difficult to obtain due to the high degree of variability associated with manual 3DE segmentation, thus, limiting the scope of ML-based solutions. Currently, the dataset belonging to the Challenge on Endocardial Three-dimensional Ultrasound Segmentation (CETUS)[1] (15), organized as part of

[1] https://www.creatis.insa-lyon.fr/Challenge/CETUS/

the 2014 Medical Image Computing and Computer Assisted Interventions (MICCAI) conference, remains the only publicly available resource. This dataset consists of expert-annotated 3DE images from 45 subjects, for which data from 15 subjects are made publicly available for training. Due to the lack of clear guidelines for endocardial contouring in 3DE, considerable effort was expended in establishing a consistent analysis regime amongst three expert observers. Despite this, large inter-expert variability was reported and an agreement was only reached after several revisions and consensus discussions (16). Nevertheless, efforts in providing a publicly accessible benchmark such as the CETUS platform represent an important step toward the development of automated 3DE analysis methods.

Alternative approaches for generating training data involve producing synthetic 3DE images via *in silico* simulations (17, 18) or generative adversarial networks (19, 20). While these methods do not require additional segmentation (as the underlying anatomy is known in such cases), synthetic datasets are often unable to adequately capture all features found in real images. Unsupervised domain adaptation strategies have also gained interest in medical imaging applications, enabling knowledge gained from higher-resolution images or data to improve the segmentation of lower-resolution or degraded images (21–23). However, as with unsupervised methods in general, it cannot be certain that the model is optimized for the target domain.

Alongside the ongoing advances in 3DE acquisition systems, more accurate and efficient analysis methods will substantially benefit patient care and management. Having acknowledged the lack of expert consensus in obtaining reference annotations, and the limitations associated with population shape priors and synthetic data, we instead leveraged subject-specific labels from CMR acquired in a heterogenous population of 134 subjects. Here, we present MITEA (MR-Informed Three-dimensional Echocardiography Analysis): an annotated 3DE dataset for the segmentation of the LV myocardium and cavity for quantification of systolic function and mass, and subsequently show how this data can be used to train a deep learning model for automated 3DE analysis. The full annotated 3DE dataset and trained model can be accessed as part of the Cardiac Atlas Project[2] (24). To date, this represents the largest publicly available 3DE dataset, and the first which uses labels derived from subject-specific CMR analyses.

2. Materials and methods

Non-invasive multimodal 3DE and CMR imaging were performed within two hours in 144 prospectively recruited participants (87 healthy subjects with no existing or history of cardiac disease; and 57 patients with acquired, non-ischemic cardiac disease), of which 134 (82 healthy subjects; and 52 patients with cardiac disease) were included in the study. Ethical approval for this research was granted by the Health and Disability Ethics Committee of New Zealand (17/CEN/226). Written informed consent was obtained from each participant.

Multimodal data belonging to 70 of these subjects have been previously presented as part of an investigation into systematic measurement biases between 3DE and CMR (10). The present study extends upon this work by: inclusion of additional disease cases for improved generalizability; inclusion of scan-rescan 3DE images to assess repeatability; and utilization of paired multimodal data for the development of automated 3DE segmentation techniques. An overview of the method for data generation is illustrated in **Figure 1**, and detailed in the following subsections.

2.1. Multimodal image acquisition

Transthoracic real-time (single-cycle) 3DE images were acquired using a Siemens ACUSON SC2000 Ultrasound System and a 4Z1c matrix array transducer (Siemens Medical Solutions, Mountain View, CA, USA) with 36 × 48 (1,728) elements. Targeted images of the LV were acquired from the apical window in a steep left lateral decubitus position during breath-holds. Parameters (including choice of fundamental or harmonic imaging, depth, gain, compression, and width of the volumetric dataset) were optimized by an experienced sonographer on a per-subject basis to maximize the image volume sampling rate, while maintaining adequate spatial resolution for analysis. To measure scan-rescan repeatability, two 3DE clips were acquired per subject (producing a total of 268 3DE datasets across the 134 included participants). All acquisitions were reconstructed into 3D Cartesian image volumes (with a rectangular bounding box and zero-values outside the pyramidal volume) using 1 mm isotropic voxels.

Multi-planar cine CMR imaging was performed on either a Siemens Magnetom 1.5T Avanto Fit ($n = 77$) or 3T Skyra ($n = 57$) scanner (Siemens Healthcare, Erlangen, Germany) with an 18-channel body matrix coil, using a retrospectively gated balanced steady-state free precession sequence under breath-holds. Acquired planes included three long-axis slices (standard two-, three-, and four-chamber views) and a short-axis stack of 6–10 slices (spanning the length of the LV from mitral valve to apex) over one cardiac cycle, with the following typical imaging parameters: TR = 3.7 ms, TE = 1.6 ms, flip angle = 45°, field of view = 360 mm × 360 mm, in-plane resolution = 1.4 mm × 1.4 mm, and slice thickness = 6 mm, in keeping with standard protocols.

[2] https://www.cardiacatlas.org/

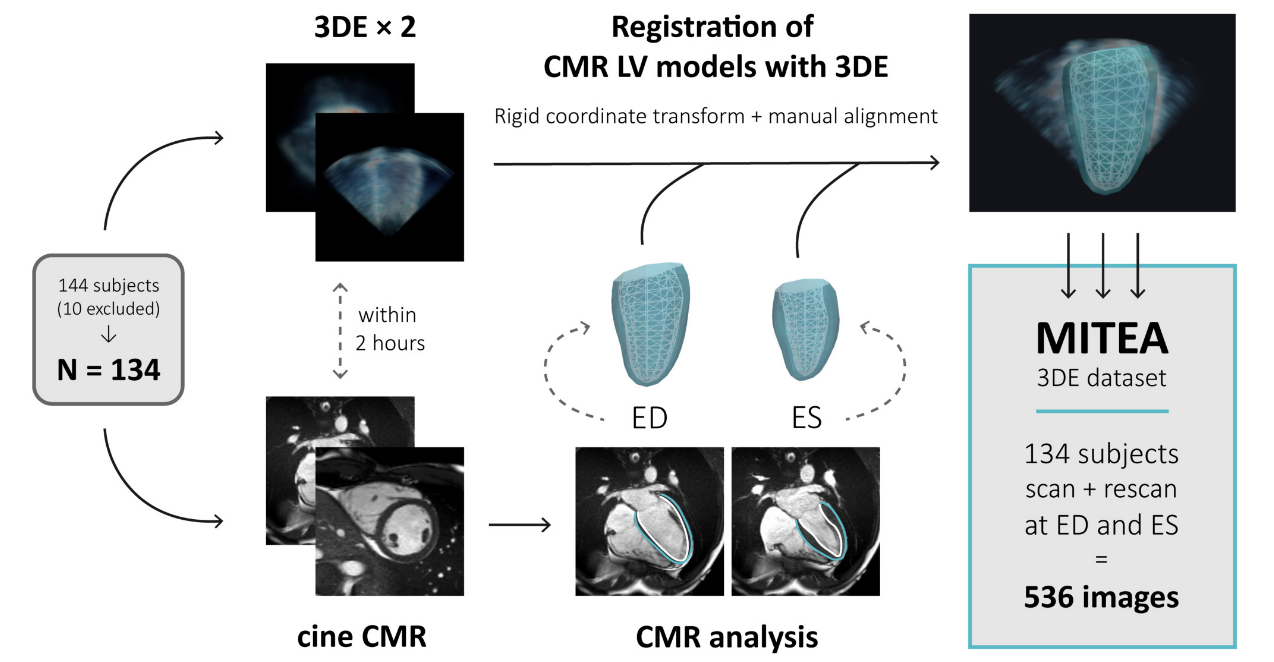

3DE × 2
Registration of CMR LV models with 3DE
Rigid coordinate transform + manual alignment
144 subjects (10 excluded)
N = 134
within 2 hours
ED
ES
MITEA
3DE dataset
134 subjects scan + rescan at ED and ES
=
536 images
cine CMR
CMR analysis

FIGURE 1
Method overview for generation of the MR-Informed Three-dimensional Echocardiography Analysis (MITEA) dataset, showing paired multimodal imaging using 3D echocardiography (3DE) and cardiac magnetic resonance (CMR) imaging. The registration of CMR-derived left ventricular geometries was performed at end-diastole (ED) and end-systole (ES) to produce subject-specific labels for the myocardium and cavity.

With these settings, an average of 29 ± 4 (range 20–44) frames per cardiac cycle were obtained for the included study population.

2.2. Image analysis

Patients were subjectively graded on a five-point 3DE image quality scale (poor, suboptimal, adequate, good, excellent) by a single expert (independent of the sonographer who acquired the images). This subjective score was based on a combination of perceived endocardial border definition (i.e., the overall sharpness of the LV cavity due to ultrasound attenuation, choice of harmonics, and selection of gains and compression), and the visibility of wall segments (relating to signal dropout and LV coverage due to probe alignment and selection of an adequate pyramidal volume size). After qualitative grading, the ratio between the mean difference and variance in signal intensity between the LV myocardium and cavity were calculated to provide a quantitative measure of CNR (25), given by:

$$\mathrm{CNR} = \frac{|\mu_{\mathrm{myocardium}} - \mu_{\mathrm{cavity}}|}{\sqrt{\sigma_{\mathrm{myocardium}}{}^2 + \sigma_{\mathrm{cavity}}{}^2}}$$

where $\mu_{\mathrm{myocardium}}$ and μ_{cavity} are the mean signal intensities in the regions belonging to the myocardium and cavity, respectively, and $\sigma_{\mathrm{myocardium}}$ and σ_{cavity} are the corresponding standard deviations.

To generate subject-specific labels from CMR, time-varying geometric models of the LV over one cardiac cycle were constructed semi-automatically by guide-point modeling (26) using *Cardiac Image Modeler* (CIM, Version 8.1, University of Auckland, New Zealand), by a single analyst. To create an initial coarse geometry and to establish the LV position and orientation, fiducial landmarks (i.e., the base of the myocardium in the long-axis slices; apical and basal centroids in the corresponding short axis slices, and insertion points of the right ventricle (RV) along the LV epicardial border in the short-axis slices, where visible) were manually identified. This was subsequently refined by interactively fitting contours to the endocardial and epicardial borders on both the long- and short-axis slices, and manually correcting in-plane breath-hold mis-registrations using the image intersections. Papillary muscles and trabeculations were included within the LV cavity (**Figure 2A**). This analysis generated a bicubic Hermite and linear finite element model of the LV (27), with the origin positioned at one-third of the distance from base to apex, with the LV long axis parallel to the *x*-axis, and the center of the RV directed toward the orthogonal *y*-axis. From the model, 145 unique points were sampled per surface (for the endocardium and epicardium) to produce a mesh consisting of 290 3D rectangular Cartesian (*x*, *y*, *z*) vertices

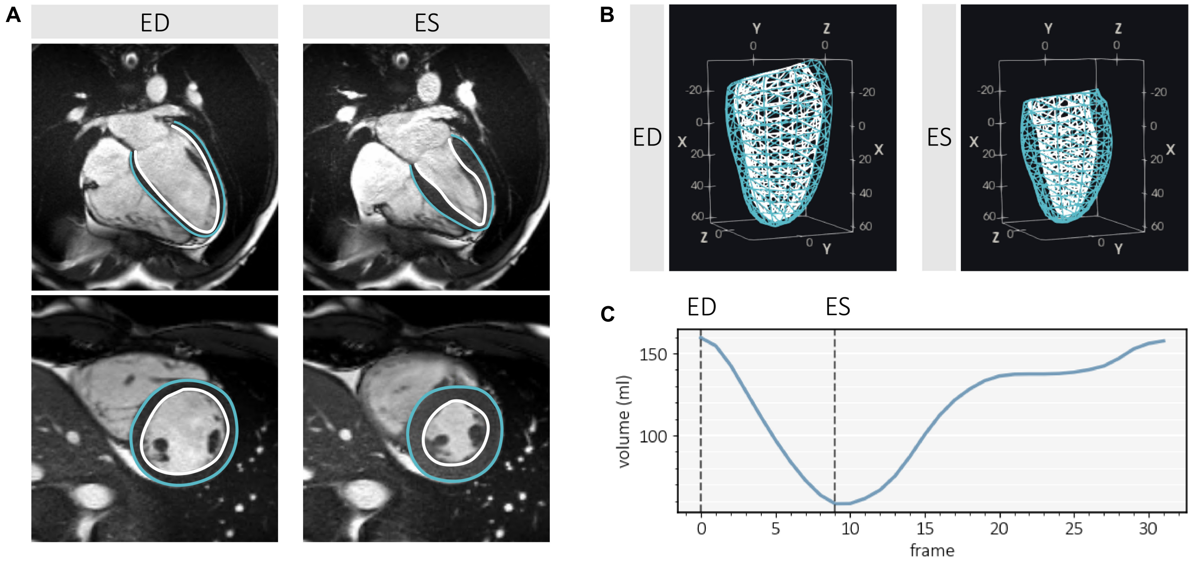


FIGURE 2
Image analysis and 3D left ventricle (LV) geometry extraction from cardiac magnetic resonance (CMR) using Cardiac Image Modeler (CIM, Version 8.1, University of Auckland, New Zealand) at end-diastole (ED) and end-systole (ES). **(A)** Contour examples of the endocardium (white) and epicardium (blue) on a 4-chamber long axis slice and mid-ventricular short axis slice, showing exclusion of trabeculae and papillary muscles from the myocardium. **(B)** 3D surface meshes (dimensions in mm) sampled from the LV finite element model. **(C)** Volume-time curve generated from CMR image analysis, indicating frame indices of interest.

representing the LV myocardium. Static 3D LV geometries were extracted at end-diastole (ED) and end-systole (ES) (Figure 2B), corresponding to the first CMR image frame, and the image frame associated with the smallest cavity volume, respectively (Figure 2C).

2.3. Multimodal registration and label generation

For each subject, registration of CMR with 3DE was performed in two steps, comprising an automated coarse alignment of the global LV position, followed by a manual refinement of the LV model within the 3DE image volume. To establish the initial transform at ED, the B-spline Explicit Active Surfaces (BEAS) algorithm (14) was used to create a fully automated segmentation of the LV from 3DE, from which a vector connecting the apex and basal centroid was extracted to represent the LV long axis orientation and position with respect to each 3DE acquisition. To differentiate between the circumferential wall segments, the direction of the RV center from the central axis was approximated as being 70 degrees from the inferior RV insertion [automatically detected based on image features as part of the BEAS segmentation (28)], as the anterior insertion is generally not well visualized in 3DE. The resultant axes were subsequently registered to the cardiac coordinate system used in the finite element model of the LV in Section "2.2 Image analysis," yielding a transformation matrix representing the rigid mapping between the 3DE image LV model coordinate systems. This transformation was subsequently applied to initially align the CMR-derived LV model to the 3DE image for each subject.

The initial alignment was refined by manually applying rigid translations and rotations using an open-source data analysis and visualization application (ParaView 5.8.0) (29) (Figure 3), by the same expert that carried out subjective 3DE image quality grading and CMR analysis. Manual registrations were performed at two frames only, representing ED and ES. For CMR, the relevant static LV geometries were extracted according to the method described in Section "2.2 Image analysis" and Figure 2C. For 3DE, ED and ES image frames were manually selected corresponding to when the cavity appeared largest and smallest. The manual refinement was performed independently for the ED frame, and further adjusted at ES, as required, to account for any changes in relative transducer angle and position over the cardiac cycle during acquisition. All manual alignments were carried out by a single observer, resulting in 536 (134 included subjects × 2 clips × 2 frames) independent alignments.

The closed meshes were subsequently converted into 3D masks of equal dimensions to the corresponding Cartesian 3DE images, containing two foreground label classes (representing the cavity and myocardium). Of note, foreground label regions were not constrained to the pyramidal volume, as shown in Figure 4.

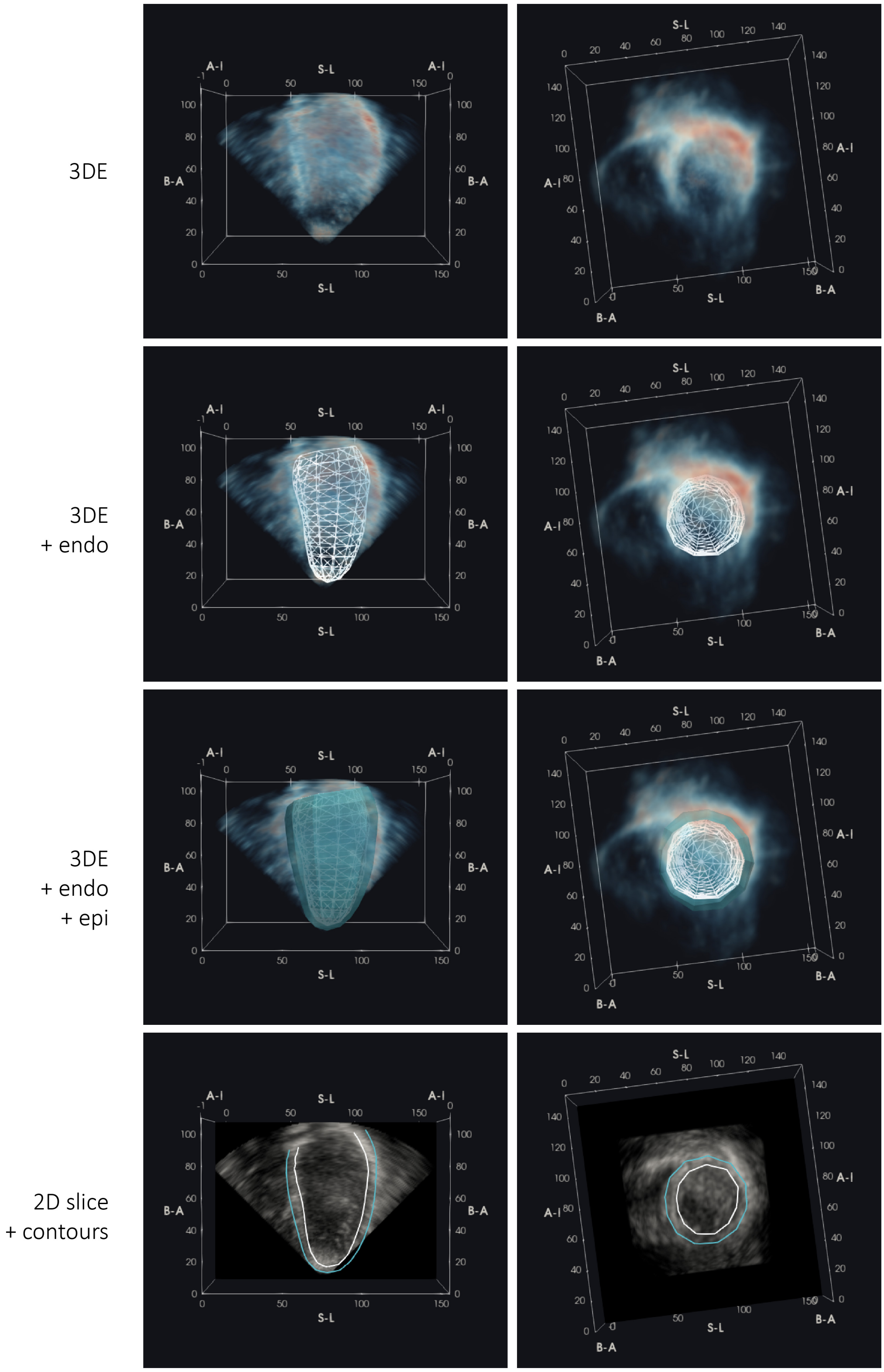


FIGURE 3
Registration of 3D echocardiography (3DE) with subject-specific geometries of the left ventricle (LV) derived from cardiac magnetic resonance (CMR) at end-diastole, showing an example 3DE image volume (visualized with an opacity transfer function on a blue-to-red colormap) and corresponding 2D mid-ventricular image slice and contours of the endocardium (endo) and epicardium (epi), viewed longitudinally and axially. Labels denote anatomical LV aspects: B-A, base-to-apex; S-L, septal-lateral; A-I, anterior-inferior. All dimensions are in mm.

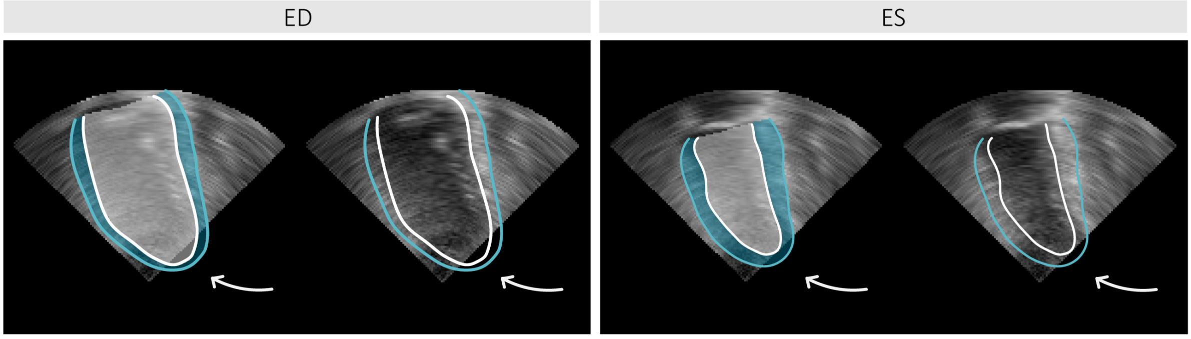


FIGURE 4
Example of an annotated 3D echocardiography (3DE) image sliced longitudinally at end-diastole (ED) and end-systole (ES), showing portions of labeled regions and corresponding contours for the left ventricular cavity (white) and myocardium (blue) falling outside the acquired 3DE pyramidal volume (as indicated by the arrows).

2.4. Deep learning segmentation experiment

To demonstrate the application of the dataset for deep learning, *nnU-Net* (30), a self-configuring network for semantic segmentation was employed for automated 3DE analysis. An 80/20 split was used for training and testing, with images from the same acquisition (i.e., ED and ES from the same cycle) and clips from the same participant (i.e., scan and rescan) grouped together. This resulted in data from 107 unique participants being included in the training set (a total of 428 paired images and labels), and data from 27 participants in the testing set (108 paired images and labels). The network was trained using fivefold cross-validation with a further 80/20 split for training and validation, producing five model instances (each trained using data from 85 or 86 participants), which were ensembled (by averaging softmax probabilities) for inference.

Using the 3D full-resolution U-Net configuration with no cascade, each fold was trained for 200 epochs (chosen empirically based on stable validation loss curves), where each epoch consisted of 250 iterations over shuffled batches of size two. Stochastic gradient descent with a large Nesterov momentum (31) ($\mu = 0.99$) and a high initial learning rate of 0.01 [reduced by $(1 - \text{epoch}_{\text{current}}/\text{epoch}_{\text{max}})^{0.9}$] using the *polyLR* schedule (32), producing an almost-linear decrease to zero, was used for optimization, with the sum of cross-entropy and Dice as the loss function. To diversify the data and increase model robustness, on-the-fly data augmentations including rotation, scaling, mirroring, and low-resolution simulation (by means of downsampling followed by upsampling), were applied during training. Training time was approximately 170 s per epoch on an NVIDIA Tesla V100 GPU with 32GB memory. With the exception of a reduction in the number of epochs (set to 1,000 by default) to reduce overfitting, the model was deployed with all other out-of-the-box parameters for pre-processing, network architecture selection, training, and post-processing. The self-configured architecture for the present dataset is shown in **Figure 5**.

2.5. Validation and performance

Model performance was evaluated on the testing set ($n = 27$ subjects, 54 acquisitions) in terms of segmentation accuracy at ED and ES using the Dice coefficient, mean surface distance (MSD), and Hausdorff distance (HD); as well as the agreement in routine clinical cardiac indices including LV end-diastolic volume (EDV), end-systolic volume (ESV), mass (LVM) (calculated as the average of mass at ED and ES), and ejection fraction (EF). Clinical measurements were also compared with those derived from conventional manual analysis using TOMTEC 4D LV-ANALYSIS 3 (TOMTEC Imaging Systems GmbH, Unterschleißheim, Germany), a commercially available, vendor-neutral software platform for 3DE quantification, performed by a single expert for the 27 test subjects (including rescans).

2.6. Statistics

Paired-sample *t*-tests were used to identify statistically significant measurement biases (calculated as $\text{index}_{\text{3DE}} - \text{index}_{\text{CMR}}$) in cardiac indices derived from 3DE (either by *nnU-Net* or manually) with respect to those obtained from corresponding CMR analyses, and Bland-Altman plots were used to visualize the agreement between paired variables. The *f*-test of equality of variances was used to assess the significance of the reduction in the standard deviation of errors when using *nnU-Net* instead of expert manual analyses in terms of measurement accuracy (with respect to CMR), as well as scan-rescan repeatability. Finally, to assess the reliability between

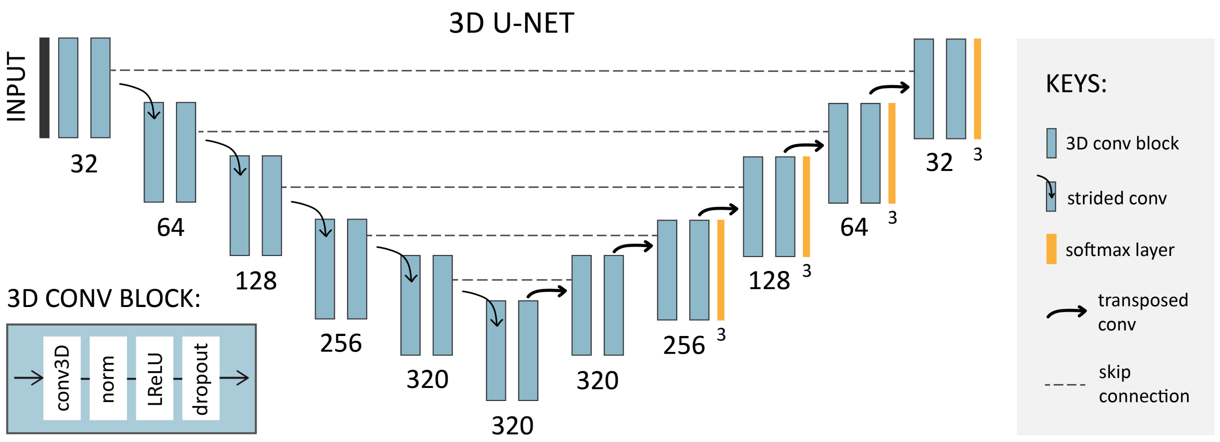


FIGURE 5
Network architecture configured by *nnU-Net* for 3DE segmentation. Each 3D convolution (conv) block consists of a plain convolution, followed by instance normalization (norm), leaky ReLU (LReLU), and dropout. Downsampling is achieved using strided convolutions (stride two), and upsampling by transposed convolutions. Numbers indicate the number of channels corresponding to each convolution block.

paired measurements, an intraclass correlation coefficient (ICC) using a two-way, mixed effects model for absolute agreement, was calculated for each index. Based on established guidelines (33), threshold values of <0.5, ≥ 0.5, ≥ 0.75, and ≥ 0.9, represented poor, moderate, good, and excellent reliability, respectively. For the quantification of absolute scan-rescan variability due to random measurement error (34), repeatability coefficients with 95% confidence (35) were also computed. All statistical tests were two-tailed and deemed significant for p-values < 0.05, and analyses were performed using IBM SPSS Statistics for Windows (Version 26.0, IBM Corp., Armonk, NY, USA).

3. Results

3.1. Population summary

Demographics (including age, sex, and body surface area) and CMR-derived LV indices for the included population are summarized in Table 1. The disease group comprised 14 patients with LV hypertrophy, 12 patients with cardiac amyloidosis, 10 patients with aortic regurgitation, eight patients with hypertrophic cardiomyopathy, six patients with dilated cardiomyopathy, and two heart transplant recipients.

3.2. Image characteristics

Images of at least suboptimal quality (n = 134) were included for analysis, leaving 10 datasets that were excluded due to poor quality. Figure 6 shows examples of 3DE images ranging from poor to excellent quality, as well as the distribution of image quality across the population. Of the 10 excluded cases, five were healthy controls, and five were patients with cardiac disease. A summary of 3DE image dimensions and acquired frames per cycle is presented in Table 2.

TABLE 1 Summary of participant demographics including age, sex, body surface area (BSA) calculated using the Mosteller formula (36), and body mass index (BMI); and indices derived from cardiac magnetic resonance imaging including left ventricular end-diastolic volume (EDV), end-systolic volume (ESV), mass (LVM), and ejection fraction (EF), for the included dataset.

	Control (n = 82)	Disease (n = 52)	Total (n = 134)
Age (years)	37 ± 16 (18–74)	62 ± 15 (18–84)	47 ± 20 (18–84)
Male sex [frequency (%)]	42 (51%)	39 (75%)	81 (60%)
BSA (m^2)	1.83 ± 0.21 (1.39–2.25)	2.01 ± 0.25 (1.46–2.72)	1.90 ± 0.24 (1.39–2.72)
BMI (kg/m^2)	24.0 ± 3.6 (16.9–34.2)	28.3 ± 5.5 (16.7–48.9)	25.7 ± 4.9 (16.7–48.9)
EDV (ml)	139 ± 31 (74–220)	166 ± 44 (101–314)	150 ± 39 (74–314)
ESV (ml)	53 ± 16 (19–103)	74 ± 38 (29–235)	61 ± 29 (19–235)
LVM (g)	110 ± 30 (58–171)	170 ± 51 (88–314)	133 ± 49 (58–314)
EF (%)	62 ± 5 (51–74)	57 ± 12 (25–78)	60 ± 9 (25–78)
HR difference (bpm)	−1 ± 7 (−22–25)	−1 ± 6 (−13–38)	−1 ± 6 (−22–38)

The difference in heart rate (HR) between 3DE and CMR acquisitions (calculated as $HR_{3DE}-HR_{CMR}$) is provided as an indication of HR variability between modalities. Continuous variables are presented as mean ± standard deviation (range).

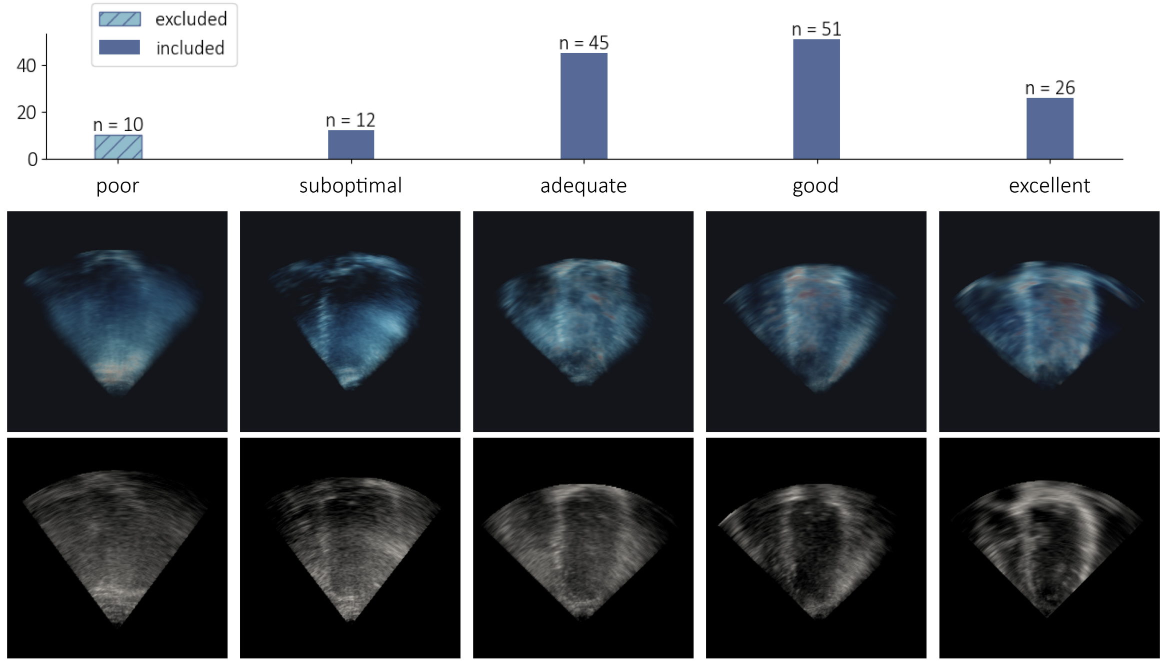


FIGURE 6
Examples of reconstructed 3D echocardiographic image volumes (visualized with an opacity transfer function on a blue-to-red colormap) and corresponding 2D mid-volume longitudinal slices (grayscale), showing variable quality (subjectively scored from poor to excellent). A total of 10 subjects were excluded from the study due to poor image quality.

TABLE 2 Summary of 3D echocardiography (3DE) image parameters including Cartesian image dimensions in *X* (elevation), *Y* (azimuth), *Z* (depth, i.e., apex-to-base) directions, the number of frames acquired per cycle, and the contrast-to-noise-ratio (CNR) associated with subjective quality scores across the included study population.

	Dimensions (mm)			Frames per cycle	CNR (dB)			
$n = 268$	*X*	*Y*	*Z*		Suboptimal	Adequate	Good	Excellent
Mean	167	168	132	36	0.526	0.649	0.831	0.930
SD	25	26	14	12	0.110	0.144	0.148	0.146
Min.	106	117	101	12	0.346	0.209	0.480	0.675
Max.	243	243	172	69	0.726	0.920	1.128	1.377

Presented values include the mean, standard deviation (SD), minimum (min.), and maximum (max.) for each parameter.

3.3. Segmentation accuracy

Figure 7 illustrates the distribution of segmentation accuracy scores obtained by the ensembled *nnU-Net* model with respect to the cavity and myocardium, evaluated on the training set ($n = 428$, consisting of data from 107 subjects × 2 clips × 2 frames) and testing set ($n = 108$ images, consisting of data from 27 subjects × 2 clips × 2 frames). Mean test scores were Dice coefficient = 0.766, MSD = 1.6 mm, and HD = 9.1 mm for the myocardium; and Dice coefficient = 0.871, MSD = 1.8 mm, and HD = 8.0 mm, for the cavity. Segmentation metrics for each of the five separate model instances evaluated on the testing set is provided in the Supplementary material. For comparison, corresponding mean scores (averaged between the reported values for ED and ES) obtained by the most accurate method for LV cavity segmentation in the fully automatic category by Barbosa et al. (37) and Queirós et al. (38) of the 2014 MICCAI CETUS challenge were Dice coefficient = 0.878, MSD = 2.4 mm, and HD = 8.2 mm.

From visual assessment, *nnU-Net* produced reasonable myocardium and cavity segmentations for all test images at both ED and ES. Mis-segmentations occurred most frequently where LV boundaries were missing from the image, with one such example illustrated in Figure 8A. Here, the reference annotations show that a substantial portion of the cavity and myocardium falls outside the acquired pyramidal volume. Where the pyramidal volume adequately encompassed the LV, segmentations were generally accurate, as shown in Figures 8B, C.

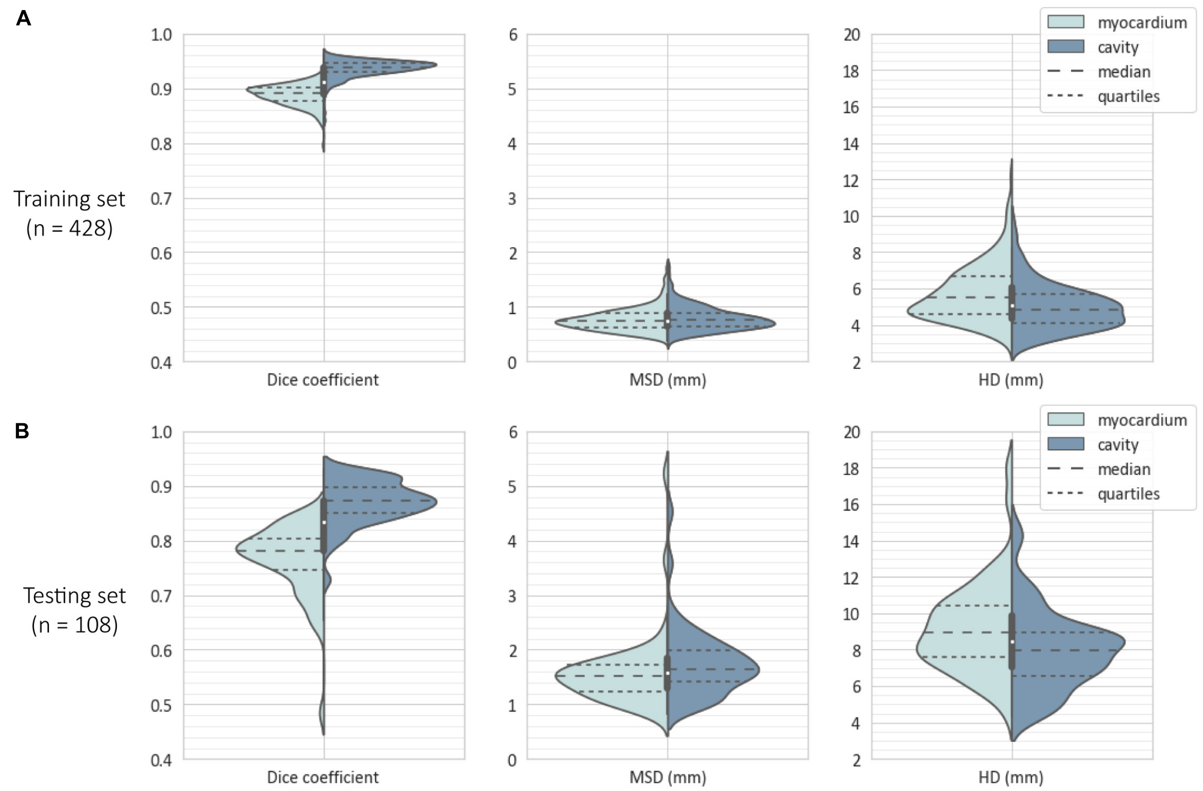


FIGURE 7
Violin plots showing the distribution and quartiles of segmentation scores in terms of Dice coefficient, mean surface distance (MSD), and Hausdorff distance (HD), evaluated on **(A)** the training set ($n = 428$ images) and **(B)** the testing set ($n = 108$ images). For each metric, distributions are split into the two foreground classes (i.e., myocardium and cavity), with the central box plot derived from the data of both classes as an estimate of the overall score.

3.4. Agreement in cardiac indices

Agreement and reliability in clinical cardiac indices between CMR and 3DE (calculated from *nnU-Net* segmentations and expert manual analyses using TOMTEC) are presented in Table 3. Higher ICC values (representing measurement reliability with respect to CMR) were observed across all cardiac indices, with significant reductions in the magnitude of bias for EDV, ESV, and LVM when using *nnU-Net* in place of expert manual analyses for recovering CMR-derived cardiac indices. Bland-Altman analyses revealed narrower 95% limits of agreement in all cardiac indices for *nnU-Net* compared to expert manual analyses, with no apparent proportional bias (Figure 9).

3.5. Scan-rescan repeatability

The variability in repeated 3DE measurements is summarized in Table 4. Both expert manual analyses and *nnU-Net* exhibited excellent reliability between scan-rescan measurements (ICC > 0.9), with the reliability of *nnU-Net* being higher for all cardiac indices. Similarly, *nnU-Net* outperformed the expert human observer in terms of significantly smaller magnitudes of variance in scan-rescan biases (again for all cardiac indices), suggesting that measurements obtained using *nnU-Net* were more consistent.

4. Discussion

Guidelines and recommendations for LV chamber quantification using echocardiography state 3DE as the preferred method of volumetric assessment (over conventional 2DE), where available and feasible (39), in keeping with the advantage of 3DE in being able to circumvent the need for geometric assumptions. Recently published normative values stratified by age, sex, and ethnic groups by the World Alliance Societies of Echocardiography (WASE) (40) further endorses the use of 3DE for the assessment of LV chamber size and function. Nevertheless, 3DE has not yet been universally incorporated into standard clinical routine due to requiring specialized expertise in both acquisition and analysis, resulting in higher costs compared to 2DE. Likewise, the generation of large amounts of expert manual annotations

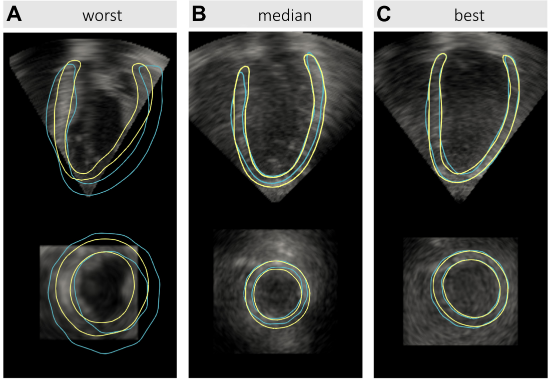


FIGURE 8
Comparison of left ventricular (LV) segmentations by *nnU-Net* (yellow) against reference labels (blue) derived from cardiac magnetic resonance (CMR) imaging for test images at end-diastole, representing the: **(A)** worst, **(B)** median, and **(C)** best model performances. For visualization purposes, 3D masks have been converted to contours representing the LV myocardium corresponding to longitudinal (top row) and axial (bottom row) slices. The resulting Dice coefficients for the myocardium and cavity were: 0.440 and 0.754, respectively, for the worst case; 0.741 and 0.924, respectively, for the median case; and 0.856 and 0.950, respectively, for the best case.

for the development of automated 3DE analysis methods has historically been a tedious and complex task. Having recognized the inter-expert variability in manual analysis (such as was experienced during the organization of the CETUS challenge), we sought to instead leverage the higher resolution and contrast of CMR in a supervised manner, to provide more objective reference labels for 3DE. Furthermore, the use of CMR-derived labels provides an implicit advantage over manual 3DE segmentations in terms of reducing intermodality measurement bias.

Using 536 annotated 3DE images from a heterogeneous population of 134 human subjects comprising healthy controls and patients with cardiac disease, the dataset was used to train a self-configuring 3D U-Net to provide automated segmentations of the LV cavity and myocardium at ED and ES. This automated *nnU-Net* model subsequently outperformed an expert human observer in terms of accuracy against CMR reference values, as well as scan-rescan repeatability, whilst exhibiting increased measurement reliability (in terms of ICC) for all measured indices. Compared to volumes obtained using conventional manual analyses, *nnU-Net* had a lower magnitude of bias between 3DE and CMR, by 12 ml for EDV, and 10 ml for ESV. Most markedly, myocardial mass estimates using *nnU-Net* were far superior to those obtained by manual analyses. The automated method produced a bias that was seven times smaller in magnitude (5 g *nnU-Net* bias compared to 35 g manual bias in Table 3) for LVM, and excellent reliability with respect to CMR (where previously only moderate reliability was attained using the manual method). While there were indeed statistically significant differences in mean EDV and EF values between *nnU-Net* and CMR, these differences (i.e., 9 ml and 2%, for EDV and EF, respectively) are clinically acceptable (41), and unlikely to influence diagnostic outcomes or treatment pathways. In terms of segmentation accuracy, *nnU-Net* achieved a comparable Dice coefficient for the LV cavity with lower MSD and HD scores compared to the highest-ranking method trained and evaluated on the CETUS dataset. However, it should be noted that these comparisons are indicative only, as results were obtained from evaluation on a different dataset.

Signal dropout [particularly at the anterior wall (10)] remains a major challenge in 3DE analysis. Furthermore, highly anisotropic speckle properties and decreasing lateral resolution (being inversely proportional to transducer proximity) obscures the boundary between the myocardium and cavity toward the base of the LV when imaged from the apical window. By leveraging subject-specific geometries from CMR, our approach provides reliable reference annotations in such regions that are otherwise unavailable. Compared to the use of population priors, subject-specific information is more likely to produce labels closer to the true LV geometry for a given image instance, which may be leveraged by computational classifiers such as convolutional neural networks, despite not being apparent to human observers. Although this is possible in the presence of low contrast or poor resolution, it remains a challenge for the ML model to predict labels in regions where image data is entirely absent, such as that illustrated in Figure 8A. This

TABLE 3 **Left ventricular end-diastolic volume (EDV), end-systolic volume (ESV), mass (LVM), and ejection fraction (EF) for the testing set (*n* = 54 clips) derived from cardiac magnetic resonance (CMR), corresponding 3D echocardiography (3DE) measurement biases [mean ± standard deviation (SD)], and single measures intraclass correlation coefficients (ICC) with 95% confidence intervals in squared brackets.**

	CMR	*nnU-Net*		Expert (manual)		Comparison	
N = 54	Mean ± SD	Bias	ICC	Bias	ICC	*t*-test	*f*-test
EDV (ml)	153 ± 52	*−9 ± 16	0.936 [0.855, 0.968]	*−21 ± 19	0.864 [0.301, 0.953]	<**0.001**	0.189
ESV (ml)	66 ± 44	−1 ± 10	0.975 [0.957, 0.985]	*−11 ± 13	0.927 [0.680, 0.972]	<**0.001**	0.104
LVM (g)	127 ± 55	5 ± 23	0.897 [0.830, 0.939]	*35 ± 43	0.532 [0.110, 0.754]	<**0.001**	<**0.001**
EF (%)	60 ± 10	*−2 ± 5	0.889 [0.795, 0.938]	2 ± 6	0.825 [0.714, 0.895]	<**0.001**	0.069

Values in bold in the Comparison column represent statistically significant differences ($p < 0.05$) between the means (*t*-test) and variances (*f*-test) of measurement biases for the expert manual and *nnU-Net* analyses. Asterisks (*) indicate statistically significant differences between 3DE and CMR using a paired *t*-test.

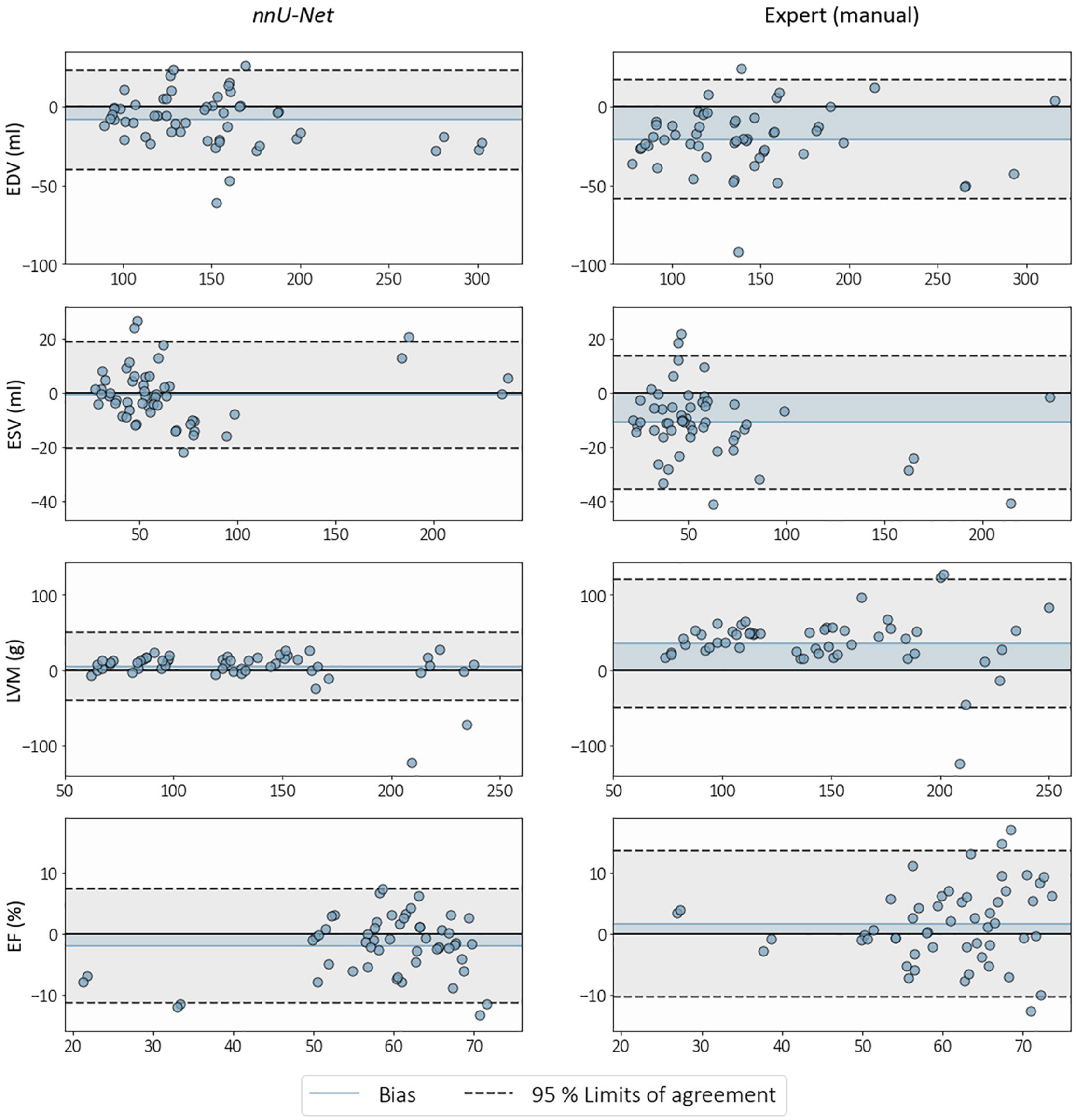


FIGURE 9
Bland-Altman plots showing biases and 95% limits of agreement between cardiac magnetic resonance (CMR) and 3D echocardiography (3DE) when analyzed by an expert and with *nnU-Net*. The horizontal axis represents the mean of measurements obtained from 3DE and CMR, against differences (calculated as 3DE–CMR) on the vertical axis, for end-diastolic volume (EDV), end-systolic volume (ESV), left ventricular mass (LVM), and ejection fraction (EF). Blue shaded regions represent the magnitude of bias from zero.

highlights the importance of image quality in terms of both texture as well as the selection of an appropriate pyramidal volume width during acquisition, the latter of which may result in a total lack of image information, and subsequent inability to recover geometric information.

The use of CMR-derived labels for 3DE relies on the assumption that there is no change in LV geometry (and associated hemodynamic status) between modalities. Although paired datasets were acquired with minimal time between CMR and 3DE scans, multimodal imaging was nevertheless performed asynchronously, with participants subject to natural physiological (e.g., heart rate) and positional (i.e., supine during CMR and lateral during 3DE) variability. Furthermore, different lung volumes during the breath-hold requirements for imaging may also influence venous return and consequently cardiac output (42). Thus, the assumption that LV volumes are identical for the same subject between scans consequently remains a limitation of the described method for the utilization of labels from a different modality. As the registration between CMR and 3DE only accounts for the rigid transformation component between imaging coordinate systems, it may be appropriate to incorporate affine components (such as scaling) to account for

TABLE 4 Scan-rescan variability in left ventricular end-diastolic volume (EDV), end-systolic volume (ESV), mass (LVM), and ejection fraction (EF) for the testing set ($n = 27$ patients) in terms of measurement biases (calculated as randomized first measurement−second measurement), average measures intraclass correlation coefficients (ICC) with 95% confidence intervals in squared brackets, and 95% confidence repeatability coefficients (RC), derived from expert manual analyses and *nnU-Net* segmentations.

	nnU-Net			Expert (manual)			Comparison
$n = 27$	Bias	ICC	RC	Bias	ICC	RC	*f*-test
EDV (ml)	1 ± 9	0.991 [0.980, 0.996]	±18	2 ± 17	0.968 [0.930, 0.985]	±34	**0.002**
ESV (ml)	−1 ± 5	0.997 [0.994, 0.999]	±9	1 ± 10	0.982 [0.962, 0.992]	±20	**<0.001**
LVM (g)	1 ± 12	0.984 [0.965, 0.993]	±24	1 ± 24	0.944 [0.877, 0.975]	±47	**0.001**
EF (%)	1 ± 3	0.987 [0.971, 0.994]	±5	−1 ± 4	0.957 [0.907, 0.980]	±9	**0.005**

Values in bold in the Comparison column represent statistically significant differences ($p < 0.05$) between the variances of the biases.

changes in LV geometry as a result of acquisition conditions. However, such changes are typically subtle for subjects at rest (43–45).

From a practical perspective, there are several advantages of using ML for 3DE analysis, including the reduction in the time required for analysis (with network inference time being approximately six seconds per 3DE image) and scan-rescan variability when compared with conventional methods, as exemplified in this study. The use of CMR in the creation of training data for automated 3DE analysis methods not only removes the measurement bias between the two modalities, but also provides more accurate and reproducible measurements (compared to manual analysis methods) to facilitate integration of 3DE into clinical practice. Lastly, the methodology surrounding the derivation of subject-specific labels from an alternative imaging modality is not limited to the LV, and similar approaches may be taken for other cardiac structures, such as the RV and cardiac atria, to enable more comprehensive examinations using 3DE.

4.1. Limitations and future work

While this work represents the largest publicly available 3DE dataset in terms of the number of labeled images, it currently stands as a single-center, single-vendor study (unlike CETUS, which includes data from three institutions and three ultrasound vendors). Similarly, reference geometries were obtained by a single observer, who performed both the CMR analysis [although interobserver variability is generally low (46)] as well as the manual refinement of CMR-to-3DE alignment. The reliance on a single observer consequently remains a limitation of this study, and further validation using an independent dataset is needed to assess the reproducibility of the label generation framework and overall robustness of the proposed method. Contributions from other institutions may also help to provide additional data variability to improve the generalizability and performance of the ML workflow presented here.

Although the use of 3D Cartesian images with isotropic spacing provides a standard format for input into most ML architectures, it is worth noting that in the case of 3DE, approximately two-thirds of the image consists of zero-values outside the pyramidal volume as a result of the rectangular bounding box. This redundancy warrants investigation into more efficient image representations and potential analysis on un-interpolated radiofrequency data, which may improve model performance.

As the present dataset is inclusive of ED and ES images only, this method may be extended to include intermediary frames and leverage temporal information (47) to enable automated full-cycle analysis. Such data would enable more in-depth analysis of cardiac motion or the assessment of diastolic function for added clinical value.

5. Conclusion

In light of the ongoing efforts in developing and evaluating automated 3DE analysis methods, we present here an annotated 3DE dataset comprising images of varying quality acquired across a range of patient demographics, representing the largest publicly available 3DE dataset to date, and the first of which leverages subject specific labels from CMR. Using this dataset, a state-of-the-art deep learning model applied to unseen 3DE images was capable of reproducing measurements derived from CMR, while outperforming an expert human observer in terms of accuracy and scan-rescan repeatability. As 3DE becomes increasingly widespread, the provision of a novel benchmark represents a critical step toward enabling the development of automated tools for enhanced efficiency and accuracy of non-invasive cardiac image analysis.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.cardiacatlas.org/.

Ethics statement

The studies involving human participants were reviewed and approved by the Health and Disability Ethics Committee (HDEC) of New Zealand (17/CEN/226). The patients/participants provided their written informed consent to participate in this study.

Author contributions

DZ devised the data generation framework, performed manual analyses and registrations, carried out the deep learning experiment, and drafted the initial manuscript. EF and GM advised on the deep learning aspects. GQ acquired the 3DE data and provided technical imaging expertise and manual segmentations. KG and VW established the pilot data acquisition protocol. TB advised on data sharing and management. JP and JD'h developed the BEAS algorithm. TS, BL, ML, PR, and RD contributed to patient recruitment and provided clinical expertise. OC advised on the interpretation and presentation of results. AY and MN designed the study, co-supervised the research, and assisted with interpretation of results. All authors have contributed significantly to the submitted work, including involvement in the research design, analysis and interpretation of data, and critical revision of the manuscript draft.

Funding

This study was funded by the Health Research Council of New Zealand (programme grant 17/608) and National Heart Foundation of New Zealand (project 1834).

Acknowledgments

The authors gratefully acknowledge the staff at the Centre for Advanced MRI at the University of Auckland for their expertise and assistance with the imaging components of this study, and Dr. Avan Suinesiaputra for facilitating the sharing of the MITEA dataset as part of the Cardiac Atlas Project.

Conflict of interest

MN was the CSO of HeartLab (NZ) Ltd. JD'h holds research contracts with GE Vingmed.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcvm.2022.1016703/full#supplementary-material

References

1. Martin-Isla C, Campello V, Izquierdo C, Raisi-Estabragh Z, Baeßler B, Petersen S, et al. Image-Based cardiac diagnosis with machine learning: A review. *Front Cardiovasc Med.* (2020) 7:1. doi: 10.3389/fcvm.2020.00001

2. Sanchez-Martinez S, Camara O, Piella G, Cikes M, González-Ballester M, Miron M, et al. Machine learning for clinical decision-making: Challenges and opportunities in cardiovascular imaging. *Front Cardiovasc Med.* (2021) 8:765693. doi: 10.3389/fcvm.2021.765693

3. Bai W, Sinclair M, Tarroni G, Oktay O, Rajchl M, Vaillant G, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *J Cardiovasc Magn Reson.* (2018) 20:65. doi: 10.1186/s12968-018-0471-x

4. Chen C, Bai W, Davies R, Bhuva A, Manisty C, Augusto J, et al. Improving the generalizability of convolutional neural network-based segmentation on CMR images. *Front Cardiovasc Med.* (2020) 7:105. doi: 10.3389/fcvm.2020.00105

5. Ruijsink B, Puyol-Antón E, Oksuz I, Sinclair M, Bai W, Schnabel J, et al. Fully automated, quality-controlled cardiac analysis from CMR. *JACC Cardiovasc Imaging.* (2020) 13:684–95. doi: 10.1016/j.jcmg.2019.05.030

6. Dorosz J, Lezotte D, Weitzenkamp D, Allen L, Salcedo E. Performance of 3-dimensional echocardiography in measuring left ventricular volumes and ejection fraction: A systematic review and meta-analysis. *J Am Coll Cardiol.* (2012) 59:1799–808. doi: 10.1016/j.jacc.2012.01.037

7. Wu V, Takeuchi M. Three-Dimensional echocardiography: Current status and real-life applications. *Acta Cardiol Sin.* (2017) 33:107–18. doi: 10.6515/acs20160818a

8. Baldea S, Velcea A, Rimbas R, Andronic A, Matei L, Calin S, et al. 3-D Echocardiography is feasible and more reproducible than 2-D echocardiography for in-training echocardiographers in follow-up of patients with heart failure with reduced ejection fraction. *Ultrasound Med Biol.* (2021) 47:499–510. doi: 10.1016/j.ultrasmedbio.2020.10.022

9. Medvedofsky D, Maffessanti F, Weinert L, Tehrani D, Narang A, Addetia K, et al. 2D and 3D echocardiography-derived indices of left ventricular function and shape: Relationship with mortality. *JACC Cardiovasc Imaging.* (2018) 11:1569–79. doi: 10.1016/j.jcmg.2017.08.023

10. Zhao D, Quill G, Gilbert K, Wang V, Houle H, Legget M, et al. Systematic comparison of left ventricular geometry between 3D-Echocardiography and

cardiac magnetic resonance imaging. *Front Cardiovasc Med.* (2021) 8:1099. doi: 10.3389/fcvm.2021.728205

11. Dong S, Luo G, Wang K, Cao S, Li Q, Zhang H. A combined fully convolutional networks and deformable model for automatic left ventricle segmentation based on 3D echocardiography. *Biomed Res Int.* (2018) 2018:1–16. doi: 10.1155/2018/5682365

12. Khellaf F, Leclerc S, Voorneveld J, Bandaru R, Bosch J, Bernard O. Left ventricle segmentation in 3D ultrasound by combining structured random forests with active shape models. In: Angelini E, Landman B editors. *Proceedings of the SPIE Medical Imaging*. Houston: (2018). 18 p. doi: 10.1117/12.2293544

13. Oktay O, Ferrante E, Kamnitsas K, Heinrich M, Bai W, Caballero J, et al. Anatomically Constrained Neural Networks (ACNNs): Application to cardiac image enhancement and segmentation. *IEEE Trans Med Imaging.* (2018) 37:384–95. doi: 10.1109/TMI.2017.2743464

14. Pedrosa J, Queirós S, Bernard O, Engvall J, Edvardsen T, Nagel E, et al. Fast and fully automatic left ventricular segmentation and tracking in echocardiography using shape-based b-spline explicit active surfaces. *IEEE Trans Med Imaging.* (2017) 36:2287–96. doi: 10.1109/TMI.2017.2734959

15. Bernard O, Bosch J, Heyde B, Alessandrini M, Barbosa D, Camarasu-Pop S, et al. Standardized evaluation system for left ventricular segmentation algorithms in 3D echocardiography. *IEEE Trans Med Imaging.* (2016) 35:967–77. doi: 10.1109/TMI.2015.2503890

16. Papachristidis A, Geleijnse M, Galli E, Heyde B, Alessandrini M, Barbosa D, et al. Clinical expert delineation of 3D left ventricular echocardiograms for the CETUS segmentation challenge. In *Proceedings of the MICCAI Chall Echocardiogr Three-Dimensional Ultrasound Segmentation*, Lyon (2014). p. 9–16.

17. Alessandrini M, De Craene M, Bernard O, Giffard-Roisin S, Allain P, Waechter-Stehle I, et al. A pipeline for the generation of realistic 3D synthetic echocardiographic sequences: Methodology and open-access database. *IEEE Trans Med Imaging.* (2015) 34:1436–51. doi: 10.1109/TMI.2015.2396632

18. Zhou Y, Giffard-Roisin S, Craene MD, Camarasu-Pop S, D'Hooge J, Alessandrini M, et al. A framework for the generation of realistic synthetic cardiac ultrasound and magnetic resonance imaging sequences from the same virtual patients. *IEEE Trans Med Imaging.* (2018) 37:741–54. doi: 10.1109/TMI.2017.2708159

19. Dong S, Luo G, Wang K, Cao S, Mercado A, Shmuilovich O, et al. VoxelAtlasGAN: 3D left ventricle segmentation on echocardiography with atlas guided generation and voxel-to-voxel discrimination. In: Frangi A, Schnabel J, Davatzikos C, Alberola-López C, Fichtinger G editors. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. MICCAI 2018. Lecture Notes in Computer Science*. Cham: Springer (2018). p. 622–9. doi: 10.1007/978-3-030-00937-3_71

20. Gilbert A, Marciniak M, Rodero C, Lamata P, Samset E, Mcleod K. Generating synthetic labeled data from existing anatomical models: An example with echocardiography segmentation. *IEEE Trans Med Imaging.* (2021) 40:2783–94. doi: 10.1109/TMI.2021.3051806

21. Chen C, Dou Q, Chen H, Qin J, Heng P. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *IEEE Trans Med Imaging.* (2020) 39:2494–505. doi: 10.1109/TMI.2020.2972701

22. Cui Z, Li C, Du Z, Chen N, Wei G, Chen R, et al. Structure-Driven unsupervised domain adaptation for cross-modality cardiac segmentation. *IEEE Trans Med Imaging.* (2021) 40:3604–16. doi: 10.1109/TMI.2021.3090432

23. Wu F, Zhuang X. Unsupervised domain adaptation with variational approximation for cardiac segmentation. *IEEE Trans Med Imaging.* (2021) 40:3555–67. doi: 10.1109/TMI.2021.3090412

24. Fonseca C, Backhaus M, Bluemke D, Britten R, Chung JD, Cowan B, et al. The cardiac atlas project–an imaging database for computational modeling and statistical atlases of the heart. *Bioinformatics.* (2011) 27:2288–95. doi: 10.1093/bioinformatics/btr360

25. Meyers B, Brindise M, Kutty S, Vlachos P. A method for direct estimation of left ventricular global longitudinal strain rate from echocardiograms. *Sci Rep.* (2022) 12:4008. doi: 10.1038/s41598-022-06878-1

26. Li B, Liu Y, Occleshaw C, Cowan B, Young A. In-line automated tracking for ventricular function with magnetic resonance imaging. *JACC Cardiovasc Imaging.* (2010) 3:860–6. doi: 10.1016/j.jcmg.2010.04.013

27. Young A, Cowan B, Thrupp S, Hedley W, Dell'Italia L. Left ventricular mass and volume: Fast calculation with guide-point modeling on MR images. *Radiology.* (2000) 216:597–602. doi: 10.1148/radiology.216.2.r00au14597

28. Pedrosa J, Heyde B, Heeren L, Engvall J, Zamorano J, Papachristidis A, et al. Automatic short axis orientation of the left ventricle in 3D ultrasound recordings. In: Duric N, Heyde B editors. *Proceedings of the SPIE 9790, Medical Imaging 2016: Ultrasonic Imaging and Tomography*. Vol. 9790 (2016). doi: 10.1117/12.2214106

29. Ahrens J, Geveci B, Law C. ParaView: An end-user tool for large-data visualization. *Vis Handb.* (2005):717–31. doi: 10.1016/B978-012387582-2/50038-1

30. Isensee F, Jaeger P, Kohl S, Petersen J, Maier-Hein K. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods.* (2021) 18:203–11. doi: 10.1038/s41592-020-01008-z

31. Sutskever I, Martens J, Dahl G, Learning G. On the importance of initialization and momentum in deep learning. In: Dasgupta S, McAllester D editors. *Proceedings of the 30th International Conference on Machine Learning*. Vol. 28 (2013). p. 1139–47. doi: 10.3390/brainsci10070427

32. Chen L, Papandreou G, Kokkinos I, Murphy K, Yuille A. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans Pattern Anal Mach Intell.* (2018) 40:834–48. doi: 10.1109/TPAMI.2017.2699184

33. Koo T, Li M. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med.* (2016) 15:155–63. doi: 10.1016/j.jcm.2016.02.012

34. Vaz S, Falkmer T, Passmore A, Parsons R, Andreou P. The case for using the repeatability coefficient when calculating test–retest reliability. *PLoS One.* (2013) 8:e73990. doi: 10.1371/journal.pone.0073990

35. Martin Bland J, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* (1986) 327:307–10. doi: 10.1016/S0140-6736(86)90837-8

36. Mosteller R. Simplified calculation of body-surface area. *N Engl J Med.* (1987) 317:1098–1098. doi: 10.1056/NEJM198710223171717

37. Barbosa D, Friboulet D, D'hooge J, Bernard O. Fast tracking of the left ventricle using global anatomical affine optical flow and local recursive block matching. *MIDAS J.* (2014) 10:17–24.

38. Queirós S, Vilaça J, Morais P, Fonseca J, D'hooge J, Barbosa D. Fast left ventricle tracking using localized anatomical affine optical flow. *Int J Numer Method Biomed Eng.* (2017) 33:doi: 10.1002/cnm.2871

39. Lang R, Badano L, Mor-Avi V, Afilalo J, Armstrong A, Ernande L, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: An update from the american society of echocardiography and the european association of cardiovascular imaging. *Eur Hear J Cardiovasc Imaging.* (2015) 16:233–71. doi: 10.1093/ehjci/jev014

40. Addetia K, Miyoshi T, Amuthan V, Citro R, Daimon M, Gutierrez Fajardo P, et al. Normal values of left ventricular size and function on three-dimensional echocardiography: Results of the world alliance societies of echocardiography study. *J Am Soc Echocardiogr.* (2022) 35:449–59. doi: 10.1016/j.echo.2021.12.004

41. Jenkins C, Moir S, Chan J, Rakhit D, Haluska B, Marwick T. Left ventricular volume measurement with echocardiography: a comparison of left ventricular opacification, three-dimensional echocardiography, or both with magnetic resonance imaging. *Eur Heart J.* (2009) 30:98–106. doi: 10.1093/eurheartj/ehn484

42. Lisanti C, Douglas D. Effects of breath-hold and cardiac cycle on the MRI appearance of the aorta and inferior vena cava in T2 abdominal imaging. *Am J Roentgenol.* (2009) 192:1348–58. doi: 10.2214/AJR.08.1646

43. Pump B, Talleruphuus U, Christensen N, Warberg J, Norsk P. Effects of supine, prone, and lateral positions on cardiovascular and renal variables in humans. *Am J Physiol Integr Comp Physiol.* (2002) 283:R174–80. doi: 10.1152/ajpregu.00619.2001

44. Ryan A, Larsen P, Galletly D. Comparison of heart rate variability in supine, and left and right lateral positions. *Anaesthesia.* (2003) 58:432–6. doi: 10.1046/j.1365-2044.2003.03145.x

45. Wieslander B, Ramos J, Ax M, Petersson J, Ugander M. Supine, prone, right and left gravitational effects on human pulmonary circulation. *J Cardiovasc Magn Reson.* (2019) 21:69. doi: 10.1186/s12968-019-0577-9

46. Bhuva A, Bai W, Lau C, Davies R, Ye Y, Bulluck H, et al. A multicenter, scan-rescan, human and machine learning CMR study to test generalizability and precision in imaging biomarker analysis. *Circ Cardiovasc Imaging.* (2019) 12:e009214. doi: 10.1161/CIRCIMAGING.119.009214

47. Painchaud N, Duchateau N, Bernard O, Jodoin P. Echocardiography segmentation with enforced temporal consistency. *IEEE Trans Med Imaging.* (2022) 41:1–1. doi: 10.1109/TMI.2022.3173669

frontiers | Frontiers in Cardiovascular Medicine

TYPE Original Research
PUBLISHED 10 March 2023
DOI 10.3389/fcvm.2023.985657

Echocardiography-based AI for detection and quantification of atrial septal defect

Xixiang Lin[1,2†], Feifei Yang[1†], Yixin Chen[3†], Xu Chen[1,2], Wenjun Wang[1], Wenxiu Li[4], Qiushuang Wang[5], Liwei Zhang[5], Xin Li[6], Yujiao Deng[1], Haitao Pu[3], Xiaotian Chen[3], Xiao Wang[1,2], Dong Luo[1,2], Peifang Zhang[3], Daniel Burkhoff[7] and Kunlun He[1]*

[1]Medical Big Data Center, Chinese PLA General Hospital, Beijing, China, [2]Medical School of Chinese PLA, Beijing, China, [3]BioMind Technology, Beijing, China, [4]Department of Pediatric Cardiac Center, Beijing Anzhen Hospital, Capital Medical University, Beijing, China, [5]Department of Cardiology, The Fourth Medical Center of Chinese PLA General Hospital, Beijing, China, [6]Department of Ultrasonography, The Sixth Medical Center of Chinese PLA General Hospital, Beijing, China, [7]Cardiovascular Research Foundation, New York, NY, United States

Objectives: We developed and tested a deep learning (DL) framework applicable to color Doppler echocardiography for automatic detection and quantification of atrial septal defects (ASDs).

Background: Color Doppler echocardiography is the most commonly used non-invasive imaging tool for detection of ASDs. While prior studies have used DL to detect the presence of ASDs from standard 2D echocardiographic views, no study has yet reported automatic interpretation of color Doppler videos for detection and quantification of ASD.

Methods: A total of 821 examinations from two tertiary care hospitals were collected as the training and external testing dataset. We developed DL models to automatically process color Doppler echocardiograms, including view selection, ASD detection and identification of the endpoints of the atrial septum and of the defect to quantify the size of defect and the residual rim.

Results: The view selection model achieved an average accuracy of 99% in identifying four standard views required for evaluating ASD. In the external testing dataset, the ASD detection model achieved an area under the curve (AUC) of 0.92 with 88% sensitivity and 89% specificity. The final model automatically measured the size of defect and residual rim, with the mean biases of 1.9 mm and 2.2 mm, respectively.

Conclusion: We demonstrated the feasibility of using a deep learning model for automated detection and quantification of ASD from color Doppler echocardiography. This model has the potential to improve the accuracy and efficiency of using color Doppler in clinical practice for screening and quantification of ASDs, that are required for clinical decision making.

KEYWORDS

artificial intelligence, deep learning, echocardiography, atrial septal defects, congenital heart disease

OPEN ACCESS

EDITED BY
Steffen Erhard Petersen,
Queen Mary University of London,
United Kingdom

REVIEWED BY
Duraisamy Balaguru,
Harvard Medical School, United States
Alison Pouch,
University of Pennsylvania, United States

SPECIALTY SECTION
This article was submitted to Cardiovascular Imaging, a section of the journal Frontiers in Cardiovascular Medicine

*CORRESPONDENCE
Kunlun He
✉ kunlunhe@plagh.org

†These authors have contributed equally to this work

RECEIVED 04 July 2022
ACCEPTED 21 February 2023
PUBLISHED 10 March 2023

CITATION
Lin X, Yang F, Chen Y, Chen X, Wang W, Li W, Wang Q, Zhang L, Li X, Deng Y, Pu H, Chen X, Wang X, Luo D, Zhang P, Burkhoff D and He K (2023) Echocardiography-based AI for detection and quantification of atrial septal defect.
Front. Cardiovasc. Med. 10:985657.
doi: 10.3389/fcvm.2023.985657

Abbreviations
ASD, Atrial septal defect; AUC, Area under the curve; A4C, Modified apical four-chamber view; CDI, Color Doppler imaging; CHD, Congenital heart disease; CNN, Convolutional neural networks; DL, Deep learning; PSAX, Parasternal short-axis view; ROC, Receiver operating characteristic; SC2A, Subxiphoid sagittal view; SC4C, Subxiphoid four-chamber view; TTE, Transthoracic echocardiography.

Introduction

It is estimated that in 2017, nearly 1.8 cases per 100 live births are diagnosed with congenital heart disease (CHD) worldwide (1). Atrial septal defect (ASD) is the second most common type of CHD, accounting for approximately 6%–10% of cases (2). Most patients with ASD are asymptomatic and may be identified as an incidental finding during routine echocardiographic examinations. Early detection of appropriately sized defects known to lead to problems later in life can prompt timely intervention and improve cardiovascular outcomes, avoiding substantial disability and mortality (3, 4).

Transthoracic echocardiography (TTE) with Doppler flow imaging is currently the most widely used noninvasive tool for detecting the presence of an ASD, especially in children (5). TTE cannot only be used to detect and quantify the size and shape of the septal defect, but can also be used to measure the degree and direction of shunting, changes of the size and function of the cardiac chambers and detect abnormal pressures and flows through the pulmonary circulation (6). However, accurate detection and quantification of ASD features relies on experienced, highly trained physicians which are in short supply, especially in rural areas (7). Furthermore, the low prevalence of disease and variability of image quality, number of acquired views and interpretation of TTE images causes low sensitivity and specificity of ASD detection (4), all of which hinder referral for treatment. Therefore, an effective solution for efficient, accurate and objective detection and grading of ASDs is critically needed.

Deep learning (DL) models have been applied for automated detection and assessment of cardiovascular diseases based on echocardiographic images and videos. Such models can complete a variety of tasks such as, image quality assessment, view classification, boundary segmentation, disease diagnosis and automatic quantification (8–12). However, there is no prior study investigating the effectiveness of a DL model for detecting ASD based on color Doppler images. Accordingly, we developed and validated a DL model for automated detection and quantification of ASDs (**Figure 1**).

Methods

Study population

This study involved algorithm development and initial testing based on a retrospective data set, and final testing from a prospective, real-world data set of consecutively acquired echocardiographic studies. 396 TTE examinations obtained between July 2020 and April 2021 from Anzhen hospital served as our training dataset. A total of 425 consecutively obtained examinations between May 2020 and Dec 2020 from the Chinese PLA General Hospital were collected as the external testing set, which including 48 ASD cases and 377 cases without ASD. The age of all cases in both training and testing datasets was less than 18. ASD diagnostic criteria were based on the 2015 ASE guideline (6), as detailed below. The ground truth for the presences of an ASD was based on the diagnosis present in the electronic medical record and echocardiographic clinical report which were provided by experienced echocardiography readers and reviewed by cardiologists who authorized the final reports. Other hemodynamically significant cardiac lesions (such as tetralogy of fallot and valvular heart disease) were excluded.

Echocardiography

Each echocardiographic study was acquired through standard methods. Four standard views were suggested by ASE guideline for detection and quantification of ASDs (6): (1) the modified apical four-chamber view (A4C); (2) the parasternal short-axis view (PSAX); (3) the subxiphoid sagittal view (SC2A); and (4) the subxiphoid four-chamber view (SC4C). These images were acquired from a diverse array of echocardiography machine manufacturers and models including Phillips iE-elite and 7C with transducer S5-1 and X5-1 (Phillips, Andover, MA, USA), Vivid E95 (General Electric, Fairfield, CT, USA), Mindray M9cv with transducer SP5-1s (Mindray, Shenzhen, Guangdong, China), Siemens SC2000 with transducer 4V1c (Siemens, Munich, Germany). All images were downloaded and stored with a standard Digital Imaging and Communication in Medicine (DICOM) format according to the instructions from each manufacturer.

View selection

We labeled 3,404 images to develop a method to classify 29 standard views, and then selected the 4 views required for detection and quantification of ASD detailed above. View selection was performed using a Xception Net neural network model according to methods that were similar to those described previously (8, 10, 13).

Segmentation

We selected 792 videos inclusive of the four standard color Doppler views required for segmentation from among the ASD cases. However, not every case had all four standard views, restricted by the limitation of retrospective data and the improper body position during examination. The atrial septum and margins of the defects were annotated with the LabelMe (**Figure 2**). In the modified apical four-chamber and subxiphoid four-chamber views, we labelled the atrial septum from atrioventricular valve to the roof of the atria (boundaries indicated by the dots in **Figure 2**). In the parasternal short-axis view, we labelled the atrial septum from aortic adventitia to the roof of the atria. In the subxiphoid sagittal view, we labelled the atrial septum from the bottom to the roof of the atria. We labelled the defect based on the width of the shunt jet detected

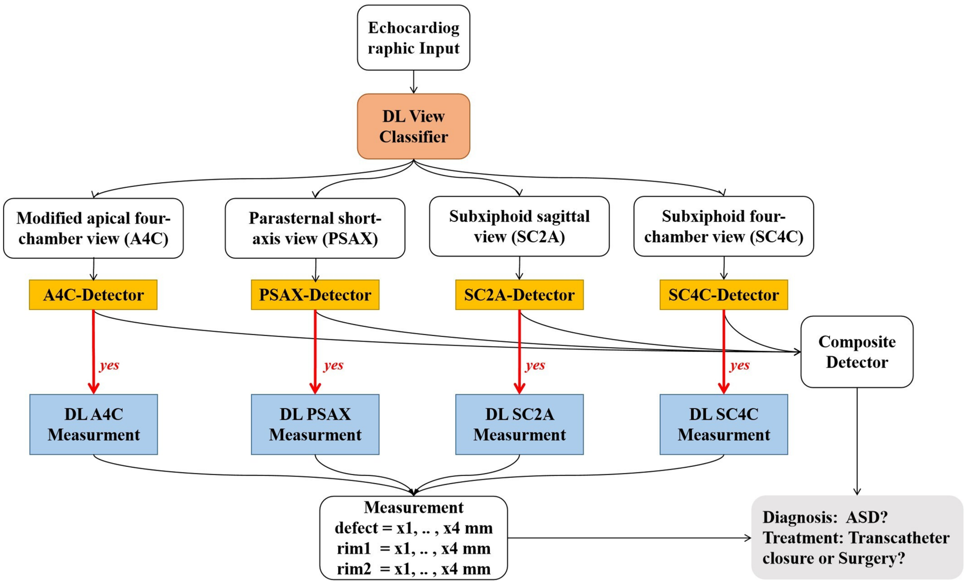


FIGURE 1
Work flow of the ensemble model. Step 1: raw echocardiographic videos are separated for classification of views (red box). Step 2: disease detection models use different views to detect the presence of ASD (orange rectangles). Step 3: if ASD is present (denoted by "yes"), metrics associated with severity of ASD are assessed (blue rectangles). DL, deep learning; ASD, atrial septal defect.

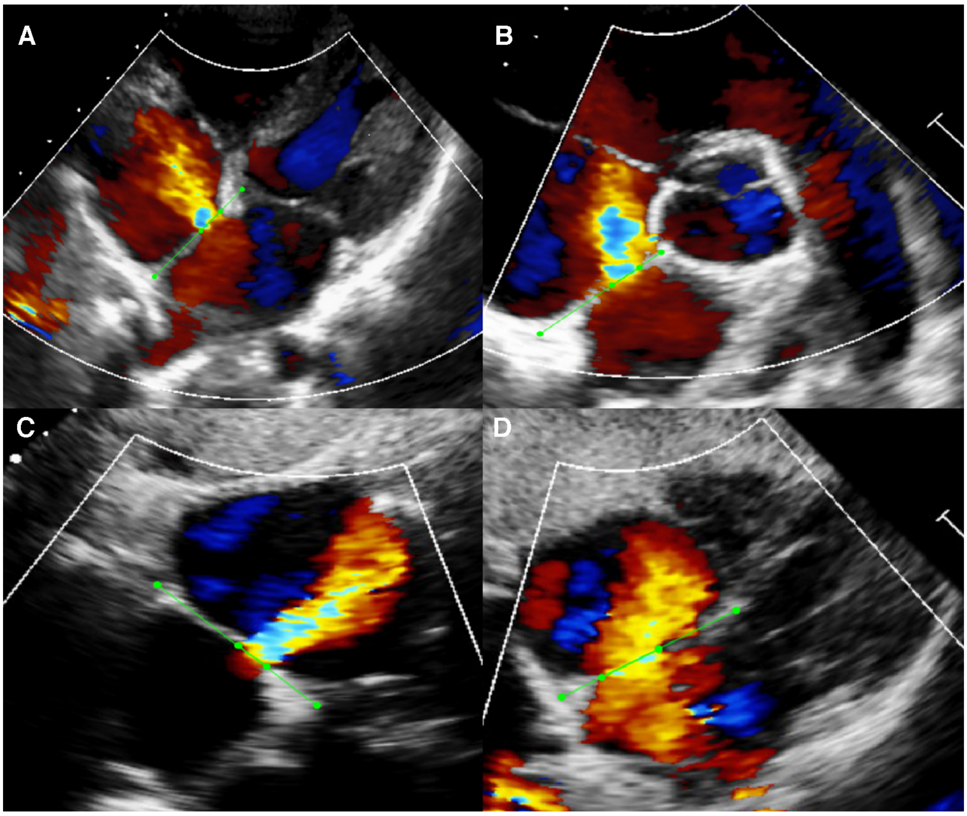


FIGURE 2
Example of manual segmentation. Green dots were manually labeled as the endpoints of atrial septum and defect with the open-source program LabelMe. (A) Modified apical four-chamber view (A4C). (B) Parasternal short-axis view (PSAX). (C) Subxiphoid sagittal view (SC2A). (D) Subxiphoid four-chamber view (SC4C).

on color Doppler flow images and the anechoic area of atrial septum in each view.

Detection of atrial septal defect

For the ASD detection task, videos were labelled as either ASD or normal based on the electronic medical record and echocardiographic clinical report. Each frame was resized to 240 × 320 pixels from 600 × 800 DICOM-formatted images. The pixel value was normalized to between 0 and 1. No clipping or interpolation operations were performed on frame numbers; therefore, the number of frames used for the analysis differed from video to video. To effectively increase the number of videos for training, we employed affine transformations including RandomShift (10%), RandomScale (10%) and RandomRotation (20°). The batch size was set to 1 because of the difference of frame number. Finally, we adopted the Adam optimizer with a weight decay of 1e-5. The learning rate was set to 3e-5. All models were trained on an Nvidia Tesla P100 GPU.

The ASD detection network architecture was shown in Figure 3A. The model was based on the ResNet architecture with modifications (14). First, we used a frame-based max pool to fuse blood flow information in each frame. Second, we used Atrous Spatial Pyramid Pooling (ASPP) to increase the visual field of the convolution feature extractor. ASPP consisted of a global average pool layer and four convolution layers with dilation coefficients of 1, 4, 8 and 12 respectively. Third, we used GroupNorm to replace BatchNorm since the batch size was 1. The loss function was binary cross-entropy. Finally, the model could provide the ASD probability of each frame in the video. Therefore, the frame with the highest probability would be selected as the keyframe of model diagnosis. We have made our code available at GitHub (15).

Quantification of atrial septal defect

For the quantification tasks, each frame was labelled with 4 points, two of which were the edges of the ASD, and the other two were the ends of the septum. Each frame was resized to 240 × 320 pixels. The pixel value was normalized to between 0 and 1. For the training stage, we randomly clipped 16 consecutive frames as an input from the video. Because of the large number of training epochs, we believed that the model had fully learned the entire information for each training sample. For validation and testing stage, each video was clipped into multiple segments. According to the majority voting principle, we made the prediction from those of segments. Each prediction took into account all the video information and did not receive the impact of randomness. We adopted the same affine transformation described above. In this case, the batch size was set to 2. We adopted Adam optimizer with a weight decay of 1e-5. The learning rate was set to 3e-5. All models were trained on an Nvidia Tesla P100 GPU.

The quantification network architecture was shown in Figure 3B. The networks architecture adopted UNet-style design (16). There were two dilemmas in using deep learning to fulfil the ASD quantification task.

Firstly, measuring the length of ASD was defined as a segmentation task, so we also adopted the structure of 3D-UNET (16). However, the performance was far less than expected. Segmenting the area of atrial septal defect was not a routine segmentation task. The common segmentation task assumes a segmentation boundary in the image, but the atrial septal defect was a disappearing region. Experienced doctors need to annotate the region based on sequent images. Therefore, we defined the task of quantifying ASD as key point detection since each key point exists in the image. We continued 3D-Unet-style and applied it to the task of key point detection.

The second point was how to make the neural network perform similarly to the senior physician. We have made two improvements: we have added different scales of auxiliary loss so that the model was from coarse to fine for key point identification; we found that the temporal convolution was unsuitable for the ASD key point detection task in echocardiography and therefore replaced it with a temporal transformer to overcome the long-range dependence dilemma of the frame dimension. The dependency dilemma can be described as follows: the key point detection of unclear video frames depends on the information obtained from the previous and next video frames. Such unclear video frames were present in human pose detection (17), for example, in the form of occlusion or overlap. However, this occlusion or overlap exists for a short period of time, usually no more than 3 consecutive frames. In echocardiography, most of the frames were not clear enough. Therefore, an experienced sonographer will prioritize the key points in clear video frames and then identify key points in other frames that were not clear enough. Convolution was a natural local attention mechanism and was not global, so using convolution to extract features will suffer long-range forgetting dilemma. As a simple example, suppose we use a $3 \times 3 \times 3$ convolution kernel for 3D convolution and we want the features of frame i to fuse the features of frame j. If $|i - j| \leq 1$, then the features of i and j were ready for fusion in the 1st convolutional layer. If $|i - j| \leq 3$, then the features of i and j need to go to the 3rd convolutional layer before they can be fused with each other. If the distance between frame i and frame j was long, the long-range forgetting dilemma will occur. In order to allow the features of clear video frames to be efficiently propagated throughout the whole video, we use the temporal transformer module. The structure of temporal transformer was shown in Figure 3C. Firstly, CNN-based extractor extracted the feature from each frame. In each temporal transformer module, the feature was transferred to three parts, query (q), key (k) and value (v). We calculated the correlation between q and k, which was called self-attention. We had 16 frames (tokens) so that the correlation map was 16×16. This map represented the correlation between any two frames. Finally, we used the correlation map processed by softmax as the weight, and sum the value. For the whole model, we only downsampled in the spatial

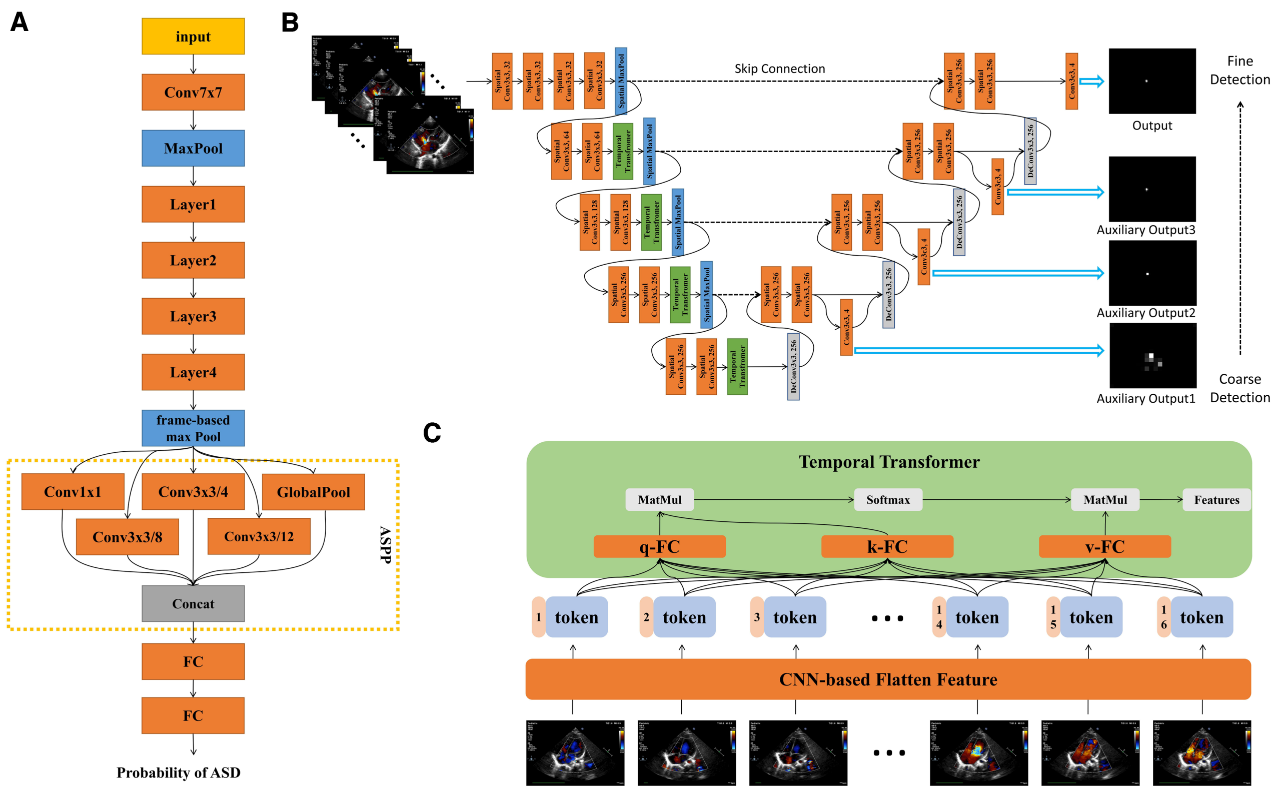


FIGURE 3
Detailed description of model architecture. (A) ASD detection neural network architecture, based on Resnet-18, combined with ASPP to increase visual field. The output of the model is the probability of ASD between 0 and 1. (B) ASD quantification neural network architecture, combining spatial convolution and temporal transformer, had three auxiliary outputs and one output from coarse to fine. Each auxiliary output is added to the feature map as an additional feature. (C) Structure diagram of temproal transformer. Each frame is encoded into a feature token by CNN. Self-attention could realize the long-distance dependence between frames. *q*, query; *k*, key; *v*, value.

dimension, and did not downsample in the temporal dimension so that for each temporal transformer block, the token number was always 16.

In addition, we have made the following adjustments: considering the cost of computation and time, we used the 2D spatial convolution and temporal transformer to replace the 3D convolution. Second, we used GroupNorm to replace BatchNorm since the batch size was small. Video $x \in \mathbb{R}^{W \times H \times F}$ (W means width, H means height and F means frame) goes through convolutional layers for feature extraction in W and H spatial dimensions. Then the W and H dimensions were merged into token T so feature map can be writen as $f \in \mathbb{R}^{W' \times H' \times F}$ or $f \in \mathbb{R}^{T \times F}$. The token of each frame was spliced with the corresponding position code, and then can be used as the input of the temporal transformer. The self-attention mechanism follows the design of ViT (18). The model can be divided into 4 stages, each containing a spatial downsampling layer, spatial convolution layers and a temporal transformer module.

The model provided an index of the "confidence" with which the septal length was estimated. Confidence was calculated as the percent of "stable frames" contained in the entire video. A frame was designated as "stable" if the absolute difference of the AI-predicted septal length from that of the prior frame divided by the average length of the 2 frames was less than 0.5. The model only calculated defect size and septal length based on the stable frames. Specifically, septal length was calculated as the average value of the lengths on all stable frames. ASD defect size was calculated the largest value among all stable frames. Accuracy of measurements of atrial septal lengths and defect sizes were compared to those made by expert echocardiographers using Bland & Altman analysis.

Statistical analysis

Analyses were performed using algorithms written in Python 3.6 from the libraries of Numpy, Pandas, and Scikit-learn. Continuous variables were expressed as mean ± standard deviation, median and interquartile range, or counts and percentage, as appropriate. Comparisons of reports and machine algorithm performances were performed using one-way analysis of variance (ANOVA), followed by the least significant difference (LSD) *t*-test. Results were regarded as statistically significant when $P < 0.05$. The models were assessed according to the area under the receiver operating characteristic (ROC) curves which plotted sensitivity vs. 1-specificity derived from the model's prediction confidence score. All calculations were performed by using IBM SPSS version 23.0.

Results

Characteristics of study population

A total of 821 patients with transthoracic echocardiographic examinations were included. The clinical and echocardiographic characteristics of included cases were summarized in Table 1. In the training dataset, patients with ASD had a median age of 3 years (IQR: 1, 10), 34.3% were male, and EF had a mean value of 70.0 ± 5.0. In the external testing dataset, patients with ASD had a median age of 1 years (IQR: 0, 9), 52.0% were male, and ejection fraction (EF) had a mean value of 64.7 ± 5.3.

Model for view selection

As summarized in Supplementary Figure S1, the deep-learning model identified four standard color Doppler views with an average accuracy of 0.99, including apical four-chamber view (0.97), parasternal short-axis view (0.99), subxiphoid frontal view (0.99) and subxiphoid sagittal view (1.0).

Model for detection of atrial septal defect

For each echo-Doppler video, the ASD detection model provided a probability level for the presence of an ASD; the frame with the highest probability was tagged as the keyframe of the video (examples shown in Supplementary Figure S2). The ROCs for the detection of an ASD in each of the 4 views for the external validation dataset were shown in Figure 4. The AUROC for ASD detection ranged from 0.901 to 0.956 for the individual views. The final diagnosis was made by the composite classifier model, which had an AUROC = 0.92. Youden's Index was used to evaluate model performance, which yielded sensitivities of 87.8% and specificities of 89.4% (Figure 4 and Table 2).

Model for quantification of atrial septal defect

Examples of segmentation model outputs were shown in the still image of Figure 5. As shown, the blue dots show where the DL model identified the ends of the atrial septum, while the orange dots show the model-identified edges of the ASD. Examples of frame-by-frame segmentation throughout entire videos, along with the model-derived measurements of the defect size and rim lengths compared to those measured by the expert physicians were shown in the videos provided in with online supplement material. These videos show results obtained from different echocardiographic views and different image qualities.

As detailed in Methods, the model provided an index of the "confidence" with which the septal length was estimated. Examples of results with different confidence levels were shown in Figure 6. The quantification model had greater performance in videos with higher confidence values; the relationships between the absolute difference between AI- and expert-determined septal length and ASD lengths as a function of the confidence values were shown in Supplementary Figure S3.

Results of the Bland & Altman analysis comparing values provided by the AI algorithm and experts' measurements were summarized in Figure 7. The mean bias for the measurement of defect size and septum length were 1.9 and 2.2 mm. We also recruited three experts to measure defect size and septum length in the test dataset. As shown in Supplementary Figure S4, the mean biases of defect size were respectively 1.5 mm, 2.3 mm, 0.3 mm, and the mean biases of septum length were respectively 0.8 mm, 2.1 mm, 1.2 mm. Despite the fact that inter-expert variability was lower than the AI model bias, the difference was insignificant. Therefore, we believed that the bias of algorithm is comparable to that encountered in current clinical practice.

Applying these automatic measurements to the indications and contraindications detailed in the 2015 ASE guidelines, we used the model to predict whether a given patient should be referred for transcatheter intervention (6). The results of the prediction were

TABLE 1 Baseline characteristics of the training and testing dataset.

	Training dataset		Testing dataset	
	ASD	Control	ASD	Control
N	198	198	48	377
Age (years)	3 (1,10)	3 (1,6)	1 (0,9)	5 (1,13)
Male patients (%)	68 (34.3)	105 (53.0)	25 (52.0)	238 (63.1)
Height (cm)	112.6 ± 34.8	107.4 ± 28.3	88.6 ± 35.5	125.2 ± 35.2
Weight (kg)	26.9 ± 21.3	22.2 ± 14.7	16.5 ± 20.5	30.0 ± 22.5
Echo parameters				
LV EF (%)	70.0 ± 5.0	70.0 ± 5.3	64.7 ± 5.3	65.6 ± 3.5
LV EDD (mm)	30.8 ± 7.4	34.6 ± 6.8	23.4 ± 9.8	33.4 ± 8.4
LV ESD (mm)	19.1 ± 6.5	21.0 ± 4.6	14.5 ± 6.9	20.7 ± 5.8
LA AD (mm)	21.8 ± 7.3	21.0 ± 5.0	16.4 ± 7.2	22.1 ± 5.6
E/A	1.7 ± 1.5	1.7 ± 0.5	1.6 ± 0.5	1.7 ± 0.5

ASD, atrial septal defect; LVEF, left ventricular ejection fraction; LV EDD, left ventricular end-diastolic dimension; LV ESD, left ventricular end-systolic dimension; LA AD, left atrial anteroposterior dimension.

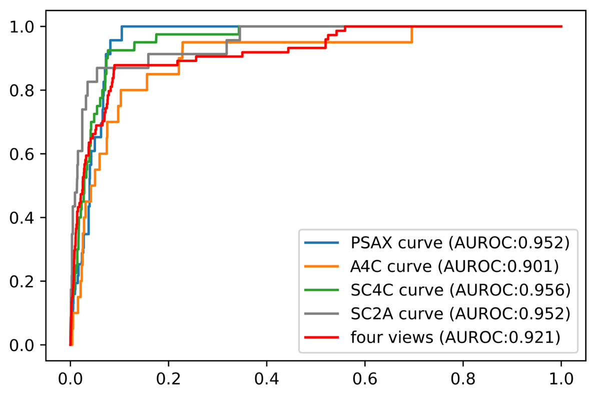


FIGURE 4
The performance of ASD detection model in the external dataset. The performance of composite classifier model (red curve) had an AUROC of 0.92. Abbreviations as in Figure 1.

TABLE 2 Model performance for ASD detection in different views.

	AUC	Sensitivity	Specificity
Composite	0.921	87.8%	89.4%
A4C	0.901	85.0%	84.4%
PSAX	0.952	95.7%	91.9%
SC2A	0.952	91.3%	83.8%
SC4C	0.956	92.5%	92.3%

Abbreviations as in Figure 1.

compared with the recommendations provided by an expert physician, who applied his own manual measurements to the guideline recommendations. The accuracy of model to predict the expert's conclusion was 85.4% (**Supplementary Table S1**).

Discussion

Echocardiography is the primary method for confirming the diagnosis of an ASD, for defining its anatomic and physiological characteristics and for deciding upon the need for and approach to treatment. However, accurate interpretation of echocardiograms for each of these purposes is in many respects subjective and time-consuming, requiring highly skilled clinicians which are not readily available in all hospitals. With the advantages of objectivity, efficiency, accuracy and consistency, deep learning (DL) models have been shown to be helpful in interpreting medical images in many fields of medicine (19–21), including echocardiography (8–12). However, ours is the first study to employed DL model for accurate detection and quantification of ASD through automated interpretation of color Doppler videos.

As in most DL models applied to echocardiography, the first step in our pipeline was echocardiographic view classification (22–24). However, newly introduced in our study is a classification model that includes color Doppler views. This model automatically selected the guideline-recommended echocardiographic views required for the detection and quantification of ASD with a high degree of accuracy.

The next step was implementation of a DL model to detect the presence of an ASD based on interpretation of the color Doppler views. This model also proved to be accurate, with high levels of sensitivity and specificity for disease detection. Similar degrees of detection accuracy were reproduced in all four echo-Doppler views examined and the AUC of the composite classifier model reached to 0.92. In addition, to address the "black box" problem and improve the interpretability, our model also automatically identified the key frame which can be provided to the clinician as a reference for final diagnosis and manual verification. Accordingly, the model has the potential to be used as a screening tool to aid doctors in identifying patients with an ASD, particularly in geographies where access to expert clinicians is limited.

Following view selection and disease detection, the final step was automated quantification of ASD size and the length of the residual rim; these are critical for determining the need for, and choice of treatment: transcatheter intervention or cardiothoracic surgery. To make these measurements, the quantification model automatically located the endpoints of the atrial septum and of the defect. In order to ensure the stability and reliability of automated quantification, the model generated an index of "confidence" with which the septal length was estimated. Naturally, the quantification model had greater performance in videos with higher confidence values. The performance of the algorithm was assessed by the bias of measurement of defect size and septum length, which provided a quantitative index of the degree of concordance between the DL model and expert physicians. Values of bias achieved by the model were low. Because the model explicitly detected and displayed the location of endpoints of the septum and defect, physicians can readily

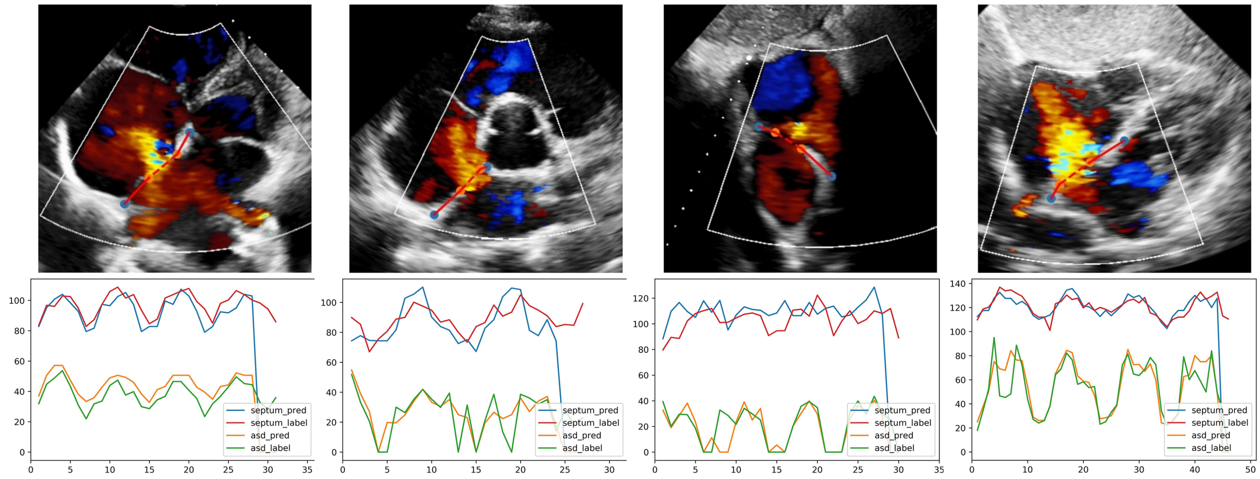


FIGURE 5
The output of quantification model in different views. The first row, one example test image of patients with ASD is shown respectively in A4C, PSAX, SC2A and SC4C view; orange and blue dots are the endpoints of defect and septum predicted by DL model. The second row, corresponding curves showing the variation of defect and septum in the video. Abbreviations as in Figure 1.

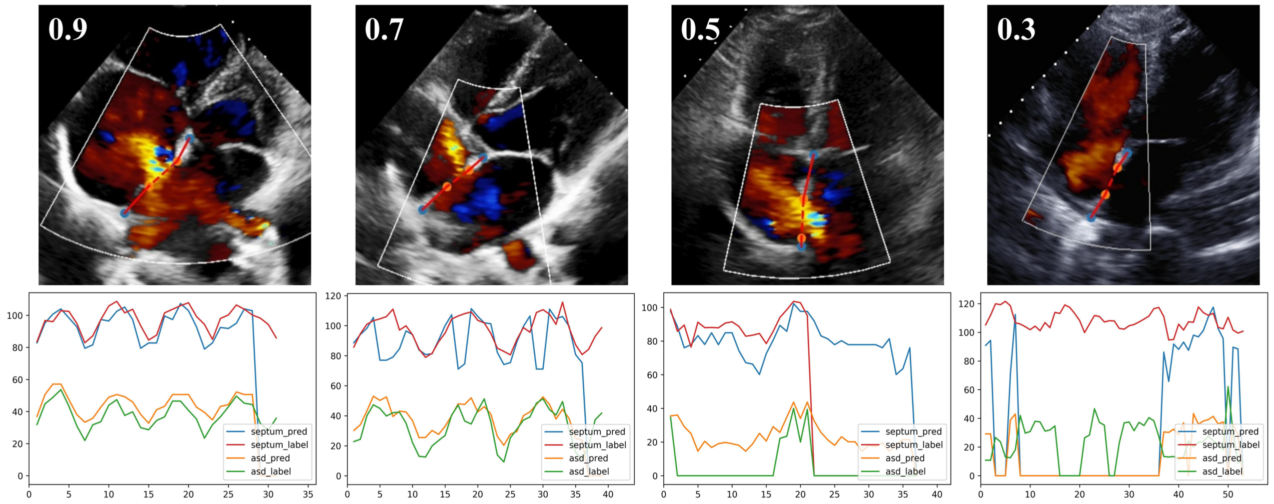


FIGURE 6
The output of quantification model with different confidence levels. The first row, one example test image of A4C view is shown respectively with the confidence of 0.9, 0.7, 0.5 and 0.3. The second row, corresponding curves showing the variation of defect and septum in the video.

verify the accuracy of the DL algorithm on a case-by-case basis. All these features are illustrated in the videos provided on the online supplemental material.

Finally, we note that whereas the metrics of defect and rim size are helpful for deciding between use of a transcatheter or surgical intervention, such decisions are not made based on these metrics alone. According to society guidelines (6) such decisions should be made based on additional metrics and other imaging approaches, such as transesophageal echocardiography and three-dimensional imaging. While this tool has potential utility in areas where access to expert physicians is limited, the method was trained and validated on images acquired by experts. So its translation to resource-limited environments might require additional adaptations to give real-time feedback on image quality if datasets are acquired by individuals with limited specialization in echocardiography of congenital defects.

Related work

Recent studies have shown remarkable performance of deep learning models in diagnosing ASDs (25–28). Wang et al. proposed an end-to-end framework which automatically analyzed multi-view echocardiograms and selected keyframes for disease diagnosis. As a result, the framework differentiated ASD, VSD and normal cases with an accuracy of 92.1% (25). Rima et al. used fetal screening ultrasound to train a DL model for these tasks, including view selection, segmentation and complex congenital heart disease detection. In the test of 4,108 fetal sonograms, the model achieved an AUC of 0.99 in distinguishing normal from abnormal hearts, which was comparable to expert clinicians' performance (26). Zhao et al. developed a variant of U-Net architecture to segment the structure of the atrial septum in magnetic resonance images of pre- and post-occlusion ASD

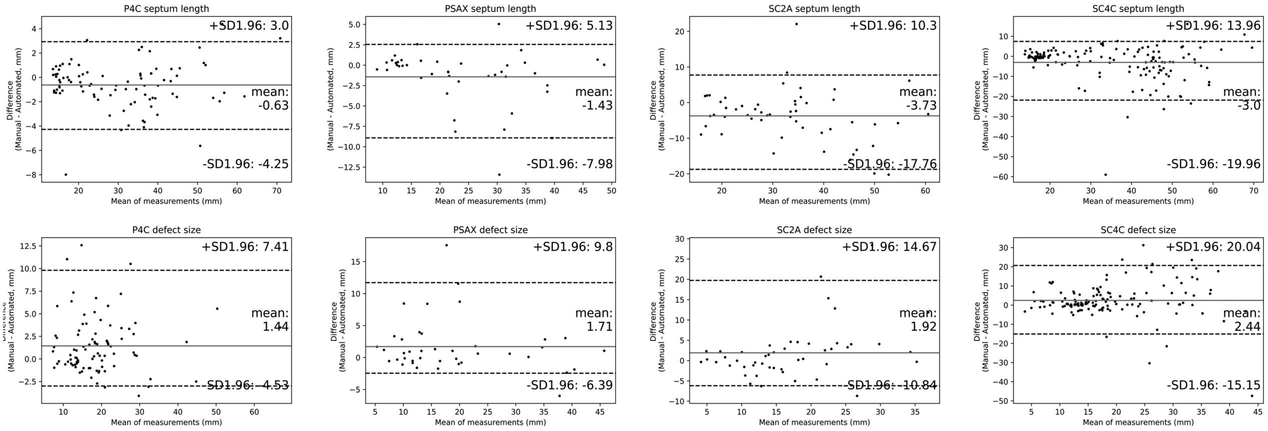


FIGURE 7
Comparisons of quantitative metrics derived from the deep learning (DL) algorithm and physician based on bland and altman analysis. Bland-Altman plots compare automated and manual measurements for septum length and defect size in A4C, PSAX, SC2A and SC4C view. Abbreviations as in Figure 1.

patients, with mean Dice index of 0.81 (27). Mori et al. proposed a DL model that used electrocardiograms (ECGs) to detect the presence of ASDs. This model outperformed 12 pediatric cardiologists in diagnosing ASD from ECG interpretations, with an accuracy of 0.89 (28). However, ours is the first study to detect the presence of an ASD based on multiple color Doppler views and to automatically identify the margins of the atrial septum and the margins of the ASD in order to provide quantitative measurements of ASD size and rim size. These represent significant advances since quantification of these anatomic features of an ASD are critical for determining treatment. Specifically, according to the 2015 ASE guideline (6), echocardiography provides important information for deciding on whether or not to treat an ASD and whether the defect is most suitably treated by transcatheter or surgical techniques. In this regard, studies have shown that ASD diameter measured directly at surgery is most accurately estimated by color flow Doppler echocardiography, while significant errors can arise if measurements are estimated from standard 2D echocardiograms alone (29).

Limitations

The results of our study need to be considered within the context of several limitations. First, all of the images were acquired by transthoracic echocardiography (TTE) rather than transesophageal echocardiography (TEE), as most of the included population were children who cannot tolerate TEE examination. Additionally, patients did not undergo cardiac computed tomography or magnetic resonance imaging, which can provide more detail information of ASD anatomy. Second, the training and testing dataset is based on images obtained from children. Despite the low prevalence, the algorithm performed very well to identify and quantify the sizes of these ASDs. This indicates that the absolute size of the heart does not influence accuracy of the model since the images are ultimately scaled to the same pixel dimensions with adequate special resolution. Third, limited by the retrospective nature, the study included a relatively small number of patients. Although our model achieved good performance in the external test set, testing of the model in a prospective multi-center cohort is warranted. Finally, the “black box” problem of our DL algorithm poses an inherent impediment to acceptance into clinical practice because of the opaqueness on how diagnoses are made. To overcome this limitation, we implemented an algorithm which provided keyframe selected by the DL model and identified the endpoints of defect and the septum on the images. This is intended to promote physician confidence in the model-based diagnoses and measurements. Even then, it is emphasized that the algorithm is intended to assistant, not replace, physician decision making.

Conclusion

We developed and validated a novel deep learning model applicable to color Doppler echocardiography for automatic detection and quantification of atrial septal defect and rim sizes. This model has the potential to improve the accuracy and efficiency of color Doppler echocardiographic screening and quantification of ASDs.

Perspectives

Competency in patient care and procedural skills

Echocardiography is the most commonly used non-invasive imaging tool for detection and quantification of atrial septal defects. Manual evaluations of echocardiographic videos required highly skilled clinical experts and is a time-consuming process.

Translational outlook

Algorithms based on deep learning approaches have the potential to automate and increase efficiency of the clinical workflow for detecting atrial septal defects and measuring the size of defect and the residual rim.

Data availability statement

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

Ethics statement

Ethics Committee (EC) approvals were obtained from each hospital for the use of deidentified echocardiographic and patient demographic data and the study was registered with the Chinese clinical trial registry (ChiCTR2000030278).

Author contributions

KH, XL and FY designed this study, analyzed and interpreted the patient data, drafted the manuscript, and guaranteed that all aspects of the work was investigated and resolved. DB and KH critically revised the important intellectual content of the manuscript. XC, WW, WL, QW, LZ, XL, YD, XW and DL collected, analysed and interpreted the data. YC, XC, HP and PZ designed the network architecture, and performed the data preparation and analysis. All authors contributed to the article and approved the submitted version.

Funding

FY received support from National Natural Science Foundation of China (82202265). KH received support from Ministry of Industry and Information Technology of China (2020-0103-3-1).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcvm.2023.985657/full#supplementary-material.

References

1. Zimmerman MS, Smith AGC, Sable CA, Echko MM, Wilner LB, Olsen HE, et al. Global, regional, and national burden of congenital heart disease, 1990–2017: a systematic analysis for the global burden of disease study 2017. *Lancet Child Adolesc Health*. (2020) 4(3):185–200. doi: 10.1016/s2352-4642(19)30402-x

2. Roberson DA, Cui W, Patel D, Tsang W, Sugeng L, Weinert L, et al. Three-Dimensional transesophageal echocardiography of atrial septal defect: a qualitative and quantitative anatomic study. *J Am Soc Echocardiogr*. (2011) 24(6):600–10. doi: 10.1016/j.echo.2011.02.008

3. Baumgartner H, De Backer J, Babu-Narayan SV, Budts W, Chessa M, Diller G-P, et al. 2020 Esc guidelines for the management of adult congenital heart disease. *Eur Heart J*. (2021) 42(6):563–645. doi: 10.1093/eurheartj/ehaa554

4. Donofrio MT, Moon-Grady AJ, Hornberger LK, Copel JA, Sklansky MS, Abuhamad A, et al. Diagnosis and treatment of fetal cardiac disease. *Circulation*. (2014) 129(21):2183–242. doi: 10.1161/01.cir.0000437597.44550.5d

5. Bartakian S, El-Said HG, Printz B, Moore JW. Prospective randomized trial of transthoracic echocardiography versus transesophageal echocardiography for assessment and guidance of transcatheter closure of atrial septal defects in children using the amplatzer septal occluder. *JACC Cardiovasc Interv*. (2013) 6(9):974–80. doi: 10.1016/j.jcin.2013.05.007

6. Silvestry FE, Cohen MS, Armsby LB, Burkule NJ, Fleishman CE, Hijazi ZM, et al. Guidelines for the echocardiographic assessment of atrial septal defect and patent foramen Ovale: from the American society of echocardiography and society for cardiac angiography and interventions. *J Am Soc Echocardiogr*. (2015) 28(8):910–58. doi: 10.1016/j.echo.2015.05.015

7. unçalp Ö, Pena-Rosas JP, Lawrie T, Bucagu M, Oladapo OT, Portela A, et al. Who recommendations on antenatal care for a positive pregnancy experience-going beyond survival. *BJOG*. (2017) 124(6):860–2. doi: 10.1111/1471-0528.14599

8. Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, et al. Fully automated echocardiogram interpretation in clinical practice. *Circulation*. (2018) 138(16):1623–35. doi: 10.1161/CIRCULATIONAHA.118.034338

9. Kusunose K, Abe T, Haga A, Fukuda D, Yamada H, Harada M, et al. A deep learning approach for assessment of regional wall motion abnormality from echocardiographic images. *JACC Cardiovasc Imaging*. (2020) 13(2 Pt 1):374–81. doi: 10.1016/j.jcmg.2019.02.024

10. Huang MS, Wang CS, Chiang JH, Liu PY, Tsai WC. Automated recognition of regional wall motion abnormalities through deep neural network interpretation of transthoracic echocardiography. *Circulation*. (2020) 142(16):1510–20. doi: 10.1161/CIRCULATIONAHA.120.047530

11. Ouyang D, He B, Ghorbani A, Yuan N, Ebinger J, Langlotz CP, et al. Video-Based ai for beat-to-beat assessment of cardiac function. *Nature*. (2020) 580(7802):252–6. doi: 10.1038/s41586-020-2145-8

12. Yang F, Chen X, Lin X, Chen X, Wang W, Liu B, et al. Automated analysis of Doppler echocardiographic videos as a screening tool for valvular heart diseases. *JACC Cardiovasc Imaging*. (2022) 15(4):551–63. doi: 10.1016/j.jcmg.2021.08.015

13. Chollet F. *Xception: deep learning with depthwise separable convolutions. 2017 IEEE conference on computer vision and pattern recognition (CVPR)* (2017).

14. Li X, Ding L, Li W, Fang C editors. *Fpga accelerates deep residual learning for image recognition. 2017 IEEE 2nd information technology, networking, electronic and automation control conference (ITNEC)* (2017).

15. https://github.com/YixinChen-AI/ASD-AtrialSeptalDefect Github.

16. Iek Z, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger OJS. Cham. 3d U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation (2016).

17. Zhe C, Simon T, Wei SE, Sheikh Y editors. *Realtime multi-person 2d pose estimation using part affinity fields. 2017 IEEE conference on computer vision and pattern recognition (CVPR)* (2017).

18. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. *An image is worth 16 × 16 Words: transformers for image recognition at scale. International conference on learning representations* (2021).

19. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal Fundus photographs. *Jama*. (2016) 316(22):2402–10. doi: 10.1001/jama.2016.17216

20. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-Level classification of skin cancer with deep neural networks. *Nature*. (2017) 542(7639):115–8. doi: 10.1038/nature21056

21. Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal VK, et al. Deep learning algorithms for detection of critical findings in head ct scans: a retrospective study. *Lancet*. (2018) 392(10162):2388–96. doi: 10.1016/s0140-6736(18)31645-3

22. Østvik A, Smistad E, Aase SA, Haugen BO, Lovstakken L. Real-Time standard view classification in transthoracic echocardiography using convolutional neural networks. *Ultrasound Med Biol*. (2019) 45(2):374–84. doi: 10.1016/j.ultrasmedbio.2018.07.024

23. Madani A, Arnaout R, Mofrad M, Arnaout R. Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digit Med*. (2018) 1:91–7. doi: 10.1038/s41746-017-0013-1

24. Kusunose K, Haga A, Inoue M, Fukuda D, Yamada H, Sata M. Clinically feasible and accurate view classification of echocardiographic images using deep learning. *Biomolecules*. (2020) 10(5):665. doi: 10.3390/biom10050665

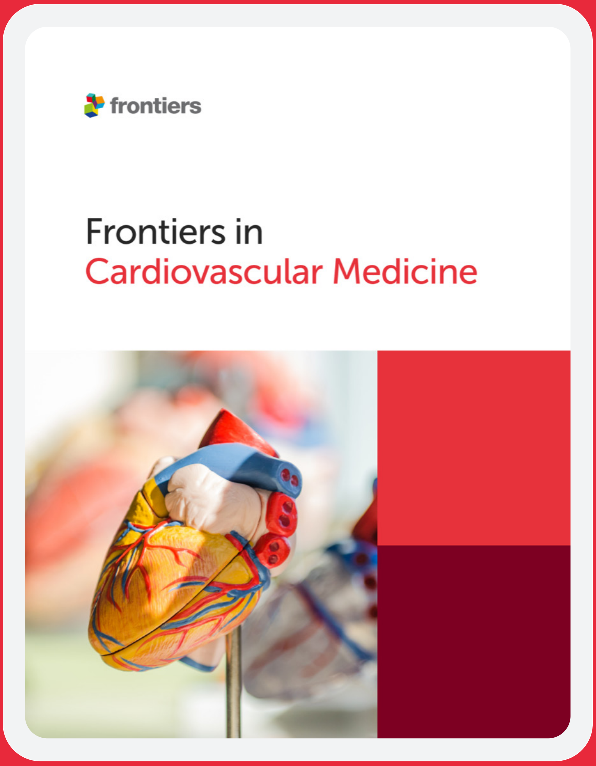
25. Wang J, Liu X, Wang F, Zheng L, Gao F, Zhang H, et al. Automated interpretation of congenital heart disease from multi-view echocardiograms. *Med Image Anal*. (2021) 69:101942. doi: 10.1016/j.media.2020.101942

26. Arnaout R, Curran L, Zhao Y, Levine JC, Chinn E, Moon-Grady AJ. An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease. *Nat Med*. (2021) 27(5):882–91. doi: 10.1038/s41591-021-01342-5

27. Zhao M, Wei Y, Lu Y, Wong KKL. A novel U-net approach to segment the cardiac chamber in magnetic resonance images with ghost artifacts. *Comput Methods Programs Biomed*. (2020) 196:105623. doi: 10.1016/j.cmpb.2020.105623

28. Mori H, Inai K, Sugiyama H, Muragaki Y. Diagnosing atrial septal defect from electrocardiogram with deep learning. *Pediatr Cardiol*. (2021) 42(6):1379–87. doi: 10.1007/s00246-021-02622-0

29. Faletra F, Scarpini S, Moreo A, Ciliberto GR, Austoni P, Donatelli F, et al. Color Doppler echocardiographic assessment of atrial septal defect size: correlation with surgical measurements. *J Am Soc Echocardiogr*. (1991) 4(5):429–34. doi: 10.1016/s0894-7317(14)80375-1



frontiers

Frontiers in
Cardiovascular Medicine

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact