UiO : University of Oslo

Andrew Gilbert

Automating Echocardiography Analysis using Deep Learning

Efficient measurement, workflow, and data generation

Thesis submitted for the degree of Philosophiae Doctor

Department of Informatics
Faculty of Mathematics and Natural Sciences

GE Vingmed Ultrasound
Cardiovascular Ultrasound R&D

2021

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 2419

ISSN 1501-7710

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

It is not in what you succeed in doing that you get your joy,
but in the doing of it.
- Jack London, Martin Eden

Preface

This thesis is submitted in partial fulfillment of the requirements for the degree of *Philosophiae Doctor* at the Department of Informatics, Faculty of Mathematics and Natural Sciences, University of Oslo. The research was conducted at the University of Oslo and at GE Vingmed Ultrasound in the framework of the Personalized In-silico Cardiology European training network. The research was conducted under the supervision of Dr. Kristin McLeod and co-supervision of Dr. Eigil Samset, Dr. Pablo Lamata, and Dr. Svein Arne Aase between May 2018 and May 2021.

This work received funding from the European Union's Horizon 2020 research and Innovation program under the Marie Sklodowska-Curie grant agreement No 764738.

Acknowledgements

Throughout this project I have been lucky to have a very knowledgeable and compassionate set of advisors that I am indebted to. Krissy, Eigil, Pablo, and Svein Arne not only taught me much about technical subjects, they have shown me how to set high goals, be efficient, work hard, and how to do all that with positivity and good cheer. They have always brought new perspectives to challenge me, and kept me focused. This project would never have been finished without them. A particular thanks to Krissy for building my confidence by always pushing me to get outside my comfort zone, whether it was running marathons or meeting deadlines. I am also grateful to Julia Schnabel for some perfect advice at key points in this project.

A big thank you to all of my colleagues at GE Vingmed, especially those in the Digital and AiVengers teams (Jurica, Qing, Xiaojuan, Mujde, Cristiana, Krissy, Eigil, Svein Arne, Siying, and everyone else). Also thank you to all of the collaborators on our projects, especially Marit and Line at Norsk Regnesentral. And of course, thank you to my colleagues at the University of Oslo, especially Børge, Tollef, and David. I'm lucky to have had all of these colleagues whose professionalism and attention to detail taught me a lot about how to take a project through to completion. Perhaps more importantly, they made it enjoyable to come to work every day.

I'm deeply grateful for the friendships, opportunities, and new insights provided by the entire PIC consortium. PIC would not have been possible or nearly as fulfilling without the dedication of Pablo Lamata who brought us all together and constantly worked to put each of us in the best position to succeed. A special thank you to my favorite classicists Ali, Filip, Joao, Jorge, Maciej, and

Manuel for the utterly chaotic PIC pre-/post-meeting trips. They were a riot and always left me with a full dose of memories and questions to ponder.

I moved to Oslo in 2018 without knowing anyone and this PhD would not have been possible without all of the friends who made Oslo an excellent place to live these past three years. Tusen takk to everyone for the volleyball matches, weekend climbing/skiing trips, and fun times which rounded out my life so well these years. And thanks to Manuel for making sure I never took anything too seriously during this crazy past year.

This thesis represents the culmination of one step of a journey and that journey has been shaped by the influences of my family who brings out the best in me. I'm eternally grateful to my parents Eric and Liza, for the values they have instilled in me, for their enduring and incredibly strong support no matter where I am, and for their sacrifices to always put me in the best position to succeed. My grandparents Ed and Marjorie have always been an especially strong positive influence on me, cultivating my intellectual curiosity and zest for life in equal parts. I always endeavour to live up to the legacy they have set. Finally, my three brothers Griff, Johnny, and Henry have been the primary formative influence on my life. My brothers are the best. I'm repeatedly challenged to keep pace with them intellectually and athletically, and I'm constantly striving to match their loyalty and humility, if not their dance moves. There's no one I'd rather start a bobsled team with.

Andy
Oslo, June 2021

Abstract

Cardiovascular ultrasound imaging (echocardiography) is the primary imaging modality used to assess cardiac morphology and function. Real-time feedback, lack of ionizing radiation, and lower cost make echocardiography ideally suited for rapid diagnostic use in patients with cardiovascular disease. However, despite its widespread use, measurements of structural and functional parameters in echocardiography have high variability. In addition, many clinics are facing an increased workload due to both increased number of patients from aging populations, and standard imaging protocols expanding as more imaging and quantification techniques become mainstream. These factors create a demand for automated tools that can increase reproducibility and efficiency.

Deep learning is a sub-field of artificial intelligence which can provide automated sophisticated analysis of natural images, offering the potential to address these demands. This thesis describes methods for applying deep learning to enable better workflows in echocardiography. Specifically, four aspects were investigated.

First, classification techniques can improve the efficiency of workflows by automatically determining which measurements and analysis should be applied to a given image. In this thesis, a highly accurate method for classifying spectral Doppler images is presented. We demonstrated how multi-modal information in each Doppler recording can be combined using a meta parameter post-processing scheme and heatmaps to encode coordinate locations. We explored the effects of various input/output combinations and proposed a confidence metric to prevent misclassifications in data types that were unseen during training. The proposed method was shown to be highly accuracy in determining the suitable measurement(s) to perform on Doppler images.

Second, automated measurements can improve the efficiency and reproducibility of quantitative analysis. We developed a deep learning method to perform multiple measurements simultaneously in 2D echocardiography. The proposed method used anatomically meaningful heatmaps as labels and a multi-component loss function to achieve high accuracy. Measurement error was comparable to intra-observer error.

Third, automated analysis techniques open up the possibility for new measurements that can enhance diagnostic power. We explored the use of septal curvature as a measure of basal septal hypertrophy, which is an early marker of remodelling in patients with hypertension. Curvature measurements led to more reproducible and robust results that better correlated to other functional parameters of remodelling related to hypertension than traditional measurements.

Fourth, the expensive nature of acquiring labeled training data and the high

inter-/intra-observer variability of labels slows the development of new automated tools in echocardiography. To address this we studied methods to automate the collection of large annotated datasets. Specifically, anatomical models were used as sources for high-quality ground truth labels and corresponding ultrasound images were synthesized using generative adversarial networks. Networks trained with synthetic images showed good performance when tested on real images, with accuracy scores matching inter-observer error.

Overall, the methods described in this thesis contributes to improved analysis in echocardiography, adding tools to increase the standard of care.

List of Papers

Paper I

Gilbert, A., Holden, M., Eikvil, L. Rakmail, M., Babić, A., Aaset, S. A., Samset, E., and McLeod, K. "User-Intended Doppler Measurement Type Prediction Combining CNNs with Smart Post-Processing". In: *Journal of Biomedical and Healthcare Informatics.* (2020), DOI: 10.1109/JBHI.2020.3029392.

Paper II

Gilbert, A., Holden, M., Eikvil, L., Aase, S. A., Samset, E., and McLeod, K. "Automated Left Ventricle Dimension Measurement in 2D Cardiac ultrasound via an Anatomically Meaningful CNN Approach". In: *Lecture Notes in Computer Science.* Vol. 11798, (2019), pp. 29-37. DOI: 10.1007/978-3-030-32875-7_4.

Paper III

Marciniak, M., **Gilbert, A.**, Loncaric, F., Fernandes, J. F., Bijnes, B., Sitges, M., King, A., Crispi, F., and Lamata, P. "Septal Curvature as a Robust and Reproducible Marker for Basal Septal Hypertrophy". In: *Journal of Hypertension.* Vol. 38, (2021), DOI: 10.1097/HJH.0000000000002813.

Paper IV

Gilbert, A., Marciniak, M., Rodero, C., Lamata, P., Samset, E., and McLeod, K. "Generating Synthetic Labeled Data from Existing Anatomical Models: An Example with Echocardiography Segmentation". In: *Transactions in Medical Imaging.* (2021), DOI: 10.1109/TMI.2021.3051806.

Contents

Chapter 1

Introduction

1.1 Motivation

Cardiovascular disease can reduce the heart's ability to distribute blood throughout the body. It accounts for 20 million deaths every year, making it the most common cause of death worldwide [1]. Cardiovascular ultrasound, or echocardiography, uses ultrasound waves to view the structure, motion, and blood flow of the heart. A fully analyzed echocardiography exam gives a holistic evaluation of heart health that is valuable when performing diagnosis and preparing a treatment plan for patients with cardiovascular disease. New processing techniques in echocardiography bring advanced visualizations of blood flow and image quality is continually improving with hardware and post-processing advancements. Given the range of metrics it provides in addition to the relative accessibility and real-time feedback, echocardiography is the most widely used imaging modality when evaluating cardiovascular health [2].

However, acquisition, measurement, and evaluation of echocardiography scans requires precision and training. Moreover, new metrics are continually being added as new research highlights the potential for improved diagnostics. The combination of more extensive exam protocols and an aging population places pressure on already overburdened healthcare systems and raises the need for automated tools that can simplify clinical workflows.

Deep learning is a sub-field of artificial intelligence which relies on multi-layered neural networks to automatically process and understand natural inputs. When given labeled training data, supervised deep learning techniques can match human performance on common image analysis tasks such as classification, regression, or semantic segmentation. These tasks map well to the requirements of automated tools in medical imaging. Initial development of these techniques within the medical domain has been challenging due to the lack of large open-source datasets and adoption has been slow due to a hesitancy to rely on black box methods in patient care settings. However, increasing availability of medical-specific datasets and techniques to understand how networks make decisions [3] have led to the move towards clinical adoption of deep learning solutions yielding high performance [4].

Apart from efficiency improvements, deep learning techniques also offer the opportunity to increase the accuracy and reproducibility of echo measurements. Echo images can be more challenging to interpret than other modalities and there are different opinions on how best to perform measurements, even on high-quality images. Automated measurement systems can standardize measurement practices. Standardization leads to more personalized treatment since decreased variability increases the statistical significance of obtained

measurements, increasing the diagnostic value of those measurements.

Despite the positive successes achieved through the application of deep learning, adapting to new applications remains a challenge. Each new application requires a new labeled dataset which is expensive and time-consuming to acquire. Augmentations, transfer learning, and domain adaptation can help reduce this cost but nonetheless, annotation remains the largest hurdle in deep learning applications. This is particularly a challenge in medical imaging where accurate annotations are time-consuming and expensive to obtain due to the high level of required expertise.

1.2 Aims of this project

The main goal of this thesis is to describe the use of deep learning techniques in echocardiography. The first two parts of this thesis detail the application of novel deep learning techniques to automate workflows and measurements in echocardiography including specific adaptations for medical data. The third part is a proposed method to unlock novel clinical metrics to improve diagnostic power by leveraging the processing capacity of deep learning. Finally, in the fourth part, novel methods are described to automatically generate synthetic data to facilitate data acquisition for training new deep learning algorithms.

1.3 Context of the project

This thesis project was completed as a part of the Personalized In-silico Cardiology (PIC) research project. The PIC project is a European Innovative Training Network; a partnership of academic, industrial, and clinical partners focused on transforming individualized cardiac care[1]. As outlined in the Digital Twin consortium position paper, attached as an appendix to this thesis, the PIC project aims to optimize diagnostics and therapy through personalized care [5]. This is accomplished through the use of both mechanistic models that provide interpretable predictions and statistical models that automatically extract parameters and find hidden patterns. The first parts of this thesis are targeted towards statistical methods for the extraction of parameters. The final part demonstrates the synergies between mechanistic and statistical models by combining structural understanding with image synthesis.

The work packages (WPs) of the PIC consortium are shown in Figure 1.1. This thesis project fits within the scope of WP2, WP3, and WP5, as described below:

- **WP2:** This thesis contributed to WP2 through the development of models to automatically extract markers and predictions that enable diagnosis.
- **WP3:** This thesis proposed methods to use the information encoded within anatomical models to generate new data sources.

[1] https://picnet.eu/: Marie Sklodowska-Curie grant agreement No 764738

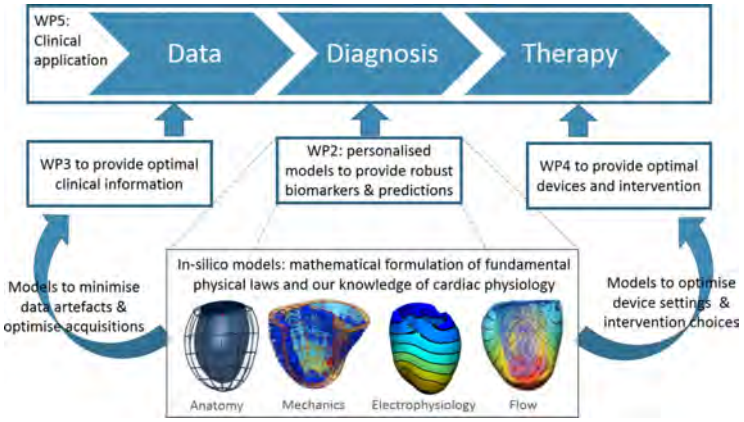


Figure 1.1: Work packages (WPs) within the PIC consortium.

- **WP5:** Several of the methods proposed in this these were integrated into clinical workflows to improve diagnosis and therapy.

The work for this thesis was primarily conducted at GE Vingmed Ultrasound (subsidiary of GE Healthcare), world leaders in cardiovascular ultrasound equipment. Many of the objectives of this project were targeted towards applications that were directly relevant to Vivid ultrasoud customers. GE Vingmed provided the opportunity to receive early feedback from key industrial and clinical leaders and several of the methods developed in this thesis were tested and integrated within the diagnostic pipeline of GE's line of Vivid echocardiography scanners[2].

[2] https://www.gehealthcare.co.uk/products/ultrasound/vivid

Chapter 2

Background

This section provides necessary background information for presenting the articles included in this thesis. First, a background on deep learning, the principle method used in this thesis, is presented. Second, details on the anatomy of the heart and cardiovascular disease are presented. Third, a brief summary of echocardiography is provided. Finally, an overview of automation techniques within echocardiography is presented.

2.1 Deep learning

Deep learning is a sub-field of artificial intelligence in which multi-layered neural networks are trained from large amounts of data to automatically interpret features in natural inputs such as images, videos, text, or audio. Powered by an exponential increase in computing power and parallel training enabled by graphics processing cards, research in deep learning has exploded in the decade since it was spearheaded by the introduction of convolutional neural networks (CNN) in 2012. The first CNNs pioneered the use of deep learning in image classification and doubled the performance on the popular ImageNet database [6]. Since then, deep neural networks have revolutionized machine performance on a number of complex tasks such as object detection [7], translation [8], speech recognition [9], game play [10], and autonomous vehicles [11].

The proficiency of deep neural networks for automatically processing complex input features into actionable insights makes deep learning a natural choice for analyzing medical images. Deep learning has been applied to a variety of tasks in medical imaging, including measurements, segmentation, object detection, classification, registration, and risk prediction [4], [12]–[15]. A brief overview of deep learning techniques relevant to this thesis is presented below.

2.1.1 Convolutional neural networks

Convolutional neural networks (CNNs) consist of a series of layers of filters where each filter is convolved over the image or the output of the previous set of filters. More specifically, a layer contains K kernels, $W = \{W_1, W_2, ...W_K\}$ and associated biases $B = \{b_1, b_2, ...b_K\}$. As shown in (2.1), each filter is used to generate a new feature map (X_k^l) through a convolution with the outputs from the previous layer (X^{l-1}), as described in (2.1). To model non-linear effects an element-wise non-linear transform (σ) is typically applied to each map (e.g. a rectified linear unit [16]).

$$X_K^l = \sigma(W_K^{l-1} \circledast X^{l-1} + b_K^{l-1}) \tag{2.1}$$

This structure makes networks more efficient than fully connected networks because detectors are learned for similar objects occurring at any point in image space. Over time, typical architectures of convolutional neural networks have evolved to achieve high performance for a wide variety of tasks. These architectures are varied, but typically follow similar patterns: a) the height/width of feature maps decrease at deeper layers in the network through pooling operations while the depth increases through the application of more filters, b) convolutional layers are interspersed with regularization, normalization, and non-linear layers, and c) skip connections pass features from shallow to deeper layers and serve as "gradient highways" during back-propagation to speed up training and enable deeper networks.

Simplified diagrams of two common network architectures, U-Net [17] and ResNet [18], are shown in Figure 2.1. U-Net is commonly used for segmentation tasks. It consists of an initial down-sampling path followed by an up-sampling path to the original input size, with skip connections bridging the two. ResNet is a common choice for classification or detection tasks. It consists of residual blocks (groups of convolutions, normalizations, and non-linear layers with skip connections). The depth of U-Net and the number of residual blocks in ResNet can be modified depending on the desired performance/speed trade-off and available data of the chosen task. Bianco et al. give a complete overview of the trade-offs of common architectures in performance, memory size, and speed [19].

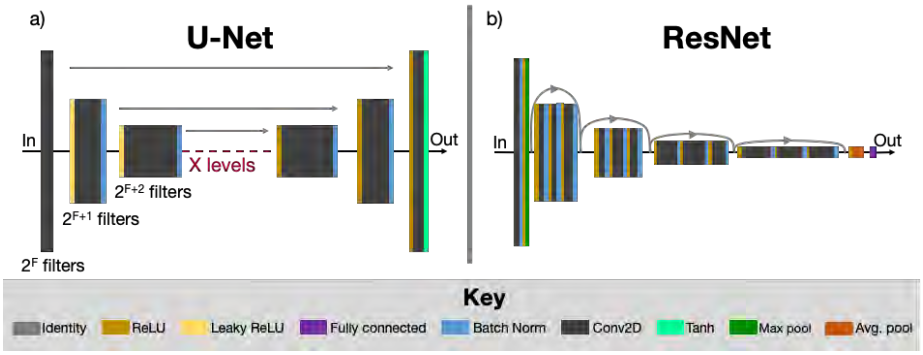


Figure 2.1: Simplified diagrams of two common network architectures a) U-Net [17] and b) ResNet (in this case ResNet-18) [18].

There are many possible variations on these common architecture structures, and custom layers have also been developed for specific applications. Two examples of relevant custom layers include coordinate convolution [20] and soft-argmax [21], [22]. Coordinate convolution involves appending channels representing image coordinates in the x and y dimension before the convolution operation and is typically done at the input to the network. This provides location information in cases where the location of a feature within an image is important. A soft-argmax layer consists of an element-wise multiplication with channels representing the x and y coordinates. This can be used to extract the

center of mass of a relevant object when the feature map is a heatmap describing that object's location. Representations of both layers are shown in Figure 2.2.

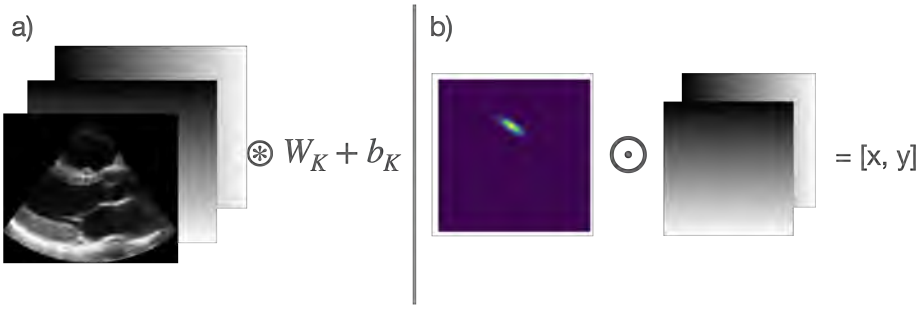


Figure 2.2: Two examples of custom neural network layers. a) In coordinate convolution, channels representing the $x-$ and $y-$coordinates are appended before the convolution operator [20]. b) In a soft-argmax layer, the center of mass of a heatmap is extracted through element-wise multiplication with coordinate channels [21]. The heatmap should be a probability distribution.

2.1.2 Optimization

Neural networks are optimized through gradient descent. Given a differentiable objective function and a labeled sample, the difference between the desired result and calculated result is measured (the loss). The contribution of each filter to the loss is determined by back-propagating through the network layers and an update to the kernel weights (W) and biases (b) is calculated from the back-propagated loss and a learning rate (α). Specifically, given loss L resulting from a prediction from a series of convolutional layers (shown in Equation (2.1)), each filter W_k^l will be updated by $W_k^l = W_k^l - \alpha * \frac{\partial L}{\partial W_K^l}$. The update $\frac{\partial L}{\partial W_K^l}$ is calculated in earlier layers through the application of the chain rule.

This process is repeated for batches of labeled samples to optimize performance across the entire dataset. Weight updates typically include a momentum term to overcome local minima and learning rate schedulers which dynamically update α through the course of training. Techniques such as dropout, normalization, or regularization are used to prevent over-fitting to the training data.

2.1.3 State-based networks

It is often advantageous for a network to maintain a state in between successive predictions. In medical imaging this can be useful when making predictions between time points (e.g. [23], [24]) or when making predictions for the same patient given new data. State-based networks such as recursive neural networks [25], long short-term memory networks [26], gated recurrent units [27],

reinforcement learning [28]–[30], or more recently transformer networks [31] are useful for accomplishing this task.

2.1.4 Generative adversarial networks

Generative adversarial networks (GANs) were first developed for image generation, but more broadly offer a different paradigm of training from the traditional optimization described in Section 2.1.2. In an adversarial setup the objective function is another network that is also dynamically improving during training, rather than a static function.

Specifically, GANs consist of two networks, a generator ($G(z;\theta_G)$) with parameters θ_G and and discriminator ($D(x;\theta_D)$ with parameters θ_D. The generator seeks to match the distribution of its output (p_g) to the distribution of a sample of real data (p_{data}) by transforming noise (z). D attempts to differentiate between real data (x) and the output of the generator ($G(z)$). As D learns to better differentiate, G must produce a more realistic output that better matches p_{data}. Mathematically, this corresponds to a minimax game over the objective V [32]:

$$\min_G \max_D \; V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[log\ D(x)] + \mathbb{E}_{z\sim p_z(z)}[log\ (1 - D(G(z)))] \quad (2.2)$$

An overview of the general GAN architecture is shown in Figure 2.3. The exact architectures and training procedures of the two networks are variable and can be optimized based on the task. For example, G can be conditioned on a separate variable [33] and this condition can take the form of another image [34] to enable image-to-image translation. Other innovations in training and network architecture have enabled the creation of high-resolution images that are virtually indistinguishable from real images [35].

A more comprehensive review of state-of-the-art GAN techniques is given by Wang et al. [36] and detailed specifically for medical imaging by Kazeminia et al. [37]. One relevant innovation is cycle-consistent GANs (CycleGANs), which rely on a paired GAN structure to perform unpaired image-to-image translation [38]. As shown in Figure 2.4 CycleGANs combine adversarial training with image reconstruction losses used in auto-encoders [39] to enable translation between two imaging domains without a dataset of paired images.

2.1.5 Deep learning workflows

The typical workflow for developing and testing a deep learning tool to be integrated into clinical practice is shown in Figure 2.5. After the identification of a relevant problem (see Section 2.4) the first step is the collection of data which defines the scope of the solution, a trade-off between robustness and resources. Increasing the variety of input data may help make the final network more generalizable in practice, but will also exponentially increase the size of the required dataset as more variations are included. Augmentations during training

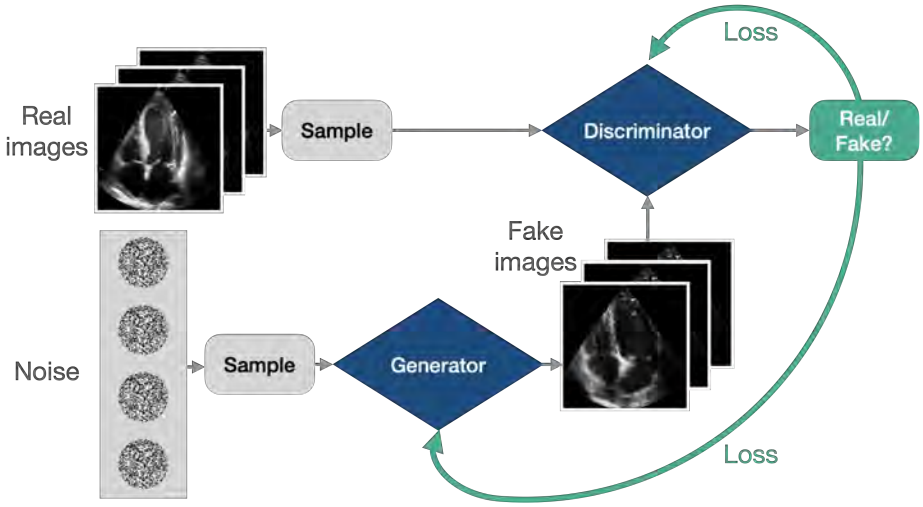


Figure 2.3: The general architecture of generative adversarial networks (GANs). A generator converts a noise sample to a fake image while a discriminator attempts to distinguish between the generated images and real images. The networks have opposite objectives for the same loss function

may help to reduce some of the required data for training, but all data types should still be included in the validation datasets. The second step is annotating the collected data to reflect the task the network should complete. Accuracy expectations can also be defined in this step by evaluating inter-observer and intra-observer errors. The third step is the training and validation of the network where the architectures and modifications described above can be applied. Finally, the tool is deployed in a clinical workflow.

2.1.6 Challenges for deep learning in medical imaging

There are several critical challenges for applying deep learning in medical imaging workflows.

First, deep learning models are increasingly data-hungry, while data collection remains a substantial (albeit sensible) hurdle within medical imaging due to privacy regulations, which makes automated collection of large datasets such as ImageNet impossible. Transfer learning from non-medical domains has been successful in some cases, while providing no benefit in others due to the large difference in feature appearance [40] [1]. The smaller changes in pixel space of medical images compared to traditional images can be an additional challenge since the important feature differences between images are subtler and noisier. Several open-source echocardiography datasets are now available [41]. However,

[1] Pre-trained models were tested in Paper II and Paper I and did not improve results

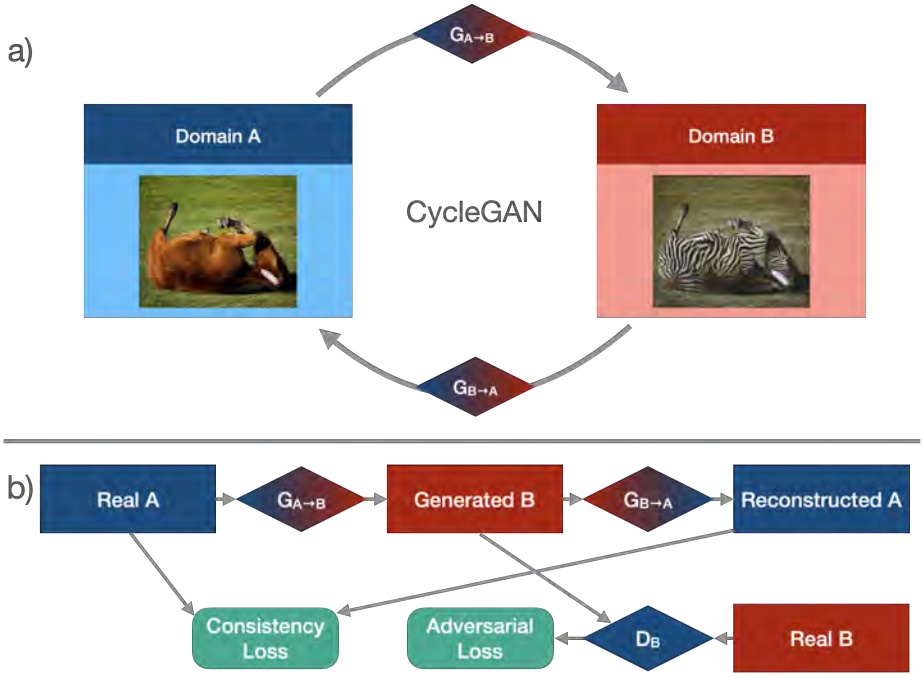


Figure 2.4: CycleGANs can be used for unpaired domain adaptation. a) A CycleGAN consists of two generative networks. One is responsible for transforming from domain A to domain B and one from B to A. b) During training an image is passed through both generative networks. The objective is both the valid reconstruction of the original input (auto-encoder loss) and the creation of a realistic image in the second domain (adversarial loss). Horse and zebra example images are from [38].

these datasets don't cover variations in machines, views, and pathologies, which can limit the ability to build tools that can be implemented in clinical practice.

Second, obtaining accurate annotations also represents a substantial challenge for medical imaging. Accurately reading medical images requires expertise and those with the required skillset are expensive. Augmentations can help in this area. Some groups have proposed the use of statistical shape modeling to modify images to include new natural shapes [42], [43] while others have proposed using GANs to generate new realistic images and labels based on a prior distribution [44]–[47]. In Paper IV we propose a method to solve this challenge by generating images from anatomical models. We use both statistical shape models to add new shape variations and GANs to generate realistic image appearances.

Third, the increase in the number of architectures, custom layers, loss functions, augmentations, and other possible adaptations has also made training

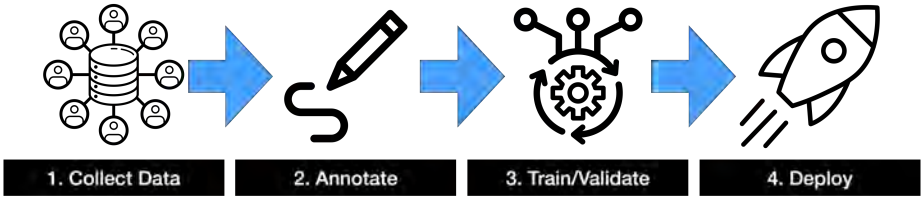


Figure 2.5: A simplified deep learning workflow. After determination of the relevant problem and desired accuracy, data is collected and annotated, the model is trained and validated, and the tool is deployed

of networks a time-consuming process due to the need to try many configurations. While some attempts have been made to automate this process [48], creation of optimal solutions requires a mix of application-specific and technical experience.

Finally, deployment remains a significant challenge as there is a wide range in the computing power of echocardiography devices. Cloud deployment is possible in some cases, but many hospitals are reluctant to use cloud-based solutions and/or don't have the proper infrastructure. Post-processing measurements and diagnostics can be shifted to the cloud as regulatory requirements and infrastructure are established, but real-time feedback applications will likely continue to require on-device implementations.

While deep learning has the potential to significantly improve echocardiography workflows, there are a number of remaining challenges in developing automatic workflows for echocardiography, as described above. As such, this thesis describes methods to address these challenges to drive forward the use of deep learning in echocardiography.

2.2 Human heart

The heart is responsible for pumping blood through the cardiovascular system which distributes oxygen and nutrients throughout the body. The heart functions through an intricate balance of the forces from electrically induced muscular contractions and the pressures between different chambers and the circulatory systems.

2.2.1 Anatomy and function

The heart pumps blood through the body's two circulatory systems: pulmonary and systemic. The pulmonary system carries blood to and from the lungs while the systemic system distributes and returns blood from the rest of the body. Pumping is controlled by four chambers and valves as shown in Figure 2.6. The right side of the heart drives deoxygenated blood into the pulmonary system toward the lungs while the left side propels blood into the systemic system.

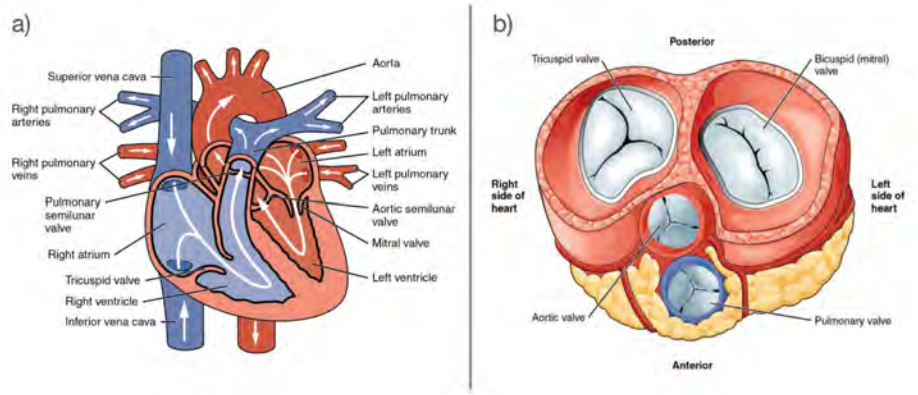


Figure 2.6: a) the basic anatomy and direction of blood flow in the heart. b) A sliced view from the top of the heart demonstrating the position and anatomy of the four valves in the heart. Images from [49].

Blood flow is accomplished through the alternating contraction and relaxation of the heart muscles surrounding the four chambers. A single round of contraction and relaxation constitutes a cardiac cycle. The cardiac cycle can be further divided into several phases as shown in Figure 2.7 and described in detail in [50].

On the left side of the heart, ventricular diastole begins with the closure of the aortic valve. During ventricular diastole, the mitral valve opens and blood flows into the left ventricle through the left atrium, expanding the size of the ventricle. At the end of ventricular diastole, atrial systole is initiated, contracting the atrium and forcing additional blood into the ventricle. Diastole ends with the closure of the mitral valve and with the left ventricle at maximum volume. In systole, the left ventricle begins a period of isovolumetric contraction with the aortic valve still closed. At the point where the pressure in the left ventricle exceeds the aortic pressure, the aortic valve opens and the blood is forced out into the aorta. The ventricle rapidly contracts and reaches its smallest volume at end-systole, when the aortic valve closes and the process begins again. A similar process occurs on the right side of the heart, although the pressures are lower.

The contraction of cardiac muscles is initiated by the electrical system of the heart. Electrical impulses originate from the sinoatrial node, which emits a signal once per cardiac cycle. This activates the contraction of the atrial musculature and also travels to the atrioventricular node, which subsequently activates the musculature of the ventricles [51]. Meanwhile, the function of the heart valves is regulated by the changes in pressure between different chambers.

2.2.2 Cardiovascular disease

Despite its complexity, the heart is generally robust. However, cardiovascular disease can cause structural or functional impairments that lead to disability,

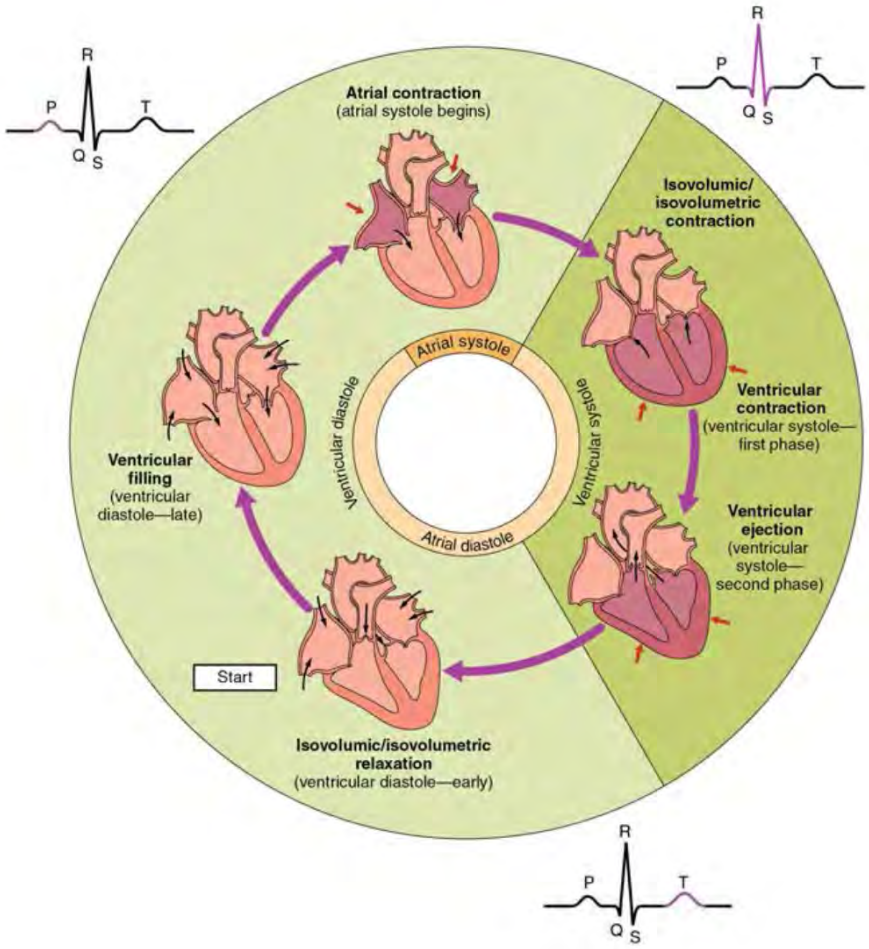


Figure 2.7: The cardiac cycle consists of one complete contraction and relaxation of ventricles and atria as shown above. The contraction is regulated by electrical impulses and the pressures and volumes in the chambers. Image from [49].

lower quality of life, or mortality. Cardiovascular health is a combination of genetics and lifestyle factors such as diet, exercise, drug use, and stress.

Cardiovascular disease includes conditions effecting all aspects of structural and functional heart health. For example, arrhythmias are caused by changes in function of the electrical system which impede proper contraction of cardiac muscles [52], [53]. Changes in valve anatomy cause leaks (regurgitation) or narrowing (stenosis) which also inhibits normal blood flow [54]. Additionally, cardiac muscles can weaken or stiffen and lead to cardiomyopathies. Cardiomyopathies are typically caused by toxins (alcohol, drugs, chemotherapy, and others) or

coronary artery disease (a buildup of plaque in the arteries bringing blood to the muscles).

All of these diseases can lead to a reduced pumping capacity of the heart in either the left or right side. While new studies have increasingly focused on the role of the left atrium [55] and the harmful effects of right heart failure [56], [57], the epidemiology, diagnosis, and treatment of heart failure has primarily focused on the role of the left ventricle. Heart failure is broadly classified in two types based on the ejection fraction of the heart, where ejection fraction describes the percentage of blood pumped out from the left ventricle during each cycle:

- **Heart failure with reduced ejection fraction** occurs when the heart pumps out a lower percentage of the blood in the left ventricle than normal. This may be caused by cardiomyopathy, impairment of the valves due to regurgitation or stenosis, high blood pressure in the arteries leading from the heart, or arrhythmias.
- **Heart failure with preserved ejection fraction** occurs when the heart still ejects a healthy/normal percentage of blood, but cannot properly relax. This reduces the volume of blood in the ventricle and thus the total volume of blood pumped. This may be caused by cardiomyopathy among other causes.

The literature covering variants, diagnosis, and treatment of heart failure is too varied to cover in detail here, but is regularly reviewed by clinical task-forces [57]–[60]. However, one marker that is particularly relevant to this thesis is hypertrophy, or abnormal thickening of the heart muscle. Left ventricle hypertrophy can be a marker of several pathologies, including hypertension, hypertrophic cardiomyopathy (HCM), sigmoid septum, aortic stenosis, or adaptation to physical training [61], which have varying prospective outcomes. For example, a sigmoid septum is not significantly correlated with cardiovascular disease, nor mortality [62]. However, distinguishing between these pathologies is difficult [63], [64] and hypertrophy itself is a major independent risk factor for mortality and thus indicates that further investigation is necessary in patients where hypertrophy is observed [65].

As described in this section, the heart is a complex, multi-faceted organ. Advanced analysis techniques are required to investigate problems that can arise due to disease or lifestyle factors. As such, this thesis describes techniques to better enable imaging and analysis of the heart to support diagnostic workflows.

2.3 Echocardiography

The complicated nature of cardiovascular disease means diagnosis requires a holistic view of the heart. This includes an analysis of cardiac structure, hemodynamics, and electrical function. Ultrasound is a method for visualizing internal structures by emitting high frequency sound waves and measuring the

strength of the reflected signals. Ultrasound has emerged as the primary method to provide a real-time imaging of the heart. A suite of analysis tools have been developed to measure specific aspects of structure, function, and flow.

Ultrasound applied to the heart is known as echocardiography and was initially developed in the early 1950s [66]. Echocardiography offers several inherent advantages as a diagnostic tool over other cardiac imaging modalities such as magnetic resonance imaging (MRI) and computed tomography (CT). First, it is a real-time imaging method that can be dynamically adjusted to focus on the relevant features in a given patient. Second, it has a lower cost and higher accessibility, and unlike CT, has no ionizing radiation. Third, echocardiography offers an improved temporal dimension compared to other modalities, with approximately 100 frames per cardiac cycle (depending on resolution and width of the scan), while MRI typically gives approximately 30 frames per cycle. Echocardiography has a lower spatial resolution, but this resolution is continuously improving and volume measurements using echocardiography exams are shown to correlate closely with those done with CT imaging [67].

Because of these advantages, echocardiography is the most widely used analysis tool for cardiologists. The community has also developed a set of standard acquisitions and measurements [68], along with a resulting set of normal values that allow clinicians to easily compare a variety of patient populations when making a diagnosis [69]. Moreover, patients will often receive many echocardiography scans over the course of their treatment, capturing the impact of the treatment on their health.

2.3.1 Image formation

Ultrasound imaging consists of the emission and reception of high frequency sounds, as shown in Figure 2.8. These waves are typically between 2-18 megahertz (MHz) for medical applications [70]. When the emitted wave hits a tissue boundary, part of the signal continues through the boundary, part is reflected, and a small part may be absorbed.

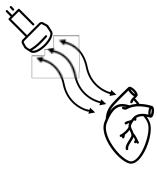


Figure 2.8: Ultrasound imaging works by emitting high-frequency sound waves from a probe and absorbing the corresponding reflected waves.

The strength of the ultrasound waves received back at the probe (echoes) are measured to form the image. The time at which the signal is received indicates the depth of the reflection, while the strength indicates the impedance of the tissue at that depth. Due to reflection, refraction, and absorption, the strength

(amplitude) of the ultrasound wave will decrease as the wave propagates through tissue, but this can be automatically compensated for in imaging systems. The strength of the received signal will also be a function of the angle of the tissue relative to the probe, and structures oriented perpendicular to the direction of the ultrasound waves will reflect a much stronger signal.

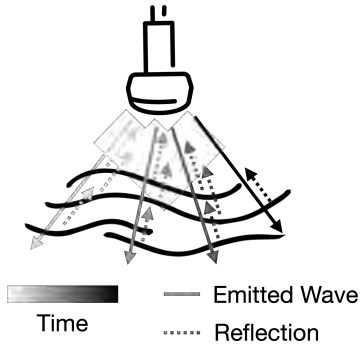


Figure 2.9: Echocardiography images are generated by emitting ultrasound in a single scan-line and progressively sweeping this scan-line across a fan-shaped region over time. The image is formed by measuring the strength and time delay of received reflections from each scan line.

An image is formed by emitting multiple ultrasound waves across a sector of scan-lines and measuring the reflected signals from each wave. In echocardiography, these waves are emitted from a small probe and directed outwards in a fan-like shape, as shown in Figure 2.9. This allows the ultrasound probe to fit between small windows (e.g. between ribs) while giving a larger viewing width at the depth of the heart. However, it also means that the lateral resolution will be much higher closer to the probe than deeper in the tissue.

While there are strong reflected signals at the boundary between two different materials, there are many individual scattering particles within tissue which also reflect some sound. These are known as speckles, and are useful in tracking the position of tissue over time. There are few scattering particles in blood so these regions will appear dark in an image.

Ultrasound waves are emitted using piezoelectric crystals. The application of an electric signal to these crystals causes them to vibrate, emitting a sound wave. The crystals can be dampened to emit only a short pulse. The same crystals can be used for recording because a sound wave hitting the crystal will cause a vibration and elicit an electrical response. Ultrasound probes are composed of arrays of piezoelectric crystals as well as acoustic focusing materials and electronics to stimulate and record from the crystals.

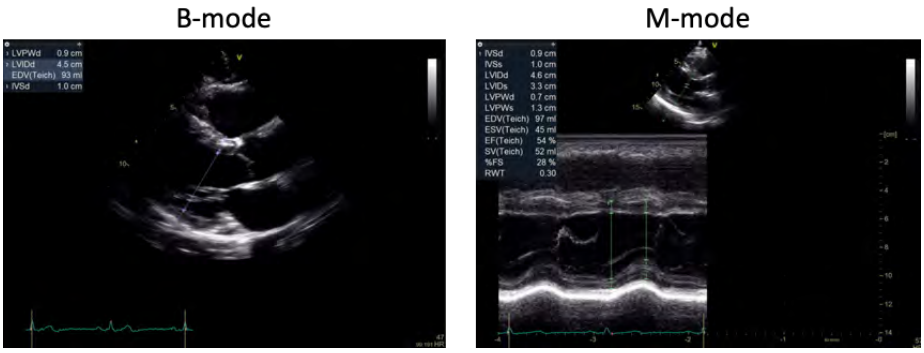


Figure 2.10: Example B-mode (left) and M-mode (right) echocardiography images. In both cases the y-axis is depth. In B-mode images, the X-axis is width and time is shown through a video loop. In M-mode images, the X-axis is time and a small B-mode frame is shown at the to orient the 1D acquisition in space. Both example images are in the parasternal long axis view (see Section 2.3.5) and demonstrate common measurements in that view (these are the measurements automated in Paper II).

2.3.2 2D imaging

There are several different modes of echocardiography imaging. In brightness mode (**B-mode**) imaging, ultrasound signals are successively emitted in a sweeping motion across a 2D plane as described above. For each emission (scan line) the echoes are recorded to create an image of a sector beneath the probe. This process is repeated to create a movie showing the motion of the heart.

Alternatively, pulses can be repeatedly emitted and recorded on the same scan line (**M-mode**). The signal is recorded in the same way, but each new pulse is visualized as a new column on the x-axis. Because only a single scan line is imaged, the temporal resolution of M-mode images is much higher. These images are used to view movement of high-velocity events, such as valve opening and closings, more clearly. Example B-mode and M-mode images are shown in Figure 2.10.

2.3.3 3D imaging

With advances in electronics, computing power, and signal processing, the same principles used for 2D B-mode imaging can be extended to create 3D real-time images. Real-time 3D is implemented in many commercial scanners and is already an important part of many workflows, such as live feedback during procedures. 3D offers the potential to better visualize complex anatomies and avoid errors caused by foreshortening (see Section 2.3.5), and will play an increasingly important role as resolution and visualization techniques improve. However, much of the analysis of echocardiography images is still based on 2D

images, where most of the previous data, tools, and guidelines are focused.

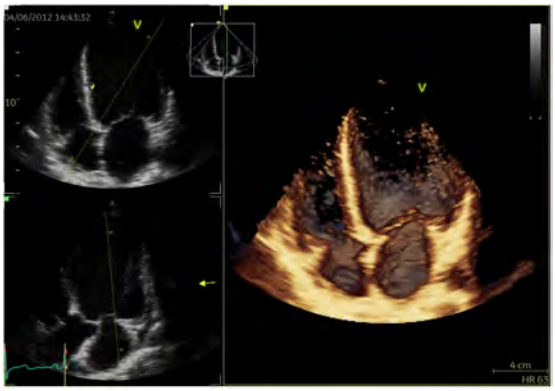


Figure 2.11: Example 3D B-mode image.

2.3.4 Doppler imaging

An alternative mode of imaging in echocardiography relies on the Doppler effect to analyze the velocity of tissue movement and blood flow. As described in Equation (2.3), the velocity (V) of a particle can be determined by analyzing the difference between the transmitted frequency (F_t) and received frequency (F_s) of a wave. This effect captures only the velocity in the direction of the transmission (hence the inclusion of $cos(\theta)$ in the denominator where θ is the angle between the wave and the motion of the particle).

$$V = \frac{c * (F_s - F_t)}{2 * F_t * cos(\theta)} \tag{2.3}$$

This effect can be harnessed in several ways. In **spectral Doppler** the frequency spectrum is analyzed at a single point over time (similar to M-mode imaging). This is useful for analyzing blood flow or tissue movement at specific regions of interest, such as valves. In spectral Doppler imaging, a cursor is positioned over the region of interest and the frequency spectrum at the given point is displayed over time. Spectral Doppler has two imaging modes: pulsed wave and continuous wave Doppler. In **pulsed wave (PW)** Doppler, short pulses of ultrasound are emitted (similar to above) and the frequency of each returning echo is recorded. A frequency spectrum from the region of interest is extracted and displayed as shown in Figure 2.14. However, due to the Nyquist limit, higher frequencies cannot be measured with PW Doppler. The Nyquist limit states that a waveform must be measured at least twice per wavelength to accurately detect the frequency. Otherwise, aliasing may occur, as shown in Figure 2.13. The frequencies measurable by PW Doppler are limited by the pulse repetition frequency, which is fundamentally limited by the speed of sound and depth of the region of interest.

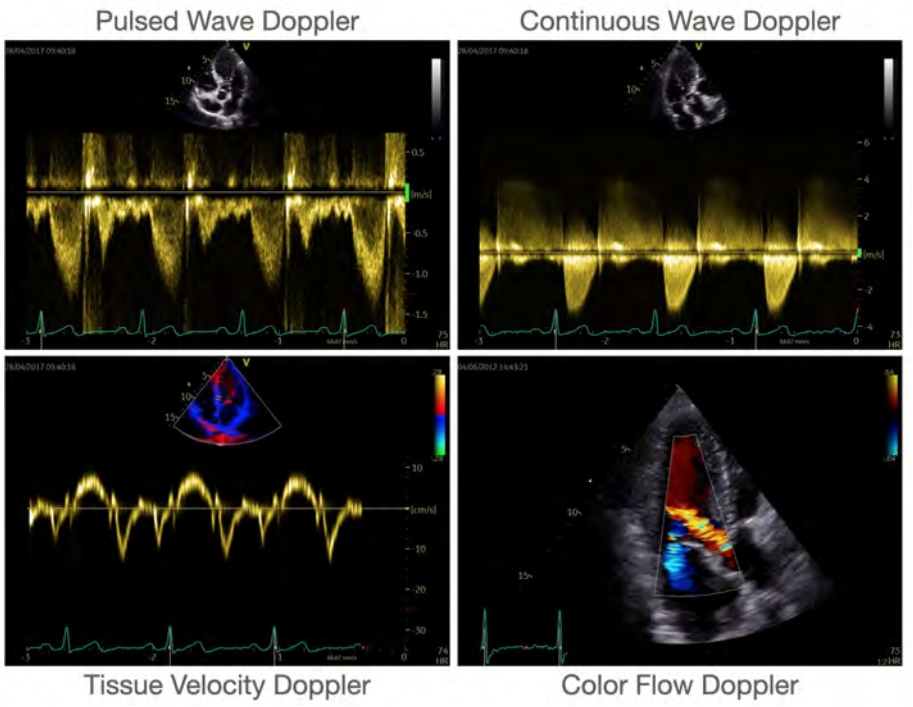


Figure 2.12: The various modes of Doppler acquisition: pulsed wave Doppler, continuous wave Doppler, tissue velocity Doppler, and color flow Doppler.

There are several methods to increase the frequency range of PW Doppler. First, the baseline can be adjusted, which corresponds to a change in the measured frequency range. This can be useful in cases where the flow is mostly positive or negative, as shown in Figure 2.14. Another method to increase the frequency range is high pulse repetition frequency Doppler. In this mode a new pulse is emitted before the echos from the previous pulse have returned to the probe. This increases the measurable frequency range at the expense of adding some depth ambiguity.

Tissue Velocity Doppler is pulsed wave spectral Doppler with the cursor positioned over tissue rather than a region of blood flow.

Continuous Wave (CW) Doppler is the other mode of spectral Doppler imaging. In CW Doppler pulses are constantly emitted and received. Since the sampling rate is much higher, CW Doppler can measure much higher frequencies. However, there is depth ambiguity since it is unclear to the receiver where the echo was reflected.

While spectral Doppler shows a frequency spectrum over time at a specific region of interest, **Color Flow Doppler** can be used to display velocities across an area. Color Doppler is a pulsed wave technique where the beam is swept across the heart (like in B-mode imaging), and the frequency for each returning

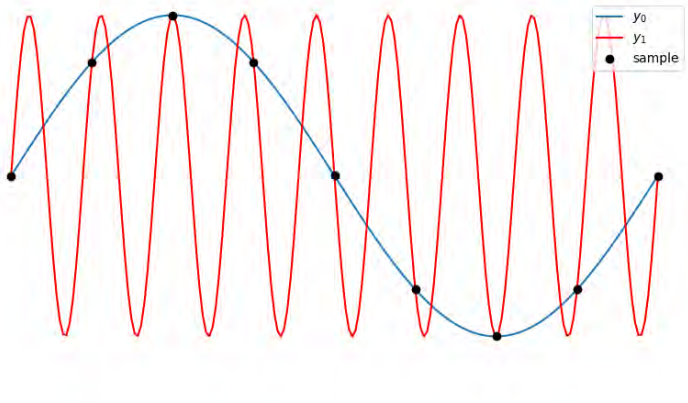


Figure 2.13: Aliasing occurs when the sampling rate is lower than the Nyquist limit. For example, given the black sampling points, either y_0 or y_1 would be a valid measured frequency.

echo is measured. The intensity information from the echo is used to generate the B-mode, while the frequencies are used to generate a color overlay corresponding to the velocity. Red is used to indicate flow towards the probe and blue shows flow away from the probe. The velocities can be filtered so that only those relevant to blood flow are displayed. Color Flow Doppler can also be applied to tissue rather than blood (Color Tissue Doppler).

2.3.5 Relevant views

During an echocardiography exam there are a set of standard B-mode views of the heart that are typically used for analysis. Views are primarily categorized by a) probe type, b) the location of the probe ("window"), c) the imaging axis of the heart, and d) the feature focused on in the image.

There are three groupings of probes. Transthoracic echocardiography (TTE) probes are used to image the heart externally from the surface of the chest and are the most commonly used probe type in echocardiography, and the focus of this thesis. TTE probes are smaller with respect to probe footprint than ultrasound probes from other applications because ultrasound signals must be sent and received through gaps in the ribs. Alternatively, transesophageal echocardiography (TEE) probes are long and flexible and made to be inserted into the patient's esophagus. TEE provides better image quality in many cases, but requires an invasive procedure, which may cause some discomfort to the patient. As such, it is avoided when possible. Finally, in intracardiac echocardiography (ICE) a miniature probe is inserted within a catheter tip within the heart. This modality is used to guide surgeons during interventional procedures.

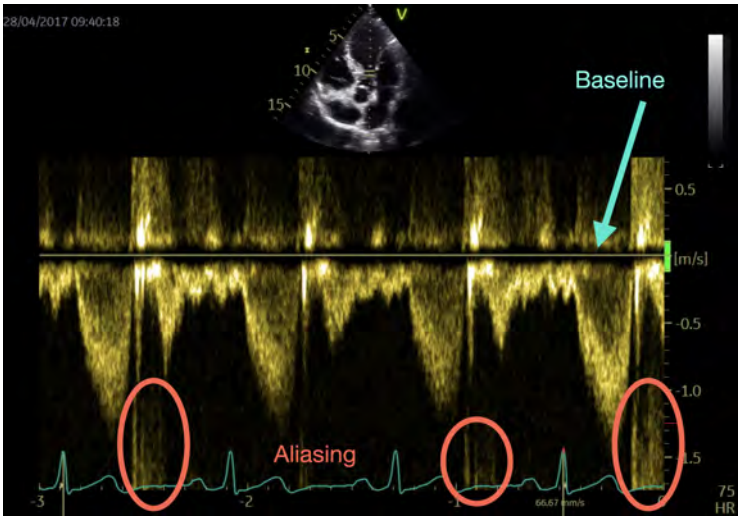


Figure 2.14: The baseline is shifted higher in this pulsed wave Doppler spectral image to avoid aliasing in the bottom part of the spectrum which will be measured. This causes some aliasing to occur in the top part of the spectrum.

There are four primary windows for TTE probes, as shown in Figure 2.15. An overview of the relevant views from each window is given below, and shown in Figure 2.16.

2.3.5.1 Apical window

The apical window can be used to analyze the structure and function of all four chambers. Together the apical views are used for measuring ejection fraction and advanced functions such as strain.

- **Apical four chamber (A4C):** The "home" view. The probe is oriented along the four-chamber plane and the interventricular septum bisects the image. The probe can also be tilted and the image zoomed to focus on either the left (**A4C - LV**)or right ventricle (**A4C - RV**) for more a detailed analysis of those chambers.

- **Apical two-chamber (A2C):** Accessed by rotating the probe 60 degrees counterclockwise from the A4C view. This view focuses specifically on the left ventricle and atrium.

- **Apical long-axis (A-LAX):** Accessed by rotating the probe 60 degrees counterclockwise from the A2C view. This view gives another view of the left ventricle and atrium, including the left ventricle outflow tract.

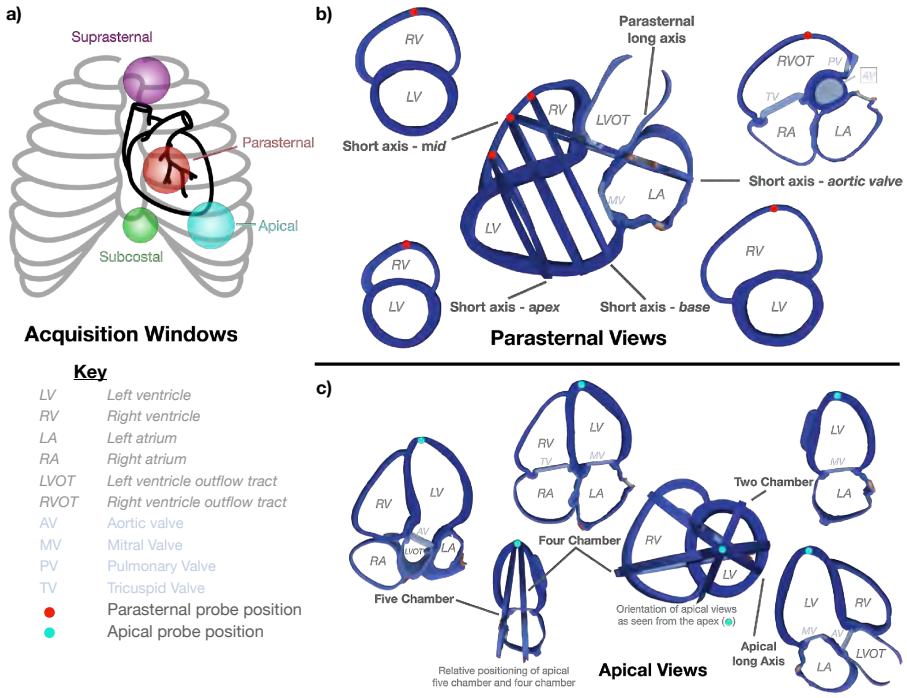


Figure 2.15: Echocardiography imaging windows and planes. a) The primary windows for transthoracic echocardiography: suprasternal, parasternal, apical, and subcostal. b) The imaging planes seen from the parasternal window. c) The imaging planes seen from the apical window.

- **Apical five-chamber (A5C):** Accessed by tilting the probe ventrally from the A4C view to include the aortic valve and left ventricle outflow tract (the fifth "chamber").

2.3.5.2 Parasternal window

The parasternal window is useful for measuring dimensions since many features are oriented perpendicular to the direction of the ultrasound waves (and axial resolution is much higher than lateral resolution). It can also be useful for analyzing structures in the anterior of the heart such as the mitral and tricuspid valves.

- **Parasternal long-axis (PLAX):** A long-axis view from the parasternal window. Most commonly used to focus on either the left ventricle or the left atrium/left ventricle outflow tract.
- **Parasternal short-axis (PSAX):** Accessed by rotating the probe 90 degrees from the PLAX view. A PSAX image can bisect the heart at the

apex of the left ventricle (**PSAX-AP**), the middle of the left ventricle at the level of the papillary muscles (**PSAX-Mid**), the base of the left ventricle (**PSAX-LV**), the mitral valve (**PSAX-MV**), or the aortic valve (**PSAX-AV**).

2.3.5.3 Subcostal window

The subcostal window provides an alternative view of many of the same imaging planes mentioned above, without the obstruction of bones or lungs. In particular, the subcostal window provides a better view of the right side of the heart, including the vena cavae.

2.3.5.4 Suprasternal window

The suprasternal window is less frequently used, but provides a view of the aortic arch.

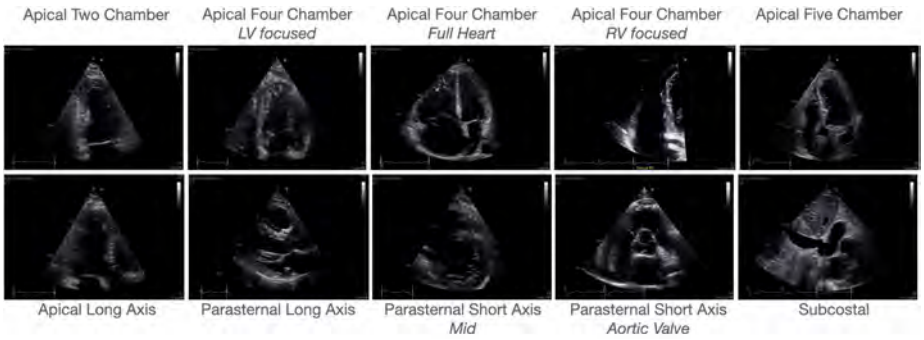


Figure 2.16: The most commonly used views in echocardiography.

2.3.6 Trade-offs

The quality of an echocardiography image will depend on several factors. The lateral resolution (resolution across the width of the image) is based on the number of scan lines and spacing between each scan line. The lateral resolution will also vary at different depths due to the widening of the sector, as well as beam aperture and focusing characteristics. The axial resolution (resolution in the direction of the beam) depends on the frequency of the ultrasound. Higher frequencies of transmission will give a better axial resolution, but higher frequency waves will not be able to penetrate as deeply into tissue.

The temporal resolution (or pulse repetition frequency) is mainly limited by the speed of sound (1540 m/s in the heart) and the width and depth of the desired image. This trade-off between spatial and temporal resolution is shown in Equation (2.4), where t is the time per frame, d is the desired depth, n is the number of scan-lines, and c is the speed of sound. As an example, to achieve 40

frames per second with 128 scan lines, the depth should be no greater than 15 cm.

$$t = \frac{2 * d * n}{c} \tag{2.4}$$

2.4 Echocardiography analysis: workflows and automation

Accurate assessment of cardiac structure, function, and hemodynamics is essential for accurate diagnosis and treatment of patients suffering from cardiovascular disease [71]. A thorough assessment requires (a) a large set of images from various views and modes, (b) measurements of important parameters within each image, and (c) a diagnosis of the patient's condition based on the images, measurements, and other patient characteristics such as age, medical history and genes.

Current guidelines recommend well over 100 different images and measurements for a basic echocardiography examination, with more required for investigating specific chambers and pathologies [72]. Moreover, new measurements are continually being developed by the clinical community to better assess patient health. This extensive protocol requires automated tools to enable a high-quality assessment of each patient. The advanced interpretation capabilities of deep learning offer the chance to automate many parts of this workflow and free up clinicians' time to focus on interpretation and advanced analyses [73]–[75].

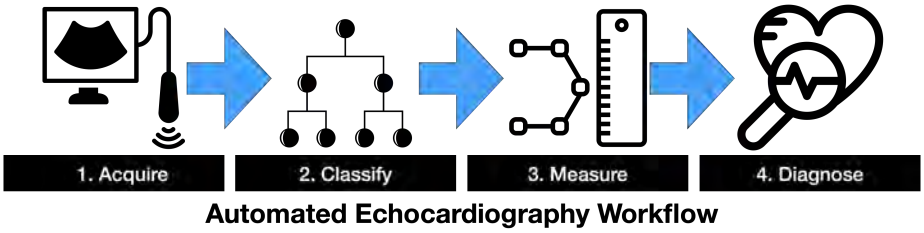


Figure 2.17: The main steps in the echocardiography workflow: acquisition of the images, classification into different image types, measurement of parameters, and analysis/diagnostics.

Figure 2.17 provides an overview of the echocardiography workflow and the following sections review each part, as well as how deep learning can be used to automatically perform the work. Note that this thesis specifically focuses on automating the classification (Section 2.4.2) and measurement (Section 2.4.3) parts of the workflow.

2.4.1 Acquisition

The first step in the echocardiography workflow is acquiring a set of images to analyze. Guidelines document the set of important views to acquire (see

Section 2.3.5), although it may vary between clinics and depend on the history and diagnosis of the patient.

During acquisition, deep learning can provide estimates of image quality [76], foreshortening detection [77], or guidance [78]. Quality assessments are useful to ensure that accurate measurements can be obtained from the acquired images. Foreshortened images will also change the measurements (see Section 2.4.5), so automated foreshortening detection can prevent these errors. While foreshortening detection so far has focused on the left ventricle, foreshortening can also significantly effect measurements of the left atrium [79] and right ventricle [80]. Guidance can help novice users find the correct view planes, and speed up the acquisition process.

2.4.2 Classification

The second step in an echocardiography workflow is to categorize each image to determine which measurements should be applied. An expert user can typically perform this action quickly, but since classification must be performed for every image, the time-savings from an automated workflow add up.

Deep learning has primarily been applied to view classification of B-mode images [73], [81], [82]. B-mode images are prioritized since M-mode images are no longer recommended for most measurements, and are only used in specialized cases to evaluate high-velocity tissue movements and volumetric assessments of flow [83].

Extending classification to include Doppler images was a part of this thesis. Paper I describes a proposed method to automate classification of Doppler spectra to enable automated measurement workflows for Doppler imaging.

2.4.3 Measurement

The third step in the echocardiography workflow is measurement of important parameters from each image. Typically, this consists of structural measurements and functional measurements. Structural measurements include dimensions, areas, or volumes, while functional measurements include velocities of wall motion, assessments of blood flow (from Doppler), or advanced quantification from speckle tracking. Speckle tracking echocardiography quantifies the movement of specific tissue regions over time by tracking each region through the distinctive pattern of speckles. Speckle tracking can be used to quantify the strain (degree of deformation) and strain rate of different cardiac muscles. As shown in Figure 2.18, basic structural and functional measurements can be combined to derive additional parameters [72].

Thus far, measurement automation has primarily focused on area measurements of the left ventricle, because an area estimate can provide a measure of ejection fraction as well as the initial region of interest necessary for left ventricle strain quantification. Typically, the methods for area estimates have consisted of a segmentation of the left ventricle in one or more apical views [73], [77], [84]–[86]. Automated segmentation methods have been shown to correlate well

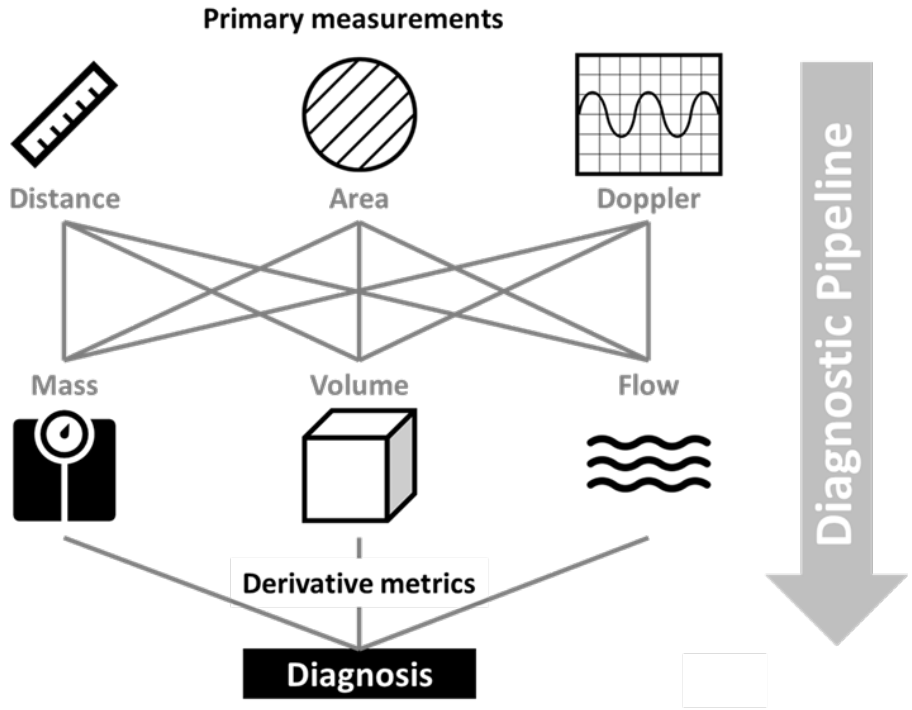


Figure 2.18: The measurement pipeline consists of primary dimension, area, and Doppler measurements. From these, derivative structural metrics (e.g. mass/volume) and functional metrics (e.g. blood flow/ejection fraction) are extrapolated and used to build a diagnosis. Advanced measurements (e.g. strain) and alternative factors (e.g. age/patient history) also contribute to the diagnosis.

with manual measurements [87]. Some groups have directly estimated ejection fraction without performing a segmentation to find volumes [75], which reduces possible sources of error, but operates as a black box for the user.

Other methods have focused on the automation of spectral Doppler measurements by segmenting the envelope of the spectrum, either through deep learning [88]–[90] or traditional methods (see [41]). Segmentation of the mitral and aortic annuli has also been automated with high accuracy for flow measurements and valve replacement planning [91]–[94].

Although an important part of the analysis workflow, 2D measurements have been automated to a lesser extent. Three recent review articles on machine learning in echocardiography did not include a section on 2D dimension measurements [41], [95], [96], and Zhang et al.'s fully automated echocardiography pipeline ignored automation of 2D measurements [73].

There are several reasons for this. First, there is a higher inter-observer error in 2D measurements than most others [97], increasing the difficulty of

accurate automation. Second, more data is typically required for training a 2D measurement network since there is less information in each sample. Third, 2D measurements often have significant variations in the measurement method. 2D measurements can be performed in B-mode or M-mode depending on operator preference[2] and measurement location can also significantly vary depending on the presence of pathologies. For example, the presence of basal septal hypertrophy can significantly effect the measurement process for interventricular septum measurements of the left ventricle [72].

As shown in Figure 2.19, basal septal hypertrophy (also known as sigmoid septum) is a localized thickening of the upper septal region. The methodology for performing intraventricular septum measurements must be adjusted based on this pathology. While the measurement is typically done at the level of the mitral valve leaflets, it should be moved more apically given a sigmoid septum. Encoding this type of pathological knowledge in an automated algorithm is difficult. Basal septal hypertrophy is relatively common in elderly patients (2-6% in patients 65-85 and 17% in patients 85+ [98]) so changes in measurement protocol are relatively common.

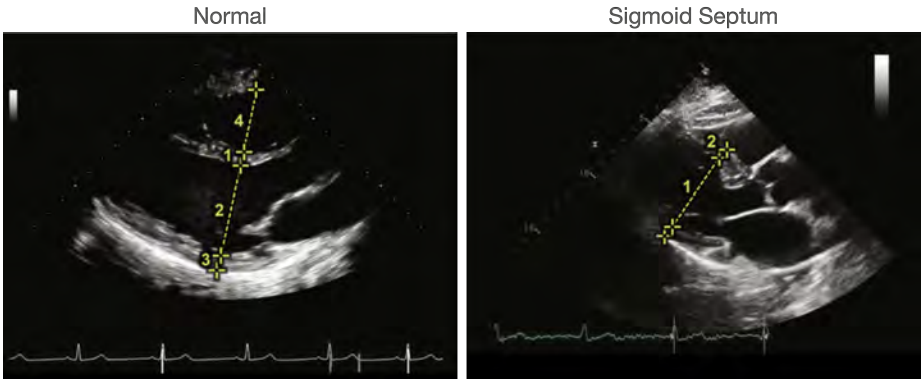


Figure 2.19: Measurements of left ventricle dimensions in the parasternal long axis vary depending on the presence of ventricular septal hypertrophy (also known as a sigmoid septum). In the normal case (left) the measurements should be conducted at the level of the mitral valve leaflets. With a sigmoid septum (right) measurements should be placed more apically. Images from [72].

This thesis focuses on adding automation of 2D measurements to the echocardiography workflow. Paper II demonstrates a method for measurements of left ventricle dimension measurements, while Paper III exhibits a new measurement for diagnosing basal septal hypertrophy.

[2]Although the guidelines recommend B-mode, many users still perform the measurements in M-mode because of habit.

2.4.4 Diagnostics

Automated diagnostics have typically followed a decision-tree approach by developing machine learning classifiers which can discover patterns between a wide range of parameters both from the B-mode and Doppler echocardiography data and patient characteristics [99]–[105]. These techniques are capable of simultaneous encapsulation of many different types of data to provide pathways for automatically flagging abnormal cases for additional review.

More recently, unsupervised deep learning has expanded the capabilities of diagnostics [106]–[111]. One strength of these unsupervised techniques is that they analyze patients on a spectrum, which better models the continuum of diseases in real life. In addition, these methods offer better opportunities to integrate parameters from many different data sources and the discovery of novel patterns through the analysis of high-dimensional data [112].

2.4.5 Challenges

Numerous challenges hinder the high-quality acquisition and analysis of echocardiography images as described in the following subsections.

2.4.5.1 Challenges for acquisition

Interference from ribs and dampening from other structures often inhibits proper visualization. The anatomy and relative positioning of every heart varies. For example, the heart is oriented more vertically in thinner patients. Additionally, the heart is beating and aside from the natural expansion/contraction that composes the beating motion, the heart translates and rotates through the cycle. These motions mean a 2D ultrasound plane will image different cardiac structures at different points in the cardiac cycle, and it is very difficult to optimize the view across the cycle. There is no acquisition protocol that works for all patients and sonographers often must be creative in probe positioning and angling to acquire images. Incorporating this knowledge into automated acquisition algorithms remains a significant challenge.

Foreshortening is also a common problem in echocardiography acqusition, where the 2D imaging plane does not bisect the largest part of the chamber. From a 2D acquisition, determining whether foreshortening occurs requires a careful analysis of the structures and motion. An example foreshoretened image of a parasternal acquisition is shown in Figure 2.20. A foreshortened image can lead to errors down the pipeline in measurement and diagnosis.

2.4.5.2 Challenges for automated analysis

Deep learning is notoriously sensitive to variations in the input space that have not been observed during training [113]. Indeed, some studies have shown neural networks are significantly more biased by the texture of images than the structure [114], although this effect varies depending on the dataset and augmentations used [115]. Changes in the input data that were not observed during training

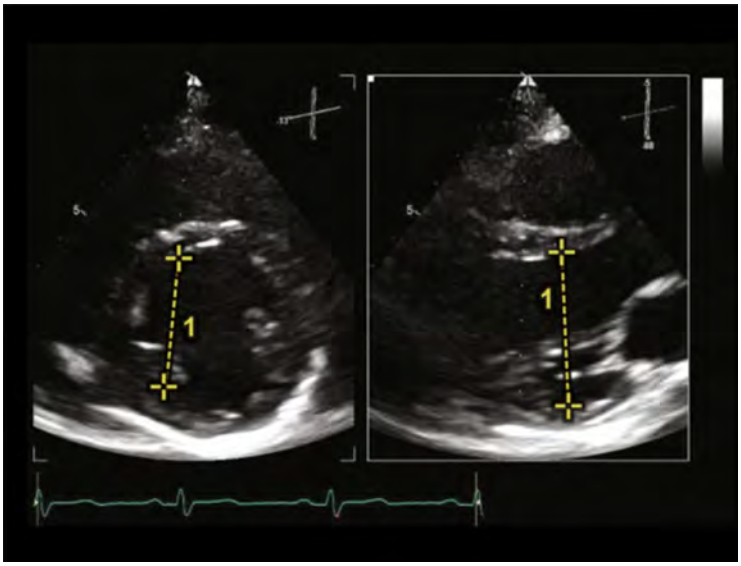


Figure 2.20: Foreshortening leads to errors in analysis. The PLAX image (right) is not positioned to bisect the largest part of the ventricle as demonstrated by the measurement shown in the bi-plane PSAX image (left). This leads to measurement errors, as the size of the ventricle is underestimated. Image from [72].

is referred to as under-specification. Under-specification can cause issues in medical imaging domain since there is variation in image textures at many levels; between different vendors, between machines produced by the same vendor, and between different imaging settings on the same machine. Indeed, significant performance drops were demonstrated when algorithms trained on one dataset were tested on a dataset from a different vendor/clinic [116].

The effects of under-specification were demonstrated with several datasets in Paper IV. Paper I used a test dataset from a separate clinic from the training set and a small drop in accuracy was observed due to differences in acquisition practices (see Paper I for details).

Echocardiography can be a powerful imaging and analysis technique, but there are opportunities to enhance this power by increasing the reproducibility and efficiency with deep learning.

Chapter 3

Summary of Contributions

This chapter provides a brief overview of the motivation and outcomes of the publications included in this thesis. Figure 3.1 shows how each of the papers fits within the context of the echocardiography and deep learning workflows described above.

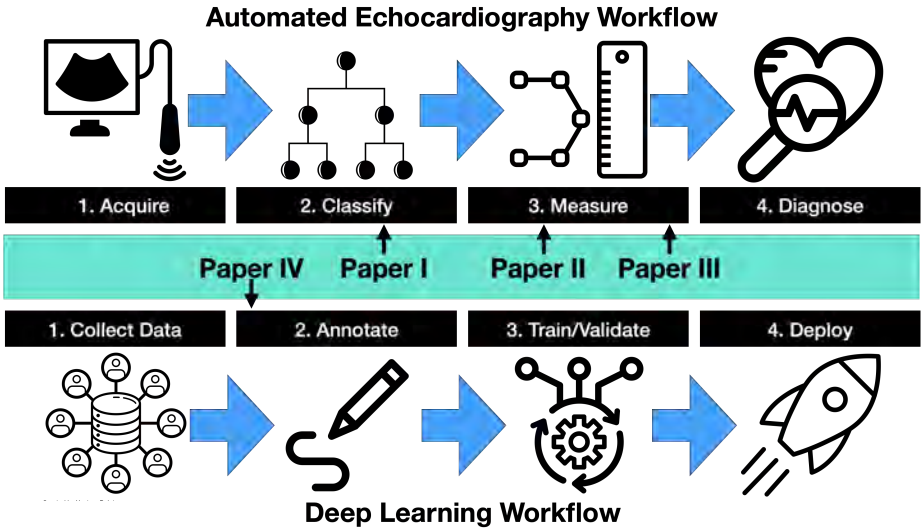


Figure 3.1: The contributions of this thesis within the spectrum of echocardiography and deep learning workflows. Paper I demonstrates automated classification and Paper II shows automated measurements. Paper III focuses on measurements as well, but from the perspective of the improved diagnostic value of a new measurement. Finally, Paper IV automates part of the data collection and annotation pipeline for creating new deep learning algorithms within echocardiography

3.1 Publications

Paper I: User-Intended Doppler Measurement Type Prediction Combining CNNs with Smart Post-Processing

Andrew Gilbert, Marit Holden, Line Eikvil, Mariia Rakmail, Aleksandar Babić, Svein Arne Aase, Eigil Samset, and Kristin McLeod, *Journal of Biomedical and Healthcare Informatics*, 2020.

As discussed in Section 2.3.4, spectral Doppler measurements are an important part of the standard echocardiography examination. These measurements give insight into myocardial motion and blood flow, providing clinicians with parameters for diagnostic decision making. Many of these measurements are performed automatically with high accuracy, increasing the efficiency of the diagnostic pipeline. However, full automation is not yet available because the user must manually select which measurement should be performed on each image.

In Paper I, we developed a pipeline based on convolutional neural networks (CNNs) to automatically classify the measurement type from spectral Doppler scans. The proposed algorithm enables a fully automatic pipeline from acquisition to Doppler spectrum measurements. We achieved 96% accuracy classifying 18 measurement types on a test set drawn from separate clinical sites. In the same way that view recognition enables automatic processing of B-mode images, Doppler spectrum classification enables automated inference of which measurements should be applied.

The principle contributions included demonstrating how multi-modal information in each spectral Doppler recording can be combined using a meta-parameter post-processing mapping scheme and heatmap-encoding at the input of the network to include coordinate locations. Additionally, the effects of network architecture and ensemble networks were explored to examine the trade-off between accuracy, speed, and memory usage for resource-constrained environments. Finally, a confidence metric was developed using the values in the last fully connected layer of the network and we showed the confidence metric can prevent many misclassifications.

Paper II: Automated Left Ventricle Dimension Measurement in 2D Cardiac ultrasound via an Anatomically Meaningful CNN Approach

Andrew Gilbert, Marit Holden, Line Eikvil, Svein Arne Aase, Eigil Samset, and Kristin McLeod, *Smart Ultrasound Imaging workshop at MICCAI 2019. Lecture Notes in Computer Science.*

Two-dimensional echocardiography measurements of the left ventricle are highly significant markers of several cardiovascular diseases and are often used in clinical care, despite suffering from large variability between observers. This variability is due to the challenging nature of accurately finding the correct temporal and spatial location of measurement endpoints in ultrasound images. These images often contain blurry boundaries and varying reflection patterns between frames.

In Paper II, we presented a convolutional neural network-based approach to automate left ventricle measurements. Treating the problem as a landmark detection problem, we proposed a modified U-Net CNN architecture to generate heatmaps of likely coordinate locations. Results showed 13.4%, 6%, and 10.8%

mean percent error on intra-ventricular septum, left ventricle internal dimension, and left ventricle posterior wall measurements respectively. These results match, or approach intra-observer expert error.

While previous works had focused on either septum or internal dimension measurement, this paper demonstrated accurate measurement of three dimensions. The additional challenge of achieving results that are logical with respect to the other measurements makes this a more difficult task than independent measurements. In addition, principle contributions included demonstration of coordinate convolution to better localize points, the use of anatomically meaningful heatmaps as labels, and a multi-component loss function which optimizes for landmark location as well as measurement size and visual appearance.

Paper III: Septal Curvature as a Robust and Reproducible Marker for Basal Septal Hypertrophy

Maciej Marciniak, **Andrew Gilbert**, Filip Loncaric, Joao F. Fernandes, Bart Bijnens, Marta Sitges, Andrew P. King, and Pablo Lamata, *Journal of Hypertension*, 2021.

Basal septal hypertrophy is an asymmetric, localized thickening of the upper interventricular septum, and constitutes a marker of an early remodelling in patients with hypertension. This morphological trait has been extensively researched because of its prevalence in hypertension, yet its clinical and prognostic value for individual patients remains undetermined. One of the reasons is the lack of a reliable and reproducible metric to quantify the presence and the extent of BSH. Paper III proposed the use of the curvature of the left ventricular endocardium as a robust feature for basal septal hypertrophy characterization, and as an objective criterion to quantify the degree of sigmoidal septum which is currently mostly subjectively analyzed via visual assessment.

Robustness and reproducibility were assessed on a cohort of 220 patients, including 161 hypertensive patients (32 with BSH) and 59 healthy controls. Results showed that compared with the conventionally used wall thickness metrics, the new marker was more reproducible (relative standard deviation of errors of 7% vs. 13%, and 8 vs. 38% for intra-observer and inter-observer variability, respectively). The correlation between the new marker and functional parameters related to basal septal hypertrophy were also better than wall thickness, with the main difference being in local deformation changes assessed by longitudinal strain (absolute rank correlation 0.417 vs. 0.341) .

The primary contribution was the demonstration of the proposed marker, called average septal curvature. Average septal curvature is defined as the inverse of the radius adjacent to each point of the endocardial contour along the basal and mid inferoseptal segments of the left ventricle. Average septal curvature is a more precisely defined and reproducible metric than thickness ratios, it can be fully automated, and better infers the functional remodelling.

Paper IV: Generating Synthetic Labeled Data from Existing Anatomical Models: An Example with Echocardiography Segmentation

Andrew Gilbert, Maciej Marciniak, Cristobal Rodero, Pablo Lamata, Eigil Samset, and Kristin McLeod, *Transactions on Medical Imaging*, 2021.

As demonstrated in Paper I and Paper II, deep learning can bring time savings and increased reproducibility to medical image analysis. However, acquiring training data is challenging due to the time-intensive nature of labeling and high inter-observer variability in annotations.

Rather than labeling images, in Paper IV we proposed an alternative pipeline where images were generated from existing high-quality annotations using generative adversarial networks. The annotations were derived automatically from anatomical models and were transformed into realistic synthetic ultrasound images with paired labels using a CycleGAN.

We demonstrated the pipeline by generating synthetic 2D echocardiography images to compare with existing deep learning ultrasound segmentation datasets. A convolutional neural network was trained to segment the left ventricle and left atrium using only synthetic images. Networks trained with synthetic images were extensively tested on four different unseen datasets of real images with median Dice scores of 91, 90, 88, and 87 for left ventricle segmentation. These results matched inter-observer results measured on real ultrasound datasets and are comparable to a network trained and tested on sets of real images from different datasets. Results demonstrated the images produced can effectively be used in place of real data for training.

In addition to the proposed pipeline and the demonstration on left ventricle segmentation, contributions included a thorough analysis of sources of error in segmentation including differences in shape, texture, and annotator style. The proposed pipeline opens the door for automatic generation of training data for many tasks in medical imaging as the same process can be applied to other segmentation or landmark detection tasks in any modality.

Appendix A: The "Digital Twin" to enable the vision of precision cardiology

Jorge Corral-Acero, Francesca Margara, Maciej Marciniak, Cristobal Rodero, Filip Loncaric, Yingjing Feng , **Andrew Gilbert**, Joao F Fernandes, Hassaan A Bukhari, Ali Wajdan, Manuel Villegas Martinez, Mariana Sousa Santos, Mehrdad Shamohammdi, Hongxing Luo , Philip Westphal, Paul Leeson, Paolo DiAchille, Viatcheslav Gurev, Manuel Mayr, Liesbet Geris, Pras Pathmanathan, Tina Morrison, Richard Cornel, Frits Prinzen, Tammo Delhaas, Ada Doltra, Marta Sitges, Edward J Vigmond, Ernesto Zacur, Vicente Grau, Blanca Rodriguez, Espen W Remme, Steven Niederer, Peter Mortier, Kristin McLeod, Mark

Potse, Esther Pueyo, Alfonso Bueno-Orovio, and Pablo Lamata, *European Heart Journal*, 2020.

Providing therapies tailored to each patient is the vision of precision medicine, enabled by the increasing ability to capture extensive data about individual patients. In Appendix A, a position paper from the Personalized In-silico Cardiology consortium (see Section 1.3), we argued that the second enabling pillar towards this vision is the increasing power of computers and algorithms to learn, reason, and build the 'digital twin' of a patient. Computational models are boosting the capacity to draw diagnosis and prognosis, and future treatments will be tailored not only to current health status and data, but also to an accurate projection of the pathways to restore health by model predictions. The early steps of the digital twin in the area of cardiovascular medicine were reviewed, together with a discussion of the challenges and opportunities ahead. We emphasized the synergies between mechanistic and statistical models in accelerating cardiovascular research and enabling the vision of precision medicine.

3.2 Innovations

Work from two of the articles presented in this thesis was integrated into the GE Healthcare software ecosystem used in echocardiography machines worldwide. The methods from the other two papers were released as open source software packages.

3.2.1 Automatic measurement in clinical software

Results from Paper I and Paper II were included as a part of the new Vivid Ultra Edition software release from GE Healthcare. AI Auto Spectrum Recognition[1] was based on the work presented in Paper I. AI Auto Measure 2D[2] was based on the work presented in Paper II with the addition of the confidence metrics described in Appendix B.

Both of these tools were received positively in early clinical use. Feedback from customers on AI Auto Spectrum Recognition included: "AI does an amazing job", "HUGE TIMESAVER", "much more accurate and easier to use than any other vendor" and "[with this tool we] may even have time for lunch". Feedback on the AI Auto Measure 2D included: "extremely happy with this new tool", "worked extremely well with little need to manipulate measurements", and "[AI Auto Measure 2D] will revolutionize [our] workflow".

3.2.2 Open source software packages

The sources for the methods presented in Paper III and Paper IV were released as open software packages. The repository for Paper III[3] contains tools for

[1] Demo available: https://gevividultraedition.com/iq/auto-measure-sr
[2] Demo available: https://gevividultraedition.com/e95/auto-measure-2d#
[3] https://github.com/MaciejPMarciniak/curvature

calculating the curvature of a line, surface and 3D mesh. The repository for Paper IV[4] contains tools for extending a set of anatomical models using principle component analysis, extracting images from the models, transforming to realistic ultrasound images using GANs, training/testing segmentation networks, and links to anatomical models that can be used for training.

3.3 Patents

Two patent disclosures were filed as a part of the work in this thesis.

Method of Performing Automated Measurements Over Multiple Cardiac Cycles

Andrew Gilbert, Gunnar Hansen, Svein Arne Aase, and Andreas Heimdal, 2020.

Automated echocardiography measurement systems enhance measurement reproducibility by combining measurements across multiple cardiac cycles and images. This patent disclosure describes a method for using deep learning confidence metrics and traditional statistical analysis to combine measurements across cycles into a global result and measure deviation across cycles. It also describes methods to pick the optimal cycle and measurement to show the user.

Systems and Methods for Adaptive Measurement of Medical Images

Andrew Gilbert, 2021.

This disclosure presents a method for adapting automated systems to user input to adapt to variations in user preference. In a measurement with many degrees of variability, such as a measurement of multiple dimensions or a segmentation, a user may want to override one part of the measurement. This disclosure describes approaches to adapt the remaining points to user corrections using heatmap encoding.

[4] https://adgilbert.github.io/data-generation/

Chapter 4

Discussion

In this thesis, methods have been presented to improve workflow and measurements in echocardiography. Additionally, systems were developed for simplifying future automation projects through the automated generation of synthetic training data. In this section we present important takeaways from this work as well as areas of future improvement.

4.1 Automation in clinical workflows: observations from applying deep learning

The primary goal of automation is to increase the efficiency and reproducibility of analyses. Given the broad spectra of potential applications, it is important to evaluate where the most impact can be obtained in automating clinical workflows.

4.1.1 Emphasis should be placed on automating normal cases with high precision

Deep learning is fragile and may have unpredictable results on unseen inputs. Given the large spectrum of potential inputs, it is more important to tackle the normal cases rather than trying to optimize predictions across a wide range of pathologies. The diversity of cases makes it impossible to automate across the disease spectrum [117] and corner cases will need to be manually reviewed anyway. Focusing on automating the normal cases with high precision will lead to high-impact tools that can increase the usability and repeatability of echocardiography examinations.

Rather than trying to handle all corner cases, automated tools can help flag potentially troublesome cases for manual review. Confidence metrics attached to the output of a network help achieve this. Sample confidence metric methods are presented in Paper I and Appendix B. Confidence measurements can help reduce uncaught errors and highlight improvement areas for automated tools.

4.1.2 Deep learning is a tool rather than a replacement for cardiologists

The rise of AI and deep learning has caused some prominent experts to claim that radiologists will be "obsolete" soon. However, as discussed above, deep learning is still far away from being able to handle the wide variety of cases seen in real life. While powerful, deep learning cannot replicate the nuanced view a cardiologist has. Moreover, there are significant practical problems to solve when implementing an automated system (e.g. [118]). Rather, medical image analysis

will continue to evolve with deep learning serving primarily as an assistive tool rather than a replacement for human interpretation. The combination of human and machine intelligence will enable more advanced analyses.

4.1.3 Perceived accuracy may be more important than measured accuracy

For clinical acceptance, perceived accuracy is often more important than real accuracy. For example, the angles and relative positioning of measurement calipers in Paper II were just as important as the accuracy when judged by experts. For example, Figure 4.1 demonstrates an example set of measurements which make sense individually, but not collectively. This means that clinical experts should be involved throughout the process when determining evaluation metrics and loss functions for new automated tools.

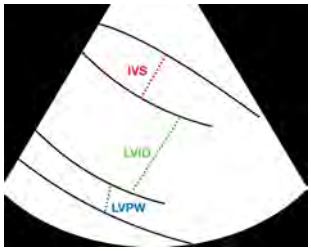


Figure 4.1: The example interventricular septum (IVS), left ventricle internal dimension (LVID), and left ventricle posterior wall (LVPW) measurements shown above are relatively accurate individually, but may be accepted collectively due to the differences in angle and positioning.

4.1.4 Network architecture does not significantly affect accuracy

One takeaway from the work presented in this thesis is that network architecture typically does not play a significant role. There are architectures that perform well for specific tasks (e.g. U-Net [17] for segmentation and ResNet [18] for classification) but accurate results can be achieved with variations on the broad themes proposed by these architectures. However, within the field there is often an over-emphasis on the effects of architecture. Network architecture experiments were conducted in Paper I and Paper II and showed only minor differences in performance. Many of the accuracy changes from varying architecture can be attributed to hyperparameter optimization, and Isensee et al. demonstrated that architectural hyperparameter optimization could be automated to achieve better accuracy [48].

4.1.5 Domain-specific adaptations are critical for success

However, the relative insignificance of network architecture does not mean that problems in medical imaging can be solved with by simply plugging in pretrained networks. A detailed understanding of the problem is required for accurate and robust application of deep learning to medical imaging. This often includes domain-specific adaptations to the input/output layers of the network, the structure of the loss function, or the setup of the proposed problem for the network to solve. For example, many tasks could be framed as either a landmark detection or a segmentation, and these approaches offer different trade-offs. As another example, the domain-specific adaptations at the input and the output of the network presented in Paper I achieved higher accuracy (96.4% vs 91.6%) with a much larger class set (18 classes vs. 3 classes) than other works which relied on fine-tuning networks pretrained on ImageNet [89].

4.1.6 Measurement automation is more important than automated diagnostics

The primary goal of echocardiography analysis is to diagnose a patient and create a treatment plan. This has traditionally been done by creating a set of measurements and evaluating those compared to normal ranges of values (an "evidence-based" diagnosis). However, there are several drawbacks to performing measurements:

- Measurements introduce a new source of potential error in the diagnostic pipeline, and as described in Chapter 2 (and demonstrated in Paper II, Paper III, and Paper IV) the variability in these measurements can be substantial.
- Some works have also shown high accuracy of direct visual assessment of parameters such as ejection fraction without needing to measure [119], [120], although others have found visual assessment to be less reliable [87]. Feedback from clinicians on automated measurement algorithms has also suggested that automatic direct visual assessment may eliminate the need for performing the measurements.
- Early work on algorithms that can jump directly from imaging (with or without other data) to diagnoses has been relatively successful [73], [93], [121]–[124].

The successes of direct diagnostic prediction and the shortcomings of measurements (automated or manual) beg the question of whether algorithms that jump directly to a diagnosis are a more impactful area of research than focusing on automating measurements. The progress of other deeper and deeper networks in other fields such as the GPT model in language modeling [125]–[127] demonstrates that perhaps the only obstacles standing in the way of this vision are larger datasets and more computing resources. The allure of completely

automated processing from images to a disease classification is appealing both to the clinicians tasked with performing copious measurements on every exam and the engineers attempting to automate those measurements.

However, even in this idealized vision of a diagnostic future, traditional measurements will have a role to play. While informative, advanced deep models are susceptible to including biases found in the initial data (e.g. racial/gender biases in [127]). Measurements offer one road for finding these biases and methods for correcting them.

Additionally, automated detection of low-level parameters is more tedious to perform than interpretation [128], [129]. Clinical feedback to GE Vingmed reports decreased variability in measurements and reduction in tedious manual tasks as the top priorities for future tools. This feedback is reflected in reports on demand, where quantitative tools make up the majority of the current and projected market for artificial intelligence tools in echocardiography [130].

Furthermore, as discussed above, deep learning algorithms are still far from being able to handle all of the corner cases seen in real patient care. Mistakes from automated diagnostics algorithms have higher consequences for patients and currently provide limited reasoning for the classifications made. Black-box solutions such as automated diagnostics also require much more significant validation trials before widespread use while automated measurements can be easily visualized and reviewed.

Measurements also offer a method for better understanding the decisions made by these models. The vision behind the development of this thesis (and the PIC consortium) is that the future of personalized cardiology relies on the simultaneous development of both a rich understanding of the mechanisms at play through mechanistic modeling and the interplay of mechanisms and correlations to outcomes through statistical modeling [5]. This vision is embodied within the digital twin where analysis of both the patients position in the population spectrum and detailed structural/functional analyses can occur. The automation of measurements increases the reliability of mechanistic modeling and the power of statistical modeling through increased accuracy and reliability.

Future workflows should rely on a combination of algorithms that can automatically provide meaningful insights about patients and those that can add the measurements and reasoning that re-enforce that insight. In this regard, algorithms that can integrate traditional measurements, markers from varying sources, and novel parameters interpreted directly from images are essential [131]. Additionally, rather than binary classifications, future algorithms should focus on encoding images and other data sources into manifolds which describe the patient's position within a spectrum of pathologies [106], [110], [111], [132]. Grading on a continuum offers a more nuanced and realistic view of how different diseases interact.

4.2 Future work

This thesis aims to contribute to the automation of measurements and workflow in echocardiography. The field is continually advancing and there remains significant new future directions to investigate.

4.2.1 Fully automatic curvature measurements

The curvature measurement approach presented in Paper III was semi-automated. The curvature measurements are built from a part of the left ventricle endocardial border, meaning curvature could be fully automated using a endocardial segmentation algorithm. We initially explored the use of deep learning segmentation tools to extract this measurement. However, as shown in Figure 4.2, we found prior segmentation networks (and the thus the training datasets) ignored basal septal hypertrophy. The networks were optimized for strain calculation and including a substantially curved endocardial border would change the longitudinal strain. This is an example of where methods for automatically adjusting annotations (such as Paper IV) could help adapt datasets to be problem-specific. Future work will focus on using these tools to fully automate the extraction of curvature measurements.

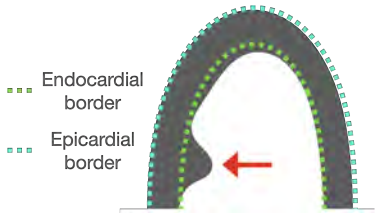


Figure 4.2: Segmentation methods trained to optimize strain measurements will ignore basal septal hypertrophy (red arrow) in the endocardial segmentation.

4.2.2 New measurements and applications

Figure 4.3 shows a current perspective of the impact, difficulty, and automation level of various aspects of echocardiography workflows. Some tools, such as view recognition and Doppler spectrum classification, have a low time-savings per use but are used frequently throughout acquisition. These tools also tend to have a higher level of current automation (see [73], [81], [82], and Paper I). Others, such as segmentation and dimension measurement are used less frequently, but have a higher time-savings per use. These tools have achieved a high level of automation for the left ventricle (e.g. [24], [45], [73], [85], [133], Paper II, and Paper IV) and the mitral valve (e.g. [94], [134]), but have not yet been thoroughly applied to other regions of the heart.

From this perspective, the greatest opportunity for new automated tools comes in two places. The first comes from automation of new dimension

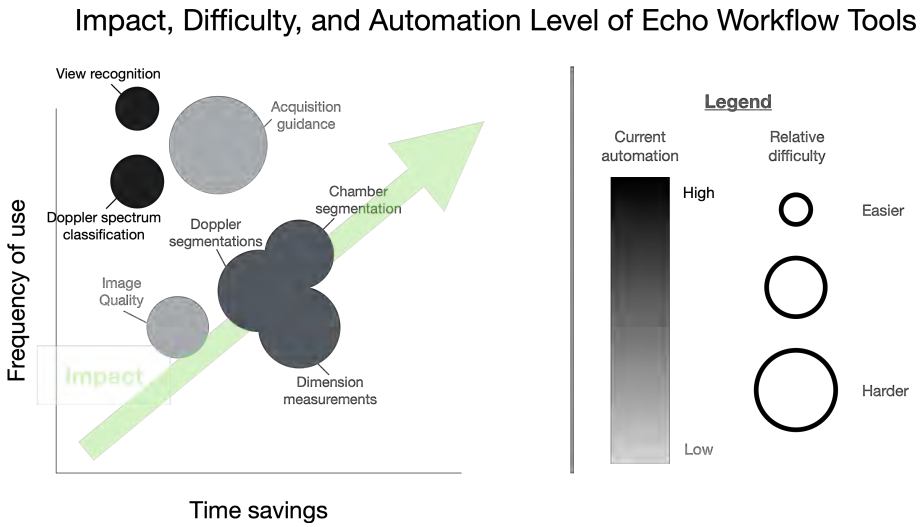


Figure 4.3: The tradeoff between the impact and difficulty of automating different parts of the diagnostic pipeline. The x-axis shows the time-savings for each use and the y-axis shows the frequency of use (so the highest impact is in the top right). The circle size shows difficulty of implementation and the shading shows the current level of automation.

measurements and segmentations for additional chambers. The approach presented in Paper II can likely be applied with high accuracy to new measurements while the pipeline presented in Paper IV can be extended to generate the necessary training data to simplify the process.

The second largest opportunity is acquisition guidance. While the time savings for expert users will be minimal, guidance can help to increase accessibility for novice users, increasing reliability and expanding the impact of echocardiography. However, it is also a difficult problem given the challenges discussed in Section 2.4.5. Accurate guidance may first require more nuanced view classification that positions acquired views more specifically within the global anatomy rather than a broadly defined class.

4.2.3 Adding the temporal dimension

Finally, one major benefit of echocardiography is the high temporal resolution. Thus far, most algorithms have not taken advantage of this attribute in their design despite the fact that many clinical observers report relying heavily on the ability to flip between frames and track structures when performing measurements. Those works that have taken advantage have relied on a variety of different approaches, including convolutional neural networks feeding into recursive networks [23], [24] or use of optical flow networks [86]. Continuing to integrate these and other new methods for temporal processing (such as

transformer networks) into the workflow for automated echocardiography will yield more accurate results and more advanced metrics.

Chapter 5

Conclusion

In this thesis, four papers have been presented to address the motivations outlined in Section 1.1, namely, to improve echocardiography workflows through deep learning-based automation. The papers presented in this thesis advance both the methodology and practical implementation of automation in clinical echocardiography. The first goal was to investigate the possibility of improving workflows in echocardiography analysis. Paper I proposed a method to automatically classify Doppler spectra to enable fully automated processing of measurements. The second goal was to increase the reliability of measurements in echocardiography. Paper II presented a method to automate measurements in 2D echocardiography while Paper III presented a novel measurement method to increase the robustness of the diagnosis of basal septal hypertrophy. The final goal was to simplify the adoption of new automated measurements. Paper IV proposed a method to automatically generate synthetic data for training new algorithms. The methods proposed above were thoroughly validated against a variety of datasets and compared to state-of-the-art techniques.

Several methods were integrated into clinically available echocardiography software. According to key economic opinion leaders, the world is now entering the fourth industrial revolution. The fourth revolution is characterized by "a fusion of technologies that is blurring the lines between the physical, digital, and biological spheres" [135]. Artificial intelligence (deep learning in particular) continues to be a key driving factor of this change across many industries, including medical imaging. However, to date many of the innovations have been constrained to the lab. This thesis helps drive forward the practical implementation of artificial intelligence in echocardiography.

The key differentiating factor from previous industrial revolutions is the exponential rate of disruption [135]. While that exponential growth has been observed with deep learning, the data requirements for training have also grown exponentially, a key limiting factor for medical imaging. We demonstrated novel methods to address this challenge and enable future automation.

Overall, the methods presented in this thesis will help to improve the standard of patient care in echocardiography through more efficient and reliable measurements, workflow, and data generation.

Bibliography

[1] Purcarea, A. *et al.*, "Cardiovascular disease risk scores in the current practice: which to use in rheumatoid arthritis?" *Journal of medicine and life*, vol. 7, no. 4, pp. 461–7, 2014.

[2] Shah, B. N., "Echocardiography in the era of multimodality cardiovascular imaging.," *BioMed research international*, vol. 2013, p. 310 483, Jun. 2013.

[3] Montavon, G., Samek, W., and Müller, K. R., "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing: A Review Journal*, vol. 73, pp. 1–15, Jun. 2018.

[4] Litjens, G. *et al.*, "A Survey on Deep Learning in Medical Image Analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[5] Corral-Acero, J. *et al.*, "The 'Digital Twin' to enable the vision of precision cardiology," *European Heart Journal*, pp. 1–11, 2020.

[6] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, 2012.

[7] Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M., "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint*, 2020. arXiv: 2004.10934.

[8] Wu, Y. *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint*, 2016. arXiv: 1609.08144.

[9] Xiong, W. *et al.*, "Achieving human parity in conversational speech recognition," *arXiv preprint*, 2016. arXiv: 1610.05256.

[10] Silver, D. *et al.*, "Mastering the game of Go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[11] Bojarski, M. *et al.*, "End to end learning for self-driving cars," *arXiv preprint*, 2016. arXiv: 1604.07316.

[12] Taghanaki, S. A. *et al.*, "Deep Semantic Segmentation of Natural and Medical Images: A Review," *Artificial Intelligence Review*, vol. 54, no. 1, pp. 137–178, 2019. arXiv: 1910.07655.

[13] Shen, D., Wu, G., and Suk, H.-I., "Deep Learning in Medical Image Analysis," *Annu. Rev. Biomed. Eng*, vol. 19, pp. 221–248, 2017.

[14] Tajbakhsh, N. *et al.*, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, 2020. arXiv: 1908.10454.

[15] Chen, C. *et al.*, "Deep learning for cardiac image segmentation: A review," *Frontiers in Cardiovascular Medicine*, vol. 7, 2020. arXiv: 1911.03723.

[16] Nair, V. and Hinton, G. E., "Rectified linear units improve restricted boltzmann machines," in *International Conference on Machine Learning*, 2010.

[17] Ronneberger, O., Fischer, P., and Brox, T., "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.

[18] Zhang, X., Ren, S., and Sun, J., "Deep Residual Learning for Image Recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2016. arXiv: 1512.03385.

[19] Bianco, S. *et al.*, "Benchmark Analysis of Representative Deep Neural Network Architectures," *IEEE Access*, vol. 6, 2018. arXiv: 1810.00736v2.

[20] Liu, R. *et al.*, "An Intriguing Failing of Convolutional Neural Networks and the CoordConv Solution," in *Advances in Neural Information Processing Systems*, 2018. arXiv: 1807.03247.

[21] Yi, K. M. *et al.*, "Lift: Learned invariant feature transform," in *European Conference on Computer Vision*, Springer, 2016, pp. 467–483.

[22] Nibali, A. *et al.*, "3D human pose estimation with 2D marginal heatmaps," *Proceedings IEEE Winter Conference on Applications of Computer Vision*, no. Figure 1, pp. 1477–1485, 2019. arXiv: 1806.01484.

[23] Jahren, T. S. *et al.*, "Estimation of End-Diastole in Cardiac Spectral Doppler Using Deep Learning," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 12, pp. 2605–2614, 2020.

[24] Sofka, M. *et al.*, "Fully convolutional regression network for accurate detection of measurement points," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2017.

[25] Rumelhart, D. E., Hinton, G. E., and Williams, R. J., "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[26] Hochreiter, S. and Schmidhuber, J., "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[27] Cho, K. *et al.*, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint*, 2014. arXiv: 1409.1259.

[28] Bertsekas, D. P. and Tsitsiklis, J. N., "Neuro-dynamic programming: an overview," in *Proceedings of 1995 34th IEEE conference on decision and control*, vol. 1, IEEE, 1995, pp. 560–564.

[29] Watkins, C. J. C. H., "Learning from delayed rewards," Ph.D. dissertation, University of Cambridge, 1989.

[30] Sutton, R. S., "Learning to predict by the methods of temporal differences," *Machine learning*, vol. 3, no. 1, pp. 9–44, 1988.

[31] Vaswani, A. *et al.*, "Attention is all you need," *arXiv preprint*, 2017. arXiv: 1706.03762.

[32] Goodfellow, I. *et al.*, "Generative Adversarial Nets," *Advances in Neural Information Processing Systems 27*, pp. 2672–2680, 2014. arXiv: arXiv: 1406.2661v1.

[33] Mirza, M. and Osindero, S., "Conditional Generative Adversarial Nets," 2014. arXiv: 1411.1784.

[34] Isola, P. *et al.*, "Image-to-image translation with conditional adversarial networks," *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2017.

[35] Brock, A., Donahue, J., and Simonyan, K., "Large scale GaN training for high fidelity natural image synthesis," in *7th International Conference on Learning Representations, ICLR 2019*, 2019. arXiv: 1809.11096.

[36] Wang, L. *et al.*, "A State-of-the-Art Review on Image Synthesis with Generative Adversarial Networks," *IEEE Access*, vol. 8, pp. 63 514–63 537, 2020.

[37] Kazeminia, S. *et al.*, "GANs for Medical Image Analysis," *Artificial Intelligence in Medicine*, 2020. arXiv: 1809.06222v1.

[38] Jay, F. *et al.*, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks Jun-Yan," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 183–202.

[39] Kingma, D. P. and Welling, M., "Auto-Encoding Variational Bayes," *arXiv preprint*, Dec. 2013. arXiv: 1312.6114.

[40] Tajbakhsh, N. *et al.*, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?" *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[41] Zamzmi, G. *et al.*, "Harnessing Machine Intelligence in Automatic Echocardiogram Analysis: Current Status, Limitations, and Future Directions," *IEEE Reviews in Biomedical Engineering*, 2020.

[42] Bhalodia, R. *et al.*, "DeepSSM: A Deep Learning Framework for Statistical Shape Modeling from Raw Images," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 11167 LNCS, pp. 244–257, 2018. arXiv: 1810.00111.

[43] Corral Acero, J. *et al.*, "SMOD - Data Augmentation Based on Statistical Models of Deformation to Enhance Segmentation in 2D Cine Cardiac MRI," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11504 LNCS, pp. 361–369, 2019.

[44] Shin, H. C. *et al.*, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11037 LNCS, pp. 1–11, 2018. arXiv: 1807.10225.

[45] Jafari, M. H. *et al.*, "Semi-supervised learning for cardiac left ventricle segmentation using conditional deep generative models as prior," *Proceedings - International Symposium on Biomedical Imaging*, vol. 2019-April, no. Isbi, pp. 649–652, 2019.

[46] Huo, Y. *et al.*, "Synseg-net: Synthetic segmentation without target modality ground truth," *IEEE Transactions on Medical Imaging*, vol. 38, no. 4, pp. 1016–1025, 2018.

[47] Amirrajab, S. *et al.*, "XCAT-GAN for Synthesizing 3D Consistent Labeled Cardiac MR Images on Anatomically Variable XCAT Phantoms," in *International Conference on Medical Image Computing and Computer-Assisted Intervention.*, Springer, 2020, pp. 128–137. arXiv: 2007.13408.

[48] Isensee, F. *et al.*, "Automated Design of Deep Learning Methods for Biomedical Image Segmentation," *arXiv preprint*, pp. 1–55, 2019. arXiv: 1904.08128.

[49] OpenStax *et al.*, *Anatomy & Physiology: OpenStax*. OpenStax College, 2013.

[50] Chan-Dewar, F., "The cardiac cycle," *Anaesthesia & Intensive Care Medicine*, vol. 13, no. 8, pp. 391–396, 2012.

[51] Silverman, M. E., Grove, D., and Upshaw, C. B., "Why does the heart beat? The discovery of the electrical system of the heart," *Circulation*, vol. 113, no. 23, pp. 2775–2781, 2006.

[52] Huikuri, H. V., Castellanos, A., and Myerburg, R. J., "Sudden death due to cardiac arrhythmias," *New England Journal of Medicine*, vol. 345, no. 20, pp. 1473–1482, 2001.

[53] Mayo Clinic, *Heart arrhythmia*.

[54] Coffey, S., Cairns, B. J., and Iung, B., "The modern epidemiology of heart valve disease," *Heart*, vol. 102, no. 1, pp. 75–85, 2016.

[55] Cameli, M., Mandoli, G. E., and Mondillo, S., "Left atrium: the last bulwark before overt heart failure," *Heart Failure Reviews*, vol. 22, no. 1, pp. 123–131, 2017.

[56] Gorter, T. M. *et al.*, "Right heart dysfunction and failure in heart failure with preserved ejection fraction: mechanisms and management. Position statement on behalf of the Heart Failure Association of the European Society of Cardiology," *European Journal of Heart Failure*, vol. 20, no. 1, pp. 16–37, 2018.

[57] Udelson, J. E. and Stevenson, L. W., "The future of heart failure diagnosis, therapy, and management," *Circulation*, vol. 133, no. 25, pp. 2671–2686, 2016.

[58] Ponikowski, P. *et al.*, "2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure," *European Heart Journal*, vol. 37, no. 27, pp. 2129–2200m, 2016.

[59] Roger, V. L., "Epidemiology of heart failure," *Circulation Research*, vol. 113, no. 6, pp. 646–659, 2013.

[60] Hollenberg, S. M. *et al.*, "2019 ACC Expert Consensus Decision Pathway on Risk Assessment, Management, and Clinical Trajectory of Patients Hospitalized With Heart Failure," *Journal of the American College of Cardiology*, vol. 74, no. 15, pp. 1966–2011, Oct. 2019.

[61] Zamorano, J. L. *et al.*, "2014 ESC guidelines on diagnosis and management of hypertrophic cardiomyopathy: The task force for the diagnosis and management of hypertrophic cardiomyopathy of the European Society of Cardiology (ESC)," *European Heart Journal*, vol. 35, no. 39, pp. 2733–2779, 2014.

[62] Pearson, A. C., "The evolution of basal septal hypertrophy: From benign and age-related normal variant to potentially obstructive and symptomatic cardiomyopathy," *Echocardiography*, vol. 34, no. 7, pp. 1062–1072, 2017.

[63] Pennacchini, E. *et al.*, "Distinguishing Hypertension From Hypertrophic Cardiomyopathy as a Cause of Left Ventricular Hypertrophy," *Journal of Clinical Hypertension*, vol. 17, no. 3, pp. 239–241, 2015.

[64] Canepa, M. *et al.*, "Distinguishing ventricular septal bulge versus hypertrophic cardiomyopathy in the elderly," *Heart*, vol. 102, no. 14, pp. 1087–1094, 2016.

[65] Katholi, R. E. and Couri, D. M., "Left Ventricular Hypertrophy: Major Risk Factor in Patients with Hypertension: Update and Practical Clinical Applications," *International Journal of Hypertension*, vol. 2011, Makaryus, A. N., Ed., p. 495 349, 2011.

[66] Edler, I. and Lindström, K., "The history of echocardiography," *Ultrasound in Medicine and Biology*, vol. 30, no. 12, pp. 1565–1644, 2004.

[67] Stolzmann, P. *et al.*, "Left Ventricular and Left Atrial Dimensions and Volumes," *Investigative Radiology*, vol. 43, no. 5, pp. 284–289, May 2008.

[68] Nagueh, S. F. *et al.*, "Recommendations for the Evaluation of Left Ventricular Diastolic Function by Echocardiography: An Update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging," *European Heart Journal – Cardiovascular Imaging*, vol. 17, no. 12, pp. 1321–1360, 2016.

[69] Lang, R. M. *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," 2015.

[70] Carovac, A., Smajlovic, F., and Junuzovic, D., "Application of ultrasound in medicine," *Acta Informatica Medica*, vol. 19, no. 3, p. 168, 2011.

[71] Ziaeian, B. and Fonarow, G. C., "Epidemiology and aetiology of heart failure," *Nature Reviews Cardiology*, vol. 13, no. 6, pp. 368–378, 2016.

[72] Mitchell, C. *et al.*, "Guidelines for Performing a Comprehensive Transthoracic Echocardiographic Examination in Adults: Recommendations from the American Society of Echocardiography," *Journal of the American Society of Echocardiography*, vol. 32, no. 1, pp. 1–64, 2019.

[73] Zhang, J. *et al.*, "Fully Automated Echocardiogram Interpretation in Clinical Practice," *Circulation*, vol. 138, no. 16, pp. 1623–1635, 2018.

[74] Zamzmi, G. *et al.*, "Harnessing Machine Intelligence in Automatic Echocardiogram Analysis: Current Status, Limitations, and Future Directions," *IEEE Reviews in Biomedical Engineering*, pp. 1–28, 2020.

[75] Asch, F. M. *et al.*, "Automated Echocardiographic Quantification of Left Ventricular Ejection Fraction Without Volume Measurements Using a Machine Learning Algorithm Mimicking a Human Expert," *Circulation: Cardiovascular Imaging*, vol. 12, no. 9, pp. 1–9, 2019.

[76] Abdi, A. H. *et al.*, "Quality assessment of echocardiographic cine using recurrent neural networks: Feasibility on five standard view planes," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10435 LNCS, no. September, pp. 302–310, 2017.

[77] Smistad, E. *et al.*, "Real-Time Automatic Ejection Fraction and Foreshortening Detection Using Deep Learning," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 12, pp. 2595–2604, 2020.

[78] Schneider, M. *et al.*, "A machine learning algorithm supports ultrasound-naïve novices in the acquisition of diagnostic echocardiography loops and provides accurate estimation of LVEF," *International Journal of Cardiovascular Imaging*, no. 0123456789, 2020.

[79] Voigt, J. U. *et al.*, "How to do LA strain," *European Heart Journal Cardiovascular Imaging*, vol. 21, no. 7, pp. 715–717, 2020.

[80] Rudski, L. G. *et al.*, "Guidelines for the Echocardiographic Assessment of the Right Heart in Adults: A Report from the American Society of Echocardiography. Endorsed by the European Association of Echocardiography, a registered branch of the European Society of Cardiology, and," *Journal of the American Society of Echocardiography*, vol. 23, no. 7, pp. 685–713, 2010.

[81] Madani, A. *et al.*, "Fast and accurate view classification of echocardiograms using deep learning," *NPJ Digital Medicine*, vol. 1, no. 1, pp. 1–8, Dec. 2018.

[82] Østvik, A. *et al.*, "Real-Time Standard View Classifcation in Transthoracic Echocardiography Using Convolutional Neural Networks," *Ultrasound in Medicine and Biology*, vol. 45, no. 2, pp. 374–384, 2018.

[83] Feigenbaum, H., "Role of M-mode Technique in Today's Echocardiography," *Journal of the American Society of Echocardiography*, vol. 23, no. 3, pp. 240–257, 2010.

[84] Jafari, M. H. *et al.*, "Automatic biplane left ventricular ejection fraction estimation with mobile point-of-care ultrasound using multi-task learning and adversarial training," *International journal of computer assisted radiology and surgery*, vol. 14, no. 6, pp. 1027–1037, 2019.

[85] Leclerc, S. *et al.*, "Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2198–2210, 2019. arXiv: 1908.06948.

[86] Østvik, A. *et al.*, "Automatic myocardial strain imaging in echocardiography using deep learning," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11045 LNCS, pp. 309–316, 2018.

[87] Knackstedt, C. *et al.*, "Fully Automated Versus Standard Tracking of Left Ventricular Ejection Fraction and Longitudinal Strain the FAST-EFs Multicenter Study," *Journal of the American College of Cardiology*, vol. 66, no. 13, pp. 1456–1466, 2015.

[88] Elwazir, M. Y. *et al.*, "Fully Automated Mitral Inflow Doppler Analysis Using Deep Learning," in *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, 2020, pp. 691–696.

[89] Zamzmi, G. *et al.*, "Echo doppler flow classification and goodness assessment with convolutional neural networks," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, IEEE, 2019, pp. 1744–1749.

[90] ——, "Fully automated spectral envelope and peak velocity detection from Doppler echocardiography images," in *Medical Imaging 2020: Computer-Aided Diagnosis*, vol. 11314, International Society for Optics and Photonics, 2020, 113144G.

[91] Queirós, S. *et al.*, "Fully Automatic 3-D-TEE Segmentation for the Planning of Transcatheter Aortic Valve Implantation," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1711–1720, 2017.

[92] Queirós, S. *et al.*, "Validation of a Novel Software Tool for Automatic Aortic Annular Sizing in Three-Dimensional Transesophageal Echocardiographic Images," *Journal of the American Society of Echocardiography*, vol. 31, no. 4, 515–525.e5, Apr. 2018.

[93] Moghaddasi, H. and Nourian, S., "Automatic assessment of mitral regurgitation severity based on extensive textural features on 2D echocardiography videos," *Computers in Biology and Medicine*, vol. 73, pp. 47–55, 2016.

[94] Andreassen, B. S. *et al.*, "Mitral Annulus Segmentation Using Deep Learning in 3-D Transesophageal Echocardiography," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 4, pp. 994–1003, 2019.

[95] Alsharqi, M. *et al.*, "Artificial intelligence and echocardiography," *Echo Research and Practice*, R115–R125, 2018.

[96] Slomka, P. J. *et al.*, "Cardiac imaging: working towards fully-automated machine analysis & interpretation," *Expert Rev Med Devices*, vol. 14, no. 3, pp. 197–212, 2017.

[97] Thorstensen, A. *et al.*, "Reproducibility in echocardiographic assessment of the left ventricular global and regional function, the HUNT study," *European Journal of Echocardiography*, vol. 11, no. 2, pp. 149–156, 2010.

[98] Diaz, T. *et al.*, "Prevalence, clinical correlates, and prognosis of discrete upper septal thickening on echocardiography: The framingham heart study," *Echocardiography*, vol. 26, no. 3, pp. 247–253, 2009.

[99] Narula, S. *et al.*, "Machine-Learning Algorithms to Automate Morphological and Functional Assessments in 2D Echocardiography," *Journal of the American College of Cardiology*, vol. 68, no. 21, pp. 2287–2295, 2016.

[100] Mahmoud, A., Bansal, M., and Sengupta, P. P., "New Cardiac Imaging Algorithms to Diagnose Constrictive Pericarditis Versus Restrictive Cardiomyopathy," *Current Cardiology Reports*,

[101] Qazi, M. *et al.*, "Automated Heart Wall Motion Abnormality Detection from Ultrasound Images Using Bayesian Networks.," in *IJCAI*, vol. 7, 2007, pp. 519–525.

[102] Leung, K. Y. E. and Bosch, J. G., "Localized shape variations for classifying wall motion in echocardiograms," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2007, pp. 52–59.

[103] Negahdar, M. *et al.*, "Automatic extraction of disease-specific features from Doppler images," in *Medical Imaging 2017: Computer-Aided Diagnosis*, vol. 10134, International Society for Optics and Photonics, 2017, 101340N.

[104] Araki, T. *et al.*, "A new method for IVUS-based coronary artery disease risk stratification: a link between coronary & carotid ultrasound plaque burdens," *Computer Methods and Programs in Biomedicine*, vol. 124, pp. 161–179, 2016.

[105] Kalinić, H. *et al.*, "Image registration and atlas-based segmentation of cardiac outflow velocity profiles," *Computer methods and programs in biomedicine*, vol. 106, no. 3, pp. 188–200, 2012.

[106] Sanchez-Martinez, S. *et al.*, "Characterization of myocardial motion patterns by unsupervised multiple kernel learning," *Medical Image Analysis*, vol. 35, pp. 70–82, 2017.

[107] Ahmad, T. *et al.*, "Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients," *Journal of the American Heart Association*, vol. 7, no. 8, e008081, 2018.

[108] Segar, M. W. *et al.*, "Phenomapping of patients with heart failure with preserved ejection fraction using machine learning-based unsupervised cluster analysis," *European Journal of Heart Failure*, vol. 22, no. 1, pp. 148–158, 2020.

[109] Sanchez-Martinez, S. *et al.*, "Machine learning analysis of left ventricular function to characterize heart failure with preserved ejection fraction," *Circulation: Cardiovascular Imaging*, vol. 11, no. 4, e007138, 2018.

[110] Cikes, M. *et al.*, "Machine learning-based phenogrouping in heart failure to identify responders to cardiac resynchronization therapy," *European Journal of Heart Failure*, vol. 21, no. 1, pp. 74–85, 2019.

[111] Cikes, M. *et al.*, "Machine-learning integration of complex echocardiographic patterns and clinical parameters from cohorts and trials," *European Heart Journal*, vol. 40, 2019.

[112] Loncaric, F. *et al.*, "Comprehensive data integration—Toward a more personalized assessment of diastolic function," *Echocardiography*, vol. 37, no. 11, pp. 1926–1935, 2020.

[113] D'Amour, A. *et al.*, "Underspecification presents challenges for credibility in modern machine learning," *arXiv preprint*, 2020. arXiv: 2011.03395.

[114] Geirhos, R. *et al.*, "Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," *7th International Conference on Learning Representations, ICLR 2019*, no. c, pp. 1–22, 2019. arXiv: 1811.12231.

[115] Hermann, K. L. and Kornblith, S., "Exploring the Origins and Prevalence of Texture Bias in Convolutional Neural Networks," 2019. arXiv: 1911.09071.

[116] Degel, M. A., Navab, N., and Albarqouni, S., "Domain and Geometry Agnostic CNNs for Left Atrium Segmentation in 3D Ultrasound," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11073 LNCS, 2018, pp. 630–637. arXiv: 1805.00357.

[117] D'hooge, J. and Fraser, A. G., "Learning about machine learning to create a self-driving Echocardiographic laboratory: Technical considerations," *Circulation*, vol. 138, no. 16, pp. 1636–1638, 2018.

[118] Heaven, W. D., *Google's medical AI was super accurate in a lab. Real life was a different story.* Apr. 2020.

[119] Shahgaldi, K. *et al.*, "Visually estimated ejection fraction by two dimensional and triplane echocardiography is closely correlated with quantitative ejection fraction by real-time three dimensional echocardiography," *Cardiovascular Ultrasound*, vol. 7, no. 1, pp. 1–7, 2009.

[120] Blondheim, D. S. *et al.*, "Reliability of Visual Assessment of Global and Segmental Left Ventricular Function: A Multicenter Study by the Israeli Echocardiography Research Group," *Journal of the American Society of Echocardiography*, vol. 23, no. 3, pp. 258–264, 2010.

[121] Diller, G. P. *et al.*, "Utility of machine learning algorithms in assessing patients with a systemic right ventricle," *European Heart Journal Cardiovascular Imaging*, vol. 20, no. 8, pp. 925–931, 2019.

[122] Madani, A. *et al.*, "Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease," *NPJ Digital Medicine*, vol. 1, no. 1, pp. 1–11, 2018.

[123] Ghorbani, A. *et al.*, "Deep learning interpretation of echocardiograms," *NPJ Digital Medicine*, vol. 3, no. 1, pp. 1–10, 2020.

[124] Siegersma, K. R. *et al.*, "Artificial intelligence in cardiovascular imaging: state of the art and implications for the imaging cardiologist," *Netherlands Heart Journal*, vol. 27, no. 9, pp. 403–413, 2019.

[125] Radford, A. *et al.*, "Improving language understanding by generative pre-training," 2018.

[126] Radford, A. *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[127] Brown, T. B. *et al.*, "Language models are few-shot learners," *arXiv preprint*, 2020. arXiv: 2005.14165.

[128] Sengupta, P. P. and Adjeroh, D. A., "Will artificial intelligence replace the human Echocardiographer?: Clinical Considerations," *Circulation*, vol. 138, no. 16, pp. 1639–1642, 2018.

[129] Seetharam, K. *et al.*, "Clinical Inference From Cardiovascular Imaging: Paradigm Shift Towards Machine-Based Intelligent Platform," *Current Treatment Options in Cardiovascular Medicine*, vol. 22, no. 3, p. 8, 2020.

[130] Signify Research, "What's new for machine learning in medical imaging," Tech. Rep., 2019.

[131] Martin-Isla, C. *et al.*, "Image-Based Cardiac Diagnosis With Machine Learning: A Review," *Frontiers in Cardiovascular Medicine*, vol. 7, Jan. 2020.

[132] Tabassian, M. *et al.*, "Diagnosis of heart failure with preserved ejection fraction: machine learning of spatiotemporal variations in left ventricular deformation," *Journal of the American Society of Echocardiography*, vol. 31, no. 12, pp. 1272–1284, 2018.

[133] Ouyang, D. *et al.*, "Interpretable AI for beat-to-beat cardiac function assessment," *Nature*, 2020.

[134] Pedrosa, J. *et al.*, "Fully automatic assessment of mitral valve morphology from 3D transthoracic echocardiography," in *2018 IEEE International Ultrasonics Symposium (IUS)*, IEEE, 2018, pp. 1–6.

[135] Schwab, K., "The Fourth Industrial Revolution: what it means and how to respond," Tech. Rep., 2020.

[136] Zhu, Y. *et al.*, "Semantic amodal segmentation," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 3001–3009, 2017. arXiv: 1509.01329.

[137] Leclerc, S. *et al.*, "Deep Learning Segmentation in 2D echocardiography using the CAMUS dataset : Automatic Assessment of the Anatomical Shape Validity," in *International Conference on Medical Imaging with Deep Learning – Extended Abstract Track*, 2019.

Some images in this thesis were built with icons from The Noun Project including creators: Garrett Knoll, Symbolon, SBTS, Vectors Point, ProSymbols, David Christensen, ArmOkay, Nick Bluth, LAFS, Erickson Duverge, and Tatyana.

Some ultrasound images shown in this thesis were visualized using EchoPAC (GE Vingmed, Horten, NO).

Papers

Paper I

User-Intended Doppler Measurement Type Prediction Combining CNNs with Smart Post-Processing

Andrew Gilbert, Marit Holden, Line Eikvil, Mariia Rakhmail, Aleksandar Babić, Svein Arne Aase, Eigil Samset, Kristin McLeod

Published in *Journal of Biomedical and Healthcare Informatics*, 2020, DOI: 10.1109/JBHI.2020.3029392.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3029392, IEEE Journal of Biomedical and Health Informatics

User-Intended Doppler Measurement Type Prediction Combining CNNs With Smart Post-Processing

Andrew Gilbert, Marit Holden, Line Eikvil, Mariia Rakhmail, Aleksandar Babić, Svein Arne Aase, Eigil Samset, Kristin McLeod

***Abstract*— Spectral Doppler measurements are an important part of the standard echocardiographic examination. These measurements give insight into myocardial motion and blood flow, providing clinicians with parameters for diagnostic decision making. Many of these measurements are performed automatically with high accuracy, increasing the efficiency of the diagnostic pipeline. However, full automation is not yet available because the user must manually select which measurement should be performed on each image. In this work, we develop a pipeline based on convolutional neural networks (CNNs) to automatically classify the measurement type from cardiac Doppler scans. We show how the multi-modal information in each spectral Doppler recording can be combined using a meta parameter post-processing mapping scheme and heatmaps to encode coordinate locations. Additionally, we experiment with several architectures to examine the tradeoff between accuracy, speed, and memory usage for resource-constrained environments. Finally, we propose a confidence metric using the values in the last fully connected layer of the network and show that our confidence metric can prevent many misclassifications. Our algorithm enables a fully automatic pipeline from acquisition to Doppler spectrum measurements. We achieve 96% accuracy on a test set drawn from separate clinical sites, indicating that the proposed method is suitable for clinical adoption.**

***Index Terms*— Convolutional neural network (CNN), deep learning, classification, ultrasound (US), Doppler**

I. Introduction

ECHOCARDIOGRAPHY is the primary method used to image the heart due to its portability, affordability, and absence of ionizing radiation. The diagnostic power of echocardiography is reflected in clinical guidelines. Echocardiography indices are included as both minor and major clinical diagnostic criteria in many protocols [1]. As computational power increases image quality improves. Consequently, the theoretical accuracy of clinical measurements also increases.

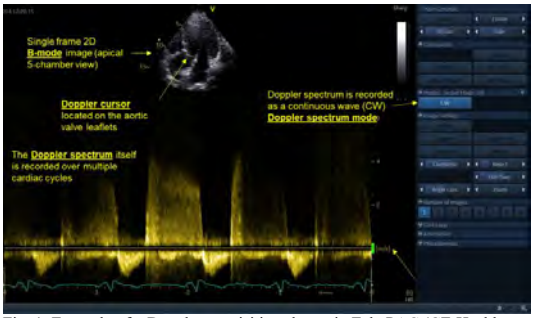


Fig. 1. Example of a Doppler acquisition shown in EchoPAC (GE Healthcare, Horten, NO) depicting the relevant information to a spectrum classification problem as a clinician would see it.

In addition to the diagnostic power, there is a growing trend to use echocardiography as a therapy guidance tool to support interventions and complement other imaging modalities. Minimally invasive valve interventions are much less risky than full surgery and are becoming the therapy of choice as techniques and prosthetics advance. Spectral Doppler imaging is the primary method to assess blood flow across valves, a crucial step for intervention planning and follow-up [2]. Therefore, spectral Doppler imaging has become an integral component of the echocardiography exam to provide a means to assess hemodynamic function in all four valves of the heart.

A. Spectral Doppler Measurements

Fig. 1 shows an example of a spectral Doppler acquisition as seen in EchoPAC (GE Healthcare, Horten, NO). There are many important features of the acquisition that are available within the raw data of each recording:

- The **Doppler spectrum** is displayed over multiple cardiac cycles for analysis and measurement.
- The **relative baseline** of the Doppler spectrum can be

This project has received funding from the European Union's Horizon 2020 research and Innovation program under the Marie Sklodowska-Curie grant agreement No 764738.

A. Gilbert and E. Samset are with GE Healthcare and also with the Department of Informatics at the University of Oslo, both Oslo, NO (email: andrew.gilbert@ge.com; eigil.samset@ge.com).

M. Holden and L. Eikvil are with the Norwegian Computing Center, Oslo, NO (email: marit.holden@nr.no; line.eikvil@nr.no).

A. Babić and M. Rakhmaiil were with GE Healthcare in Oslo, NO for their contribution to this work (email: mariya.rakhmayil@ge.com and aleksandar.babic@gmail.com).

S. A. Aase, and K. McLeod are with GE Healthcare in Oslo, NO (email: sveinarne.aase@ge.com; and kristin.mcleod@ge.com).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3029392, IEEE Journal of Biomedical and Health Informatics

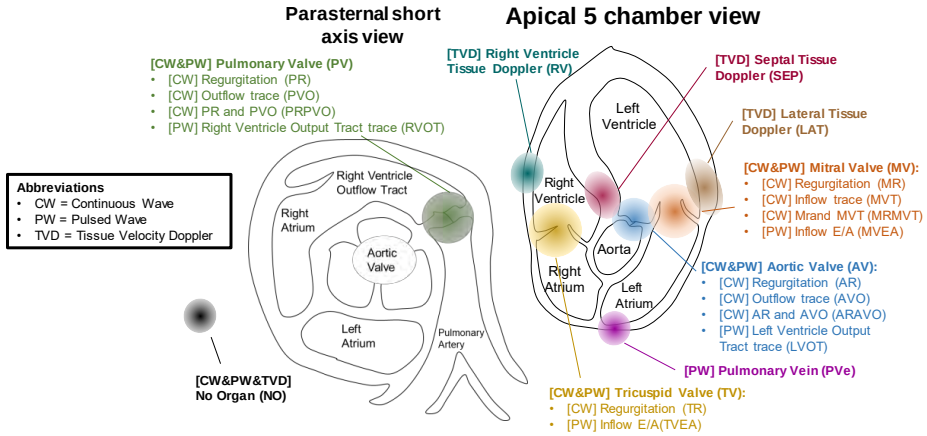


Fig. 2. Each of the Doppler measurement types sorted by the location of the cursor position. Each color corresponds to a different region in anatomical space. The mode of each measurement in shown in front. Apical 5 chamber and parasternal short axis views are shown here for illustrative purposes only, to demonstrate the relative positions of the classes. Doppler spectra are typically acquired from a variety of echocardiographic views (see Appendix A for details) and part of the challenge of this problem is that the spatial relationship between structures demonstrated above will change depending on the view used for image acquisition. The No Organ (NO) class refers to images of air and ultrasound gel.

adjusted by the user during acquisition to focus on a specific part of the spectrum and prevent aliasing.

- The **mode** provides information on how the Doppler spectrum was acquired. Spectral Doppler incorporates three main imaging modes: Continuous Wave (CW) Doppler, Pulsed Wave (PW) Doppler, and Tissue Velocity Doppler (TVD). CW is used to measure high velocity blood flow across valves, PW provides flow analysis at specific spatial points, and TVD provides quantifiable myocardial velocities.
- The 2D **B-mode** (brightness mode) image shows the orientation of the probe with respect to the physical anatomy of the heart. Doppler spectra can be obtained from a variety of probe positions and angles depending on the desired measurement. The scan converted B-mode image is displayed here to orient the user.
- The **Doppler cursor,** visible on top of the B-mode image, indicates the spatial location of the spectrum. This parameter is interpreted in the context of the B-mode image. See Fig. 2 for a visual overview of how the cursor location corresponds to specific points in anatomical space. In the TVD classes the cursor is focused directly on the tissue, while in the CW and PW classes the cursor is focused on an area of blood flow. Exact positioning will depend on the desired measurement, operator preference, and individual patient anatomy.

Together, this information identifies the Doppler spectrum and therefore which measurements should be performed.

B. Clinical Need for Measurement Type Classification

Accurate automatic classification of Doppler measurement types can be combined with already available automated measurement techniques (e.g. [3], [4]) to provide fully automated analysis of Doppler spectra. Specifically, in a fully automated workflow, as soon as a Doppler exam is acquired the classification system is triggered and determines the measurement type. The system then triggers the corresponding automated measurement algorithm to display the measurement with no additional user interaction. This workflow is more efficient, allowing clinicians to spend more time on difficult measurements.

Furthermore, many clinics have petabytes of patient data in their archive systems from tracking patients over time. Thus, if used in combination with automated measurement techniques, one application of automatic Doppler measurement type classification is to perform rapid historical analysis on past exams in a robust and standardized manner. All information used in the proposed classification system is readily available in hospital archives if those archives store the raw data for each patient. Knowing a patient's progression from previous checkpoints can provide further information to support therapy planning. Therefore, historical analysis would provide clinical value through objective study of measurements over time. Another application is continually performing analysis on patients, which could bring statistical power to the development and augmentation of clinical guidelines.

C. Related Work

1) Ultrasound Classification

Doppler measurement type classification is unique because of the heterogeneity of data available in each classification example. As shown in Fig. 1, each recording contains image data, spectral data, modal parameters, a baseline position, and Doppler cursor coordinate locations. Previously, many of these items have been automatically classified individually, borrowing techniques from non-medical domains. Processing of spectral data has been a common task for several decades in speech recognition [5], and these techniques have been applied to Doppler spectra as well. For example, Wright *et al.* used artificial neural networks to classify Doppler spectra from arteries [6]. Meanwhile, automatic image classification has also

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3029392, IEEE Journal of Biomedical and Health Informatics

become increasingly common as CNNs have surpassed the accuracy of humans on many tasks. Recently, these techniques were applied to echocardiographic B-mode images to automatically classify cardiac views with high accuracy [7], [8].

2) Multi-modal Learning

In non-medical fields, several groups have also looked at how data from different modalities can be combined. Ngiam *et al.* showed how a deep autoencoder could be trained with both video and speech data to generate a shared representation [9]. Ephrat *et al.* demonstrated how video and speech data could be encoded separately and then combined in a bidirectional long short-term memory network to solve the cocktail party problem of singling out a single speaker in a noisy audio track [10].

While many deep learning techniques have successfully made the transition from non-medical to medical applications [11], applying multimodal learning techniques remains a challenge because there are several orders of magnitude difference in the amount of available data. For example, Ephrat *et al.* were able to use >2000 hours of automatically annotated data. The annotation of such a volume of data in the context of Doppler spectra is challenging due to the lack of available simulated data. Transfer learning and fine tuning have previously been applied to solve data magnitude problems in medical imaging [12]. However, it is of limited use here since task objectives are different, and the relationship between the modalities (Doppler spectrum to B-mode) varies for each Doppler measurement class.

3) Confidence

One challenge in ultrasound imaging is that images acquired in clinical settings are not necessarily in standard views. During training, models are exposed to only a subset of possible views that might be seen in a clinical workflow. This is a concern in the given classification problem where misclassifications are more costly than doing nothing. Therefore, an algorithm to classify such images needs a mechanism to handle non-standard cases. This can be either collecting large datasets that can cover all possible views (even those that are non-standard) or a mechanism to bail-out when the image doesn't fall in the label set, such as via confidence metrics with a set threshold for acceptance.

Several groups have looked at how networks can give a confidence prediction along with an output label. It is well known that CNNs are prone to overfitting and cannot generalize well from the training set to unseen inputs [13]. Previously, Bayesian models have been used to provide a better estimate of model uncertainty by encoding model weights as a probability distribution. However, Bayesian techniques often come with increased parameter count and a higher computational cost to adequately model random distributions [14]. Monte Carlo dropout (MC-dropout) uses dropout at test time to approximate Bayesian inference with a lower computation cost [15]. Other methods such as temperature scaling [16] or histogram binning [17] calibrate fully trained network outputs without changing inference. Parameters are learned on the validation set to map network outputs to a true confidence distribution. These methods have the advantage of maintaining inference time and increasing the interpretability of the results without sacrificing the accuracy of the model.

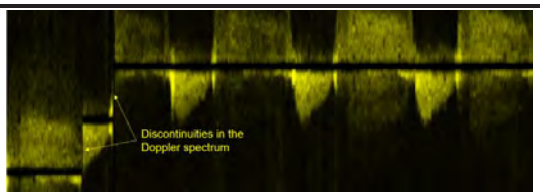


Fig. 3. Discontinuities arise when the operator shifts the baseline during acquisition. This is common practice when acquiring several measurements.

D. Contributions

After an analysis of the data, the spectral information was eliminated from the processing pipeline. This reduced the input to a B-mode image, Doppler cursor coordinate location, baseline position, and mode parameter. Although spectra provide useful information (and are used by clinical experts when labelling images), there are many variations in the collection of spectra that make it difficult to use in a network. For example, as shown in Fig. 3, spectral data can have discontinuities in the baseline as the user changes the parameters during acquisition. Spectral data is also variable length, which effectively shrinks or expands the features in the output image. Dealing with variable length would require an even larger dataset, since CNNs are not magnitude invariant. To avoid adding unnecessary complexity, the method developed here does not rely on spectral data. Instead, the method is focused on the integration of the latter four parameters. The spectra can be eliminated because we develop a novel pipeline which breaks the problem into a series of simpler pieces. We create an alternate way of uniquely identifying the spectra using these pieces. Our pipeline is outlined in Fig. 4 with references to the relevant section numbers for each piece. In brief, the principle contributions of this work are four-fold:

(1) **Heatmap encoding:** We show how to encode spatial features at the input of CNNs when multi-modal data includes coordinate locations as features.

(2) **Multi-head output:** We borrow techniques from multi-task learning to develop a multi-head learning strategy that integrates mode information to prevent misclassifications and reduce network size.

(3) **Decision tree mapping:** We use decision trees to incorporate user-defined imaging parameters in order to simplify the task of the CNN and better predict user intentions for the desired measurement type.

(4) **Confidence Thresholds:** We demonstrate how neural network layers, besides just the final layer, can be used to define a confidence metric that will disregard many images that differ from the training set. Our method requires no extra trained parameters, uses a fully nonlinear mapping between the output values and the network confidence estimate, and can be dynamically modified at inference time depending on the desired tradeoff between ignored and error rates.

To the best of our knowledge, this is the first work to use CNNs to classify Doppler measurement types. We achieve high

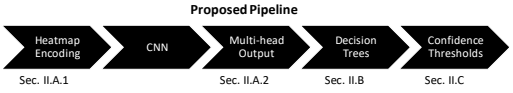


Fig. 4. The pipeline of the proposed method with relevant section numbers.

accuracy on the task, while maintaining a small memory footprint and close to real-time performance. Moreover, several of the methods developed in this work may be applicable to other classification problems, especially in medical imaging.

II. Methods

Through conversations with clinical experts, 18 of the most common measurement types for adults were identified. Three additional types of Doppler measurements were identified but were excluded from the current algorithm design due to infrequent clinical use. Specifically, among a dataset of over 7000 random Doppler images collected from a clinical site, just 30 images came from these three classes combined. Including these measurement types would be more likely to confuse the network and would require a significant effort to collect sufficient training data. Additionally, including the classes would not result in noticeable clinical time savings after algorithm implementation since they are used infrequently. Two steps are taken to account for measurement types not covered in our label set. First, to avoid making a classification on images scanned without a visible B-mode image on screen, a no organ (NO) class was added which consisted only of images where air and varying amounts of ultrasound gel were scanned. The Doppler cursor, baseline, and other parameters were chosen to cover a variety of possible inputs for the NO images. Second, a confidence metric was designed (Sec. II.C) to discard images from other classes. A full discussion of each of the measurement types is outside the scope of this work, but Fig. 2 shows a diagram of the relative cursor positions as well as the abbreviations for each type. An outline of each measurement's use and acquisition is available in [1], and reports specific to CW and PW [18], and TVD [19] mode measurements are also available. In addition, a further description of each measurement type is presented in Appendix A of the supplementary material.

The proposed method performs a classification on these Doppler measurement types. The relevant anatomical region is determined using a CNN as described in Sec. II.A. Sec. II.A.1 explains heatmap encoding at the input of the network while Sec. II.A.2 describes a multi-head output approach to divide the classification according to the imaging mode. A decision tree to simplify the network's classification task is presented in Sec. II.B. A confidence metric is defined in Sec. II.C to avoid misclassifications for low-confidence cases such as images from other measurement types or images with poor quality. Finally, the design of the dataset used for training and testing is outlined in Sec. II.D.

A. Determining Cursor Location with CNNs

As shown in Fig. 1, a single Doppler recording is composed of many multi-modal features. Given the information in the format of Fig. 1, an expert observer can mentally integrate the relevant information and classify the type of measurement that should be made. However, it would be unrealistic to expect a network to perform a classification given only an image such as Fig. 1 because some of the most important pieces of information are not emphasized in the image. For example, the Doppler cursor is very important to the classification because it indicates the location of the Doppler spectrum within the heart, but it is only a small marker on the image.

Instead, all the relevant data is extracted individually from each recording. The mode is recorded as either CW, PW, or TVD. The relative baseline is extracted as a float in the range from 0 to 1, where the default (unchanged) location is 0.5. The raw B-mode data is extracted as a 512×256 image, since the depth dimension is usually much larger in the raw data. Note that the non-scan converted (beam space) data is used directly rather than the scan-converted (probe space) data that is shown to the user. The added step in the pipeline to scan-convert the images yields no gain in this application where the Doppler cursor position relative to the heart structures is the key piece of information. Scan-converted images could equivalently be used.

As shown by the different colors in Fig. 2, the measurement types can be grouped into 9 locations in anatomical space. Since the relationship between cursor coordinates and image features would be similar for each of these locations, all measurements from the same anatomical location are merged into the same class for the CNN. Thus, the task of the CNN is only to figure out which anatomical location the measurement came from, the rest is handled during post-processing as described in Sec. II.B. The only inputs into the network are the B-mode image and the cursor coordinate.

1) Heatmap Encoding

The position of the cursor is extracted relative to the original B-mode image as a coordinate pair. In the proposed approach the coordinates are encoded as a heatmap. The coordinates are not directly used because Liu *et al*. showed CNN's are typically poor at learning mapping between coordinates in cartesian coordinate space and pixel space [20]. Additionally, in landmark detection problems, the current state of the art is to extract landmark coordinates from heatmaps of likely locations produced by the network [21]. Intuitively, using heatmaps works because there is a linear mapping from the coordinate space of the input image to the output heatmap. Logically, networks should also perform better if landmarks at the input are encoded as heatmaps instead of input as coordinates. To encode the Doppler cursor location as a heatmap, we generate a 2D normal gaussian probability density function with a standard deviation of 10 pixels centered at the cursor coordinate. The heatmap is generated in 512×256 resolution to match the original raw data, and then appended to the input image as an additional channel. Image and heatmap are both rescaled to 256×256, which has the effect of compressing the gaussian vertically. This allows the expected spatial distribution of the landmark to more closely match the physical dimensions of the raw data. An example heatmap is shown in Fig. 5. Finally, both image and heatmap are cropped to 224×224. During training, random crops are used for augmentation. Center crops are used during validation and testing.

2) Multi-Head Network

As shown in Fig. 2, other than Pulmonary Vein, the CW and PW mode measurements share the same anatomical locations.

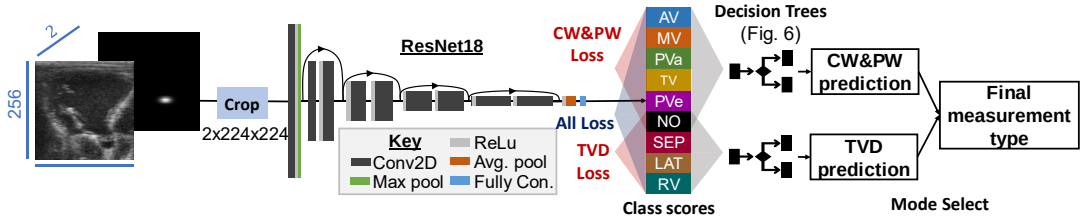


Fig. 5. The heatmap of the Doppler cursor location is appended as a channel to the non-scanconverted B-mode image and both are cropped to 224×224 before being input to the network. A ResNet18 [28] is presented here, but a variety of network architectures are tested (section III.A.3). For all networks, the last fully connected layer (of size 9) is split into two groups (heads). The No Organ (NO) class is input to both heads so the CW&PW head is 6 classes and the TVD node is 4 classes. During training the loss can be backpropagated from each head individually, or together from all classes (section II.A.2) During inference each head is passed to its own classifier and decision tree (see Fig. 6 for decision trees). The mode parameter is used to select between the two outputs to yield a final class. See Fig. 2 for class abbreviations.

The CW and PW locations are also completely distinct from the TVD mode measurements (except for the no-organ synthetic class). Because the mode parameter is always set by the user, simply training a network to classify all outputs would lead to unnecessary errors. The network should never classify a TVD mode image into a CW/PW measurement type since the mode is known at classification time, but without explicit separation this misclassification may occur. To solve this, the set of anatomical location classes can be split into unique sets, one for CW&PW and one for TVD. One approach would then be to train a different network for each mode and call the network for the relevant mode during inference. However, this approach doubles the memory footprint of any implementation, which is a downside for integration into a resource-constrained environment.

An alternative solution is to frame this as a multi-task learning problem. Multi-task learning integrates the information from several related tasks into a single network by implementing a separate network branch for each task. Often in multi-task networks, information from one task improves performance on another. The approach has proven to be successful in a variety of deep learning applications [22]. Our method is a slight variant of multi-task learning adapted for this problem. In detail, the network presented here has only a single branch with multiple classification heads operating on the final layer. Only one head is relevant for every given input sample. Therefore loss is backpropagated only from the classification head with a mode matching that sample whereas in multi-task learning loss can be taken from all branches during training. Input to each classification head are the values from the last fully connected layer for the classes that belong to that mode. The total loss for a given minibatch is shown in (1) where $f(x;\theta)$ is the output of a network with input x and parameters θ, x_{TVD} and x_{CWPW} are the TVD and CWPW samples in the minibatch respectively, and CE is cross-entropy loss. The λ values control the weighting between heads.

$$L = \lambda_1 CE(f(x_{TVD};\theta)) + \lambda_2 CE(f(x_{CWPW};\theta)) \quad (1)$$

During inference, the CNN yields predictions from both heads, but only the value from the relevant mode is read by the calling function. Due to these differences we instead call this a "multi-head" network. With this design choice, we exploit the information about the different modes by including separate heads and loss functions for the CW&PW and TVD groupings of anatomical locations. The architecture of the multi-head network is shown in Fig. 5.

B. Decision Tree Mapping

After finding the anatomical location with the CNN, the next step in the pipeline is determining the final measurement type. The mode and relative baseline position parameters extracted from the spectrum linearly separate the measurement types in each anatomical region because users change those parameters based on which type of measurement they wish to acquire. Therefore, decision trees are used for post-processing the output of each head as shown in Fig. 6. One possible error is introduced in this scheme when a CW image is classified as a Pulmonary Vein (PVe). In preliminary experiments this was never an issue, but occurrence in a clinical setting would require manual re-classification.

Decision trees are a better solution than feeding the parameters into the network because it avoids unnecessary mistakes and enables the CNN to use classes that are based solely on regions in anatomical space. Otherwise, the CNN would likely be confused between classes from the same anatomical region. Additionally, several of the original smaller classes do not have enough images for a network to properly converge. Grouping the classes increases performance.

C. Confidence Metric

Correct classifications from the network will yield significant time savings for clinical users by automatically launching the tool associated with that Doppler measurement type, where available. However, incorrect classification comes with a cost as the user will have to navigate back in the menu and select the correct measurement type. As automation continues to permeate clinical workflows, this cost may become larger. Initial misclassification could trigger unrelated measurements and automated tools. Moreover, there may be images in a clinical setting that are different from those seen during training. Thus, it is important for the network to have a bail-out mechanism on images with high uncertainty.

Our approach relies on the last fully connected layer before the softmax classifier, named the "pre-softmax" layer. This layer was chosen because raw network estimates for all classes are readily available before distortion by the multiple heads. The pre-softmax values for each example in the validation set are recorded after the network weights are trained and frozen. The recorded values are divided into quantiles. That is, rather than learning a mapping from outputs to true confidence (as was done in [16] and [17]), a series of cutoff values are found for each confidence level. During test time, the quantile is set based

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3029392, IEEE Journal of Biomedical and Health Informatics

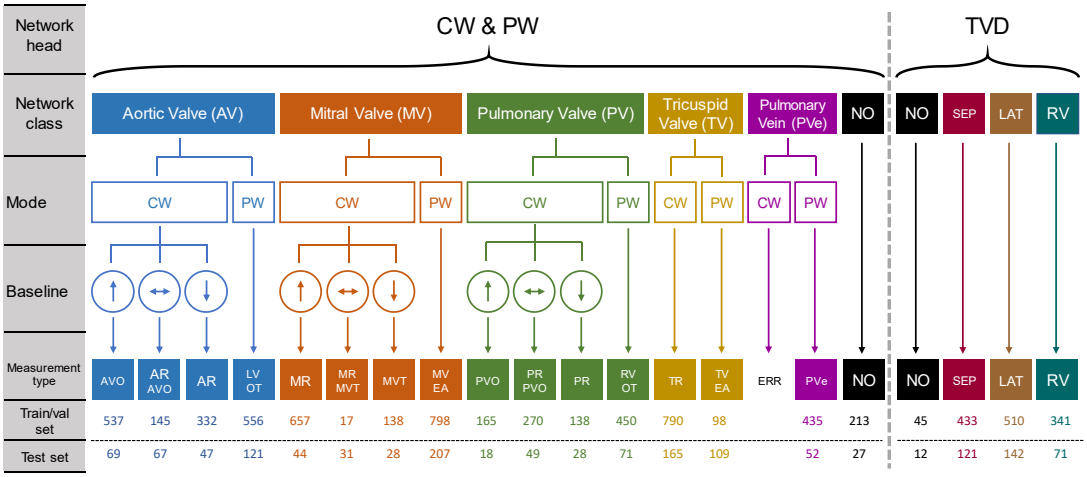


Fig. 6. Decision trees from the classes output by the network to the final classes determining the measurement type. The network has two heads, the CW&PW head and the TVD head. Outputs from each head are mapped to final classes using the mode and baseline parameters. The mode is then used to decide between the CW&PW head output and the TVD head output (Fig. 5). For the baseline: (↑) indicates a baseline was moved upwards – towards positive values, exposing a larger range of negative velocities, (↔) indicates a baseline is in center position, and (↓) indicates the baseline was moved downwards – towards negative values, exposing a larger range of positive velocities. If nothing is indicated for mode or baseline, then those parameters are not used for that mapping (all values map to the same class). For example. every image in the TVD head is guaranteed to be mode TVD so the mode is not relevant. Training, validation, and test set sizes are shown below each class. See Fig. 2 for locations of each class and acronym definition. ARAVO, MRMVT, and PRPVO are combinations of AR/AVO, MR/MVT, and PR/PVO respectively.

on the desired tradeoff between error rate and ignored rate. The maximum output value is found as usual, but if the pre-softmax value for that class falls below the given threshold then the image is labeled as low confidence and ignored.

To validate the chosen approach, results were compared to several other methods of determining confidence. MC-dropout has proven to be one method for approximating Bayesian inference in a computationally efficient manner [15]. An MC-dropout version of the model is implemented following the approach in [23], where 50% of the neurons from the last fully connected layer are randomly dropped during each inference run. In addition, combining the predictions of ensembles of neural networks has given superior classification performance [24], [25]. However, ensembled outputs can also be used as a measure of algorithm confidence. An ensembled confidence is implemented by ignoring cases where networks within the ensemble predict differing classes.

D. Dataset

The training and validation dataset consisted of exams previously collected by GE Healthcare for internal tool development. All exams were fully anonymized and came from a single clinical site. Exams were collected to try to maintain a high number in each class, but more images were available for classes that occur more frequently in clinical practice than those that occur infrequently. Thus, the dataset is slightly unbalanced because it reflects the clinical distribution of the data. Note that all classes of the same color in Fig. 2 are grouped together for training the network and split later in post-processing as shown in Fig. 6. The final set was 7081 images where individual class sizes are shown in Fig. 6.

Exams from seven institutions were used for the test set. The test institutions were spread over six different countries and three different continents. All test set institutions were different from the training set institution. This was done for two reasons. First, since images are fully anonymized, it is impossible to guarantee that two images from the same institution are not from the same patient. It is crucial for accurate test statistics that the training and test sets contain unique patients. Second, every institution has slightly different acquisition practices and patient populations, leading to small differences in the distribution of the images. Thus, to get a result that reflects real performance "in-the-wild" it is important to test on data from separate institutions. The test set contained 1479 images and class distributions are also shown in Fig. 6. All images were labeled by a clinical expert experienced in Doppler spectrum analysis and reviewed for accuracy by two other experts. Roles were swapped between sets, so a different expert did the initial labeling for the test set.

While gathering the training and validation sets, there were 298 images that had insufficient image quality for an expert to classify them. These images were categorized as the *unknown* set to analyze the confidence metric. Additionally, 30 images were identified that belonged to the three measurement classes not included in this network because they appear infrequently in clinical practice. These images were put into the *extra* set to analyze the confidence metric. Anonymization procedures removed all patient information, so the number of patients in the datasets is unknown.

E. Testing

To validate reproducibility, the combined training and validation set is used to estimate five different models. Each model is trained using 90% of the dataset with the remaining 10% set aside as validation. The model with the best

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3029392, IEEE Journal of Biomedical and Health Informatics

#	Architecture	Input	Networks	Classification Heads	Accuracy			F_1 Score (std)	Size (MB)	Time (ms)
					TVD	CWPW	Total (std)			
E1	ResNet18	I	2	One head per network	52.5%	70.8%	67.3% (0.4%)	63.1% (0.7%)	1480	3.5
E2	ResNet18	H	2	One head per network	83.9%	66.2%	68.8% (0.3%)	70.3% (1.3%)	1480	3.5
E3	ResNet18	I + H	2	One head per network	96.9%	95.3%	95.7% (0.3%)	96.3% (0.2%)	1480	3.5
E4	ResNet18	I + H	1	One head	90.8%	94.5%	93.7% (0.4%)	94.3% (0.3%)	**740**	3.5
E5	ResNet18	I + H	1	Training: one head Testing: two heads	98.4%	**95.7%**	**96.4%** (0.3%)	**97.1% (0.3%)**	**740**	3.5
E6	ResNet18	I + H	1	Two heads	**98.8%**	95.0%	95.8% (0.9%)	96.6% (0.7%)	**740**	3.5

Table 1. Comparison of experimental results for different input and output settings. In the Input column, I indicates only an image, H indicates only a heatmap, and I + H means the image with heatmap concatenated as shown in Fig. 5. In the Classification heads column "one head" refers to a single softmax classifier with all classes, and "two heads" refers to the multi-head approach detailed in Fig. 5. Accuracy is total correct images over total images (weighting classes with more images more) and F_1 score is an average across the individual F_1 scores of each class (weighting each class equally).

performance on each validation set during training is saved for testing on the independent test set. The train/validation/test split as a percentage of the total data is 74%/9%/17%. The five different validation sets are non-overlapping. Quantile cutoff limits for the confidence metric are extracted by averaging results across the five validation sets. This setup is used for all presented approaches and metrics are averaged from evaluating all five trained models on the unseen test set. Using five different models is important to (a) better estimate accuracies on the test set, (b) provide more robust quantile cutoff limits extracted across a larger set of unseen examples, and (c) obtain different models for the ensemble-based confidence method.

III. Results

To evaluate the effects of each design decision, a series of experiments were constructed with metrics measuring accuracy, speed, and memory usage. Classification accuracy was measured as defined in (2) where N_I is the total number of images in the test set, C is the set of classes, and TP_i is the number of true positives for class i. The F_1 score [26], was measured as a combination of recall and precision (3). The recall of class i, R_i, is given by $R_i = TP_i/(TP_i + FN_i)$ where FN_i is the number of false negatives. The precision of class i, P_i, is given by $P_i = TP_i/(TP_i + FP_i)$ where FP_i is the number of false positives. N_c is the number of classes.

$$Accuracy = {}^{1}/_{N_I} \sum_{i \in C} TP_i \tag{2}$$

$$F_1 = \frac{2 * \sum_{i \in C} R_i * \sum_{i \in C} P_i}{N_C * (\sum_{i \in C} R_i + \sum_{i \in C} P_i)} \tag{3}$$

Note that accuracy was obtained by micro-averaging (weighting by class frequency), while F_1 was obtained by macro-averaging (weighting each class equally). Although micro-averaging will bias results towards classes with more images, it also reflects the clinical reality since classes with more images will appear more in clinical use. The memory size and inference time measurements were implemented following [27]. We first examine the effect of different input and output settings in the experimental setup (Table 1) and then look at the performance of various network architectures (Table 2). For the first experiments a ResNet18 network (architecture shown in Fig. 5) was chosen because the residual connections in ResNet speed up training and improve accuracy when training deeper networks [28]. Specifically, ResNet18 has a smaller footprint than other networks and less parameters, which helps avoid overfitting on data-limited tasks.

A. Cursor Location with CNNs

1) Heatmap Encoding

First, the effect of adding the cursor heatmap was evaluated. A network was trained using only the B-mode image as an input (E1 in Table 1), using only the heatmap as an input (E2), and then with the cursor heatmap appended to the B-mode image (E3). In all three cases, separate networks were trained for each mode (CW&PW vs. TVD). As expected, with only a single input channel (either image or heatmap) there were low classification accuracies. The TVD network in E2 (heatmap only) did achieve 84% accuracy which shows the heatmap provides quite a bit of information for TVD cases. This is intuitive since TVD images are almost always acquired from the same echocardiography view (apical four chamber). Therefore, the position of the cursor remains relatively constant within each class and different between them. Conversely, results were worse with only a heatmap for the CW/PW mode. In this mode views (and therefore cursor locations) change within classes. Results showed a significant improvement with both the image and heatmap passed to the network (95.7% accuracy).

2) Multi-head Networks

Second, the effect of the multi-head approach was tested. As a baseline approach one network was trained with a single classification head on all 9 classes (E4). There are 9 classes instead of 10 here because with a single classification head only one NO class is needed. Results showed a 2% drop in accuracy compared to E3, indicating that not splitting the classes creates a harder task for the network. However, the memory footprint was also cut in half. To test whether the multi-head approach could achieve the same accuracy as two separate networks (E3) with the footprint of a single network, the multi-head architecture was applied at test time to the network trained in E4 (E5 in Table 1). Next, a single network was trained using the multi-head approach during both training and testing (E6). Experiments E3 – E6 all used the same input information, but with different methods of integrating the mode information.

Results showed that using a single head at training, but multiple heads at test time (E5) resulted in the best performance

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3029392, IEEE Journal of Biomedical and Health Informatics

#	Architecture	Accuracy	F_1 score	Size (MB)	Time (ms)
E5	ResNet18 [28]	96.4%	97.1%	740	**3.5**
E7	ResNet34 [28]	96.5%	97.1%	911	6.2
E8	ResNet50 [28]	96.2%	97.0%	966	8.8
E9	SqueezeNet-v1.1 [31]	94.2%	94.8%	584	3.9
E10	ShuffleNet [30]	96.0%	96.8%	**579**	12.0
E11	MobileNet-v2 [32]	96.0%	96.8%	582	7.4
E12	NASNet-A-Mobile [29]	**96.7%**	**97.3%**	653	46.3
E13	GoogLeNet [25]	96.1%	96.8%	702	10.0
E14	DenseNet-121 [33]	96.5%	**97.3%**	653	22
E15	DenseNet-169 [33]	96.2%	97.0%	735	30.9
E16	DenseNet-201 [33]	96.5%	**97.3%**	828	36.8
E17	BN-Inception [34]	96.2%	96.8%	697	13.9
E18	DualPathNet-68 [35]	96.4%	97.2%	735	27.7
E19	Xception [36]	96.4%	97.2%	891	8.0
E20	Inception-v3 [37]	96.0%	96.7%	908	15.8

Table 2. Comparison of experimental results with different network architectures. Accuracy and F1 scores are means across five networks trained using the setup described in section II.E. All experiments used the same input/output configuration as E5 in Table 1.

(96.3% accuracy), slightly better than using separate networks (E3). This result indicates that the network can use the information from other modes to improve performance like multi-task learning. It was also better to use the multi-head architecture only during testing (E5) than both training and testing (E6), possibly because the harder task of training with all 9 classes forced the networks to better differentiate between classes.

3) Network Architecture

Third, the impact of network architecture was tested. The top 15 architectures in terms of Top-1 accuracy density from the benchmark analysis of Bianco et al. [27] were evaluated. Top-1 accuracy density is classification accuracy (in their work on ImageNet) divided by number of parameters and is a measure of a network's performance relative to its size. Details on the implementation of all networks are provided in Appendix B. The E5 configuration was used for all architecture experiments.

Results are presented in Table 2. The NasNet-A-Mobile [29] (E12) had the best accuracy (96.7%), while ShuffleNet [30] (E10) had the smallest size (579 MB), and ResNet18 (E5) had the fastest inference time (3.5 ms). All models achieved accuracies >94% indicating that our method is resilient to changes in architecture. The chosen architecture will depend on the desired tradeoff between accuracy, memory size, and speed, but we judged the optimal approach considering all factors to be the ResNet18 architecture (E5). ResNet18 has just slightly lower accuracy (96.3%) than NasNet-A-Mobile but is more than 10 times as fast.

The remainder of the results are presented using the single network from E5 with the best performance on its validation set. Using validation sets is not a fair comparison since every network has a different validation set. However, as shown in Table 1, the variance between results is quite low so the choice of network does not significantly affect the results. The confusion matrix on the test set for this network is shown in Fig. 7. Tricuspid Valve (TV) had the lowest accuracy with 93%. Despite the uneven distribution of training data between classes, the method does not overfit to the classes with more images.

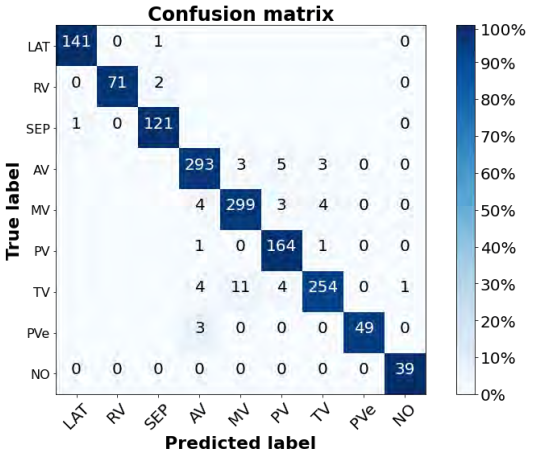


Fig. 7. Confusion matrix on the final test set using E5 The boxes where no number is shown will never be misclassified because separate classification heads are used. Colors are normalized to class size (percentages). Abbreviations are described in Fig. 2.

B. Decision Tree Mapping

Using the decision tree presented in Fig. 6 the output of the network was mapped to the final measurement type classes. The accuracy for each type is shown in Table 3. To check what kind of mistakes were occurring, an error analysis was performed for the two measurement types with the lowest accuracy: Aortic Regurgitation (AR) and Tricuspid Regurgitation (TR). For AR, there were 4 misclassified images. Of these, 3 were acquired from the apical parasternal long axis view (A full description of echocardiographic views is available in [1]). It is logical the network might miss these cases since there were few AR images acquired from this view in the training set because AR measurements are typically taken from the apical 5 chamber view. However, operators at different clinics may have different preferences, leading to the discrepancy between training and test sets. The last image was judged to have been misclassified by the labeler on re-analysis.

There were 16 misclassified TR images. Of these misclassifications 13 images were from the right ventricle inflow view. This view was not included in the training set because the labeler for the training set did not have experience with this view and thus ignored those images. Therefore, the network never learned the patterns associated with these images. For one of the remaining images the class could not be determined during re-analysis, it was initially classified as TR because TR measurements were encoded in the file. The

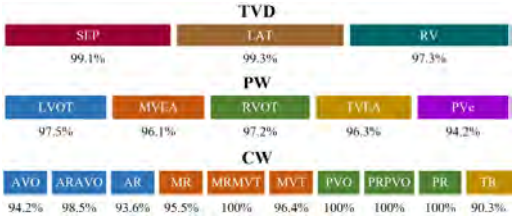


Table 3. Classification accuracy for the selected network for each measurement type sorted by mode. Abbreviations are described in Fig. 2.

remaining 2 images were simply missed by the network.

C. Confidence Metric

To test the validity of the proposed confidence metric, the pre-softmax set of cutoff values were extracted. Quantile cutoff limits were extracted for each class from 0%-10% quantiles in 0.2% step sizes. The quantile is the ignored percentage on the validation set: a quantile of 5% indicates that the 5% of images with the lowest pre-softmax values would be ignored. Quantiles were tested only on the single network but were obtained by averaging across all five networks and validation sets to enhance the robustness.

The test set was split into two pieces, the first containing the 1431 correctly classified images (*correct* set) from E5 and the second containing the 51 misclassifications (*incorrect* set). The quantiles were used when running inference on these two sets of images as well as the full *test* set and the *unknown* and *extra* sets put aside during labeling. For each set, the ignored rate was recorded while iterating through the quantile values. For the test set, the error rate was also recorded. Results are shown in Fig. 8, with ignored rates on the left axis and error rate (in red) on the right axis.

The confidence metric results may indicate some overfitting, as well as that the validation and test sets came from different distributions. If the network is not overfit and the images are from the same distribution, the quantile should map 1:1 to the ignored test percentage. However, results showed at the 0.5% quantile 5% of the test images are ignored. A difference in distribution is expected since the images came from separate clinics and matches what was found in the error analysis.

Results also demonstrated the confidence metric accurately detected which images came from outside our training distribution and eliminated misclassified images at a much higher rate than correctly classified ones. For example, a user setting the confidence threshold at the 8% quantile mark would have to manually label 16.5% of their images but would achieve an accuracy of 99% on those images automatically classified because ~80% of the *incorrect* images would be ignored.

Both the *unknown* and *extra* image sets were also ignored at a much higher rate, confirming that the metric identified out of distribution samples. Ignored rates showed an approximately logarithmic relationship with increasing quantile values for all three of the *incorrect, unknown,* and *extra* sets. For all three, >20% of images were ignored at the 1% quantile.

Results were compared to the MC-dropout versions of the model. Each MC-dropout model was run 100 times, and quantile limits were set for values from the pre-softmax layer

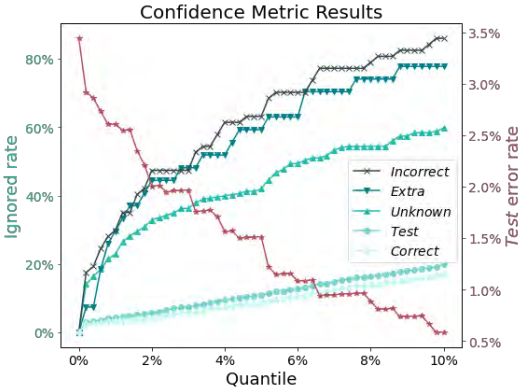


Fig. 8. Results of confidence metric experiments with ignored rates (green lines) on the left axis and error rate of the test set (red line) on the right axis. Ignored rate refers to the percentage of images which the network did not label because the output was beneath the cutoff for that quantile. *Test* refers to the test set of images. *Correct* is the subset of test images correctly classified by the network in E5 and *incorrect* is the subset of misclassifications. *Unknown* is the set of images which were unidentified during labeling. *Extra* is the set of images from classes not included in this classification scheme. An ideal confidence metric should completely ignore the *incorrect* and *extra* sets while ignoring none of the *correct* set. Ignoring the *unknown* set indicates the network accurately reflects expert confidence, but it is possible the network has detected features unseen to the observer. Best viewed in color.

and softmax layer of the normal model, and from the mean and variance of the pre-softmax and softmax layers from the MC-dropout model. Results were similar for all implementations, with a slightly higher ignored rate for all sets when using the pre-softmax layer from either model. These results indicate that the choice of how to extract quantile values does not play a large role in the resulting confidence metric. The advantages of using the pre-softmax values are that it is simpler to implement and no additional experiments need to be run.

Metrics from the ensembled approaches were also compared to selected quantiles of the proposed confidence metric in experiments C1-C5 in Table 4. The ensemble networks contained all five originally trained models. The predicted class came from either the majority vote (C4) or only images where all networks agreed while other images were ignored (C5). These ensembled approaches were compared to representative quantiles from Fig. 8: the 0% quantile (C1 – the single chosen network from E5), the 5.2% quantile (C2 – selected as the closest error rate to C5), and the 8% quantile (C3 – first quantile with error <1%). Using majority voting (C4) provided a small improvement in error rate (3.2%) compared to C1 (3.4%) without ignoring any images. Selecting only matching outputs (C5) provides a significant decrease in error rate (1.3%) although 4.8% of images were ignored. To achieve an equivalent error rate with the proposed confidence metric (C2), more than twice the number of images were ignored. The downside of the ensembled approach is five networks must be initialized, significantly increasing the memory consumption.

D. Implementation Details

All parameters were extracted automatically from private Dicom tags using Python 3.6. Pre-processing, training, and

testing were carried out on an Ubuntu machine with Python 3.6, PyTorch 0.4, and an NVIDIA Titan X GPU. All networks were trained for 60 epochs. The learning rate was set to 0.1 for all experiments other than E9 where it had to be reduced to 0.01 to get the network to converge. Learning rates were reduced by a factor of 10 every 20 epochs. All experiments used standard gradient descent with momentum (coefficient 0.9) and weight decay (0.0005). We used normalization to center the dataset during training and inference. The B-mode images were normalized to [0,1]. The heatmaps were already in the range [0,1] because they are probability maps., Both images and heatmaps were mean-centered by subtracting the mean value of all pixels in all images in the training set. This value was 0.3 for the B-mode image and 0.0062 for the heatmap image. Cross-entropy loss was used for all experiments and the classes were not weighted since the results did not suffer from class imbalance. Preliminary experiments with a weighted loss function did not improve results. While training the multi-head network, λ_1 and λ_2 (loss function hyperparameters from (1)) were set to 0.18 and 0.82 respectively, but results were not sensitive to changes of λ's within normal ranges. TVD classification is an easier task (as shown in Table 1) so $\lambda_2 > \lambda_1$ even though there were fewer TVD images.

IV. Discussion

Our results indicate that highly accurate Doppler spectrum measurement type classification is possible in echocardiography *without* using the spectrum data. Accurate classification despite the omission of the Doppler spectra proves the network is learning the relationship between user input and anatomical structures. The spectrum data can be ignored because each class correlates to a unique physical location within the heart after using our mapping scheme. Note that since each Doppler spectrum can be acquired from a variety of different views, this does not imply each class corresponds to a unique location in the input image. Accurate classification requires understanding of both the B-mode image and the cursor location. Highly accurate results have already been achieved on view recognition tasks in echocardiography (e.g. [7], [8]) indicating effective understanding of the B-mode image through CNNs. Our results take this a step further. We show heatmaps are an effective way to encode physical location information for CNNs, demonstrating the ability to connect anatomical structural information (B-mode image) to relevant user input (Doppler cursor location) in a classification.

Since deep-learning algorithms deployed in clinical settings must frequently compete for resources, methods for decreasing resource utilization were analyzed. Results demonstrated that a multi-head classification could reduce the memory footprint when the classification task can be split into separate problems by external parameters. The multi-head networks (E5/E6) maintained similar accuracy levels to those of separate networks (E3) and higher accuracy than a single network trained with all classes (E4). The final implementation achieved sub-4*ms* inference time, indicating near real-time performance.

Our approach demonstrated high accuracies across varying network architectures. Additionally, training several networks from differing dataset splits showed consistent results with low variances in F_1 scores and accuracies. The repeatability of our results across architectures and data splits is another strength of the contributions.

#	Architecture	Error rate	Ignored rate	Size (MB)	Time (ms)
C1	Single ResNet18, 0% quantile	3.4%	0%	**740**	**3.5**
C2	Single ResNet18, 5.2% quantile	1.2%	11.4%	**740**	**3.5**
C3	Single ResNet18, 8% quantile	0.9%	16.5%	**740**	**3.5**
C4	Ensemble ResNet18, Majority vote	3.2%	0%	3700	**3.5**
C5	Ensemble ResNet18, Matching output	1.3%	4.8%	3700	**3.5**

Table 4. Comparison of selected quantiles to ensemble methods using 5 networks. Size and time estimates for ensemble approaches assume that inference can be run for all networks simultaneously which depends on the implementation.

We also conducted an error analysis of the mistakes made by the network. Encouragingly, the error analysis showed the network is accurately learning the image and heatmap patterns included during training. The errors seen were mostly due to differences between the training and test sets in echo views or mismatch in labeling practices. Because the network accurately learned the patterns it was exposed to, accuracy could continually be improved by gathering additional training data that covers misclassified cases.

Misclassifications can be costly in a medical setting. They can lead to confusion when analyzing patient data and mistrust in artificial intelligence-based tools. To attempt to reduce misclassifications, several measures were taken. First, a No Organ (NO) class was included in the training dataset to avoid classifying images of air and gel into another class. Second, cutoff limits were set based on output values from the last fully connected ("pre-softmax") layer for each class. Images with a score below the cutoff were ignored rather than classified. Overall, results indicated that the proposed confidence metric can significantly reduce the error rate by ignoring missed images in the test set at a much higher rate than correctly classified ones. The confidence metric also ignores the *unknown* and *extra* image sets at an approximately three times higher rate than those from the test set. Our method demonstrates one way to handle inputs from unseen distributions in a classification problem. Moreover, it allows a user to easily set the quantile limit depending on the desired tradeoff between the error and ignored rates.

Results testing ensemble networks showed these methods ignore fewer images for the same error rate compared to quantile cutoffs. Ensemble methods are a more robust confidence predictor for environments without resource constraints. Moreover, ensemble methods could easily be combined with the quantile cutoff approach discussed above to provide a robust, tunable ignored vs. error rate tradeoff.

In future work, we hope to extend this method to public Dicom data (and thus multi-vendor). This is much easier because we don't use the Doppler spectra in our pipeline; the B-mode image, Doppler cursor location, mode, and baseline are

the only data needed. Although we extracted this information from raw data in the present work, all of it is also available from public Dicom tools. The information could thus be extracted from any vendor's Doppler data. The main difference in using publicly available data (and thus scan-converted images) is some overlay on the public B-mode image, potentially including color-flow. However, with a little effort we anticipate the ability to overcome this and make our method fully multi-vendor.

V. Conclusion

In this work, we demonstrated a CNN-based method for the automated classification of Doppler measurement types. An example integration within a clinical workflow is presented in Appendix C of the supplementary material. Notable performance gains were shown on the task by encoding the Doppler cursor as a heatmap and introducing a post-processing mapping scheme to simplify the problem. These methods would also be applicable to other tasks including location information as an input parameter and/or with linearly separable classes. We design a confidence metric capable of discarding a large proportion of images with high uncertainty to create a more reliable classification system. Our method performs fast and accurate classification of Doppler measurement types. In the same way automatic echocardiographic view recognition unlocked fully automated processing of many B-mode images, our work unlocks fully automated Doppler spectra analysis, bringing increased efficiency and statistical power to clinical workflows.

Acknowledgement

The authors thank Pablo Lamata and Julia Schnabel for discussions and valuable feedback on the methods.

Supplementary Materials

Supplementary materials are available here: https://adgilbert.github.io/cardiac-doppler-type-classification/

References

[1] C. Mitchell *et al.*, "Guidelines for Performing a Comprehensive Transthoracic Echocardiographic Examination in Adults: Recommendations from the American Society of Echocardiography," *J. Am. Soc. Echocardiogr.*, vol. 32, no. 1, pp. 1–64, 2019.

[2] A. Alfirevic, "Back to the Future—Importance of Spectral Doppler," *J. Cardiothorac. Vasc. Anesth.*, vol. 33, no. 5, pp. 1467–1470, 2018.

[3] L. Mo, D. Becker, K. McCann, M. Honda, and S. Ishiguro, "Method and Apparatus for Automatic Tracing of Doppler Time-Velocity Waveform Envelope," 5,935,074, 1999.

[4] U. Lempertz and E. Sokulin, "Cardiac Auto Doppler White Paper," 2017.

[5] L. Deng *et al.*, "Recent advances in deep learning for speech research at Microsoft," in *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc.*, 2013, pp. 8604–8608.

[6] I. A. Wright, N. A. J. Gough, F. Rakebrandts, M. Wahabs, and J. P. Woodcock, "Neural network analysis of blood flow Doppler ultrasound signals: a pilot study," *Eur. J. Ultrasound*, vol. 6, no. 5, p. S11, 2004.

[7] A. Madani, R. Arnaout, M. Mofrad, and R. Arnaout, "Fast and accurate view classification of echocardiograms using deep learning," *NPJ Digit. Med.*, vol. 1, no. 1, pp. 1–8, Dec. 2018.

[8] J. Zhang *et al.*, "Fully Automated Echocardiogram Interpretation in Clinical Practice," *Circulation*, vol. 138, no. 16, pp. 1623–1635, 2018.

[9] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," pp. 1–9, 2011.

[10] A. Ephrat *et al.*, "Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation," *arXiv:1804.03619*, 2018.

[11] G. Litjens *et al.*, "A Survey on Deep Learning in Medical Image Analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.

[12] N. Tajbakhsh *et al.*, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[13] S. Falkner, A. Klein, and F. Hutter, "BOHB: Robust and Efficient Hyperparameter Optimization at Scale," *arXiv:1807.01774*, 2018.

[14] Yarin Gal, "Uncertainty in Deep Learning," 2016.

[15] Y. Gal and G. Zoubin, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," *arXiv:1506.02142*, 2015.

[16] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *Proc. of the 34th Int.Conf. on Machine Learning*, 2017, pp. 1321–1330.

[17] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proc. of the 8th Int. Conf. on Knowledge Discovery and Data Mining*, 2002, pp. 694–299.

[18] N. B. Schiller *et al.*, "Recommendations for Quantitation of the Left Ventricle by Two-Dimensional Echocardiography," *J. Am. Soc. Echocardiogr.*, vol. 2, no. 5, pp. 358–367, 1989.

[19] K. K. Kadappu and L. Thomas, "Tissue doppler imaging in echocardiography: Value and limitations," *Hear. Lung Circ.*, vol. 24, no. 3, pp. 224–233, 2015.

[20] R. Liu *et al.*, "An Intriguing Failing of Convolutional Neural Networks and the CoordConv Solution," in *Advances in Neural Inf. Proc. Sys.*, 2018.

[21] A. Nibali, Z. He, S. Morgan, and L. Prendergast, "Numerical Coordinate Regression with Convolutional Neural Networks," *arXiv:1801.07372*, 2018.

[22] S. Ruder, "An Overview of Multi-Task Learning in Deep Neural Networks," *arXiv:1706.05098*, 2017.

[23] Y. Geifman and R. El-Yaniv, "Selective Classification for Deep Neural Networks," in *Advances in Neural Inf. Proc. Sys.*, 2017.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, 2012.

[25] C. Szegedy *et al.*, "Going deeper with convolutions," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, vol. 07-12-June.

[26] C. J. van Rijsbergen, *Information Retrieval*. London: Butterworths, 1979.

[27] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark Analysis of Representative Deep Neural Network Architectures," *IEEE Access*, vol. 6, 2018.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[29] B. Zoph, V. Vasudevan, J. Shlens, and Q. V Le, "Learning Transferable Architectures for Scalable Image Recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8697–8710.

[30] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[31] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size.," *arXiv:1602.07360*, 2016.

[32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520.

[33] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger,

"Densely connected convolutional networks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, vol. 2017-Janua, pp. 2261–2269.

[34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning (ICML)*, 2015, vol. 1, pp. 448–456.

[35] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual path networks," in *Advances in Neural Information Processing Systems*, 2017, vol. 2017-Decem, pp. 4468–4476.

[36] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions."

[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, vol. 2016-Decem, pp. 2818–2826.

Supplementary Materials for User-Intended Doppler Measurement Type Prediction Combining CNNs With Smart Post-Processing

Appendix A
Measurement Type Desciption

A detailed description of all the Doppler measurement types referenced in the article is provided below. For full information on the relevant measurements as well as the referenced echocardiography (echo) views the reader is referred to the America Society of Echocardiography guidelines [1].

Aortic Valve (AV): The aortic valve can be measured from either Apical 5 Chamber (A5C) or Apical Long Axis (APLAX). There are several measurements in the aortic valve region:

- **Aortic Regurgitation (AR):** A measure of the severity of the regurgitant flow across the aortic valve measured in Continuous Wave (CW) Doppler. The regurgitant flow is in the positive part of the spectrum, so the baseline is typically shifted down.
- **Aortic Valve Outflow (AVO):** A measure of the outflow through the aortic valve measured with CW Doppler. The outflow is in the negative part of the spectrum, so the baseline is typically shifted up.
- **AR / AVO (ARAVO):** If the baseline is unchanged then either AR or AVO can be measured in the image.
- **Left Ventricle Outflow Tract (LVOT):** A measure of the velocity of blood flow through the LVOT measured with Pulsed Wave (PW) Doppler. This can be combined with a 2D measurement of LVOT area to give the cardiac output, a common measure of left heart function.

Mitral Valve (MV): The mitral valve can be measured from either Apical 2 Chamber (A2C), Apical 4 Chamber (A4C) view, or APLAX. There are several measurements in the mitral valve region:

- **Mitral Regurgitation (MR):** A measure of the severity of the regurgitant flow across the mitral valve measured with CW Doppler. Blood flow is in the opposite direction as the aortic valve so regurgitant flow is in the negative part of the spectrum (baseline shifted up).
- **Mitral Valve Trace (MVT):** A measure of the inflow through the mitral valve with CW Doppler, called trace because the measurement consists of tracing the outline of the blood flow. The blood flow is in the positive part of the spectrum (baseline shifted down).
- **MR / MVT (MRMVT):** If the baseline is unchanged then either MR or MVT can be measured in the image.
- **Mitral Valve E/A (MVEA):** A measure of left ventricle function given by the ratio of peak velocity of blood flow across the mitral valve during the early diastole phase of the heart cycle (E wave) to the peak velocity during the atrial contraction (A wave). The E/A ratio is measured with PW Doppler.

Pulmonary Valve (PV): The pulmonary valve can be measured from the Parasternal Short Axis (PSAX) view. There are several measurements taken in the pulmonary valve region:

- **Pulmonary Regurgitation (PR):** A measure of the severity of the regurgitant flow across the pulmonary valve measured with CW Doppler. Like the aortic valve measurements, the regurgitant flow is in the positive part of the spectrum so the baseline is shifted down. **Pulmonary Valve Outflow (PVO):** A measure of the outflow through the pulmonary valve measured with CW Doppler. The outflow is in the negative part of the spectrum, so the baseline is typically shifted up. **PR / PVO (PRPVO):** If the baseline is unchanged then either PR or PVO can be measured in the image. **Right Ventricle Outflow Tract (RVOT):** A measure of the velocity of blood flow through the RVOT measured with PW Doppler. Like LVOT measurements, this can be combined with 2D measurements of RVOT area to give a metric for right heart function.

Tricuspid Valve (TV): The tricuspid valve can be measured from A4C, PSAX, or Right Ventricle Outflow view which is similar to Parasternal Long Axis (PLAX). There are several measurements taken in the tricuspid valve region:

- **Tricuspid Regurgitation (TR):** A measure of the severity of the regurgitant flow across the tricuspid valve measured with CW Doppler. The regurgitant flow is typically in the positive part of the spectrum so the baseline is shifted down. However, since this is the only CW measurement for this valve class, the baseline is not considered in our approach.
- **Tricuspid Valve Inflow (TVEA):** Like MVEA, but for the right side of the heart. TVEA is a measure of the ratio between peak velocity at the early diastole phase (E wave) to peak velocity during the atrial contraction (A wave) measured with PW Doppler.
-

Pulmonary Vein (PVe): The pulmonary vein can be measured from the A4C view and PW Doppler is used to measure pulmonary venous blood flow velocities.

Septal Tissue Doppler (SEP): The motion of the basal septal wall can be measured from A4C using Tissue Velocity Doppler (TVD) to assess systolic and diastolic function.

Lateral Tissue Doppler (LAT): The motion of the basal lateral wall can be measured from A4C or A5C using TVD to assess systolic and diastolic function.

Right Ventricle Tissue Doppler (RV): The motion of the basal right ventricular wall can be measured from A4C or A5C using TVD to assess systolic and diastolic function.

No Organ (NO): A image containing no tissue information and random Doppler cursor locations (and thus random Doppler signals) so the network will not misclassify images where the probe has not yet been placed on the body.

Three additional measurement classes were found: **Ascending Aorta**, **Descending Aorta**, and **Hepatic Vein**, but were not included in this measurement classification problem because they are used infrequently in clinical practice.

Appendix B
Network Implementations

DenseNet, GoogLeNet, Inception-v3, MobileNet-v2, ResNet, ShuffleNet, and SqueezeNet architectures were implemented following the architectures available from the PyTorch torchvision library [1]. BN-Inception, DualPathNetwork, NasNet-A-Mobile, and Xception architectures were implemented following the benchmark analysis GitHub repository[2] available from Bianco et. al. [2]. For all networks the first convolutional layer was changed to have 2 input channels instead of 3. All networks were trained from scratch since preliminary experiments with pretrained networks (excluding the first layer since it is a different size) yielded worse results. GoogLeNet and Inception-v3 contain auxiliary paths that can be used to speed up training and were weighted equally in our implementation. The input images for the Inception-v3 network were changed to 299×299 (resized after 224×224 crop) following the constraints of that architecture.

Appendix C
Example Integration of Automatic Measurement Type Recognition

Below is an example of how the automatic measurement type recognition introduced in this manuscript can increase the efficiency of clinical workflows. Fig. 1 shows an example workflow for conducting a Doppler Spectrum measurement before and after including automatic measurement type recognition. The user must manually click through menus to find the corresponding measurement without measurement type recognition, while that step is performed automatically once the algorithm is implemented.

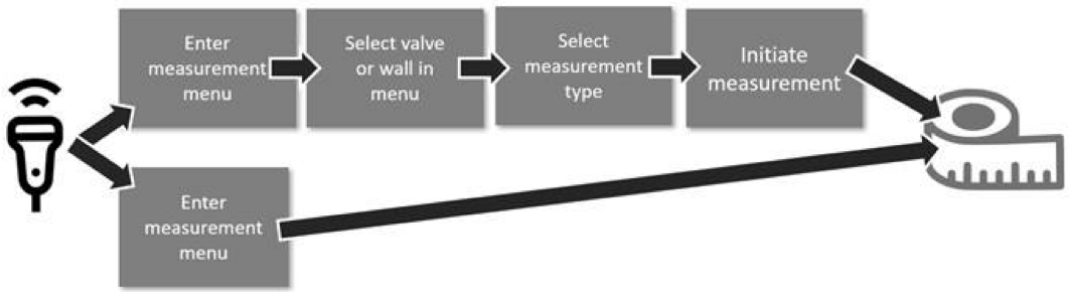


Fig. 1. An example measurement workflow *without* (above) and *with* (below) automatic measurement type recognition

References

[1] C. Mitchell, P. S. Rahko, L. A. Blauwet, B. Canaday, J. A. Finstuen, M. C. Foster, K. Horton, K. O. Ogunyankin, R. A. Palma, and E. J. Velazquez, "Guidelines for Performing a Comprehensive Transthoracic Echocardiographic Examination in Adults: Recommendations from the American Society of Echocardiography," *Journal of the American Society of Echocardiography*, vol. 32, no. 1, pp. 1–64, 2019.

[2] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark Analysis of Representative Deep Neural Network Architectures," *IEEE Access*, vol. 6, 2018. [Online]. Available: https://arxiv.org/pdf/1810.00736.pdf

[1] https://pytorch.org/docs/stable/torchvision/models.html

[2] https://github.com/CeLuigi/models-comparison.pytorch

Paper II

Automated Left Ventricle Dimension Measurement in 2D Cardiac ultrasound via an Anatomically Meaningful CNN Approach

Andrew Gilbert, Marit Holden, Line Eikvil, Svein Arne Aase, Eigil Samset, Kristin McLeod

Published in *Smart Ultrasound Imaging workshop at MICCAI 2019*[a]*. Lecture Notes in Computer Science*, 2019, volume 11798, pp. 29-37. DOI: 10.1007/978-3-030-32875-7$_4$.

[a]Best presentation award

II

Automated left ventricle dimension measurement in 2D cardiac ultrasound via an anatomically meaningful CNN approach

Andrew Gilbert[1,2], Marit Holden[3], Line Eikvil[3], Svein Arne Aase[1], Eigil Samset[1,2], and Kristin McLeod[1]

[1] GE Vingmed Ultrasound, GE Healthcare
[2] Department of Informatics, University of Oslo
[3] Norwegian Computing Center

Abstract. Two-dimensional echocardiography (2DE) measurements of left ventricle (LV) dimensions are highly significant markers of several cardiovascular diseases. These measurements are often used in clinical care despite suffering from large variability between observers. This variability is due to the challenging nature of accurately finding the correct temporal and spatial location of measurement endpoints in ultrasound images. These images often contain fuzzy boundaries and varying reflection patterns between frames. In this work, we present a convolutional neural network (CNN) based approach to automate 2DE LV measurements. Treating the problem as a landmark detection problem, we propose a modified U-Net CNN architecture to generate heatmaps of likely coordinate locations. To improve the network performance we use anatomically meaningful heatmaps as labels and train with a multi-component loss function. Our network achieves 13.4%, 6%, and 10.8% mean percent error on intraventricular septum (IVS), LV internal dimension (LVID), and LV posterior wall (LVPW) measurements respectively. The design outperforms other networks and matches or approaches intra-analyser expert error.

Keywords: ultrasound, echocardiography, landmark detection, deep learning, convolutional neural networks

1 Introduction

Ultrasound imaging is the primary imaging modality used to assess cardiac morphology and function. Compared to other imaging modalities (e.g. MRI and CT), ultrasound imaging has a lower cost, is easier to perform, and, unlike CT, does not produce ionizing radiation. This makes it ideally suited for rapid diagnostic use for patients with cardiovascular disease. A diagnosis is made by acquiring a set of images from different views of the heart and extracting measurements of heart function from those images. Some of the most frequent measurements in patient care settings are measurements of the left ventricle (LV)

from the parasternal long-axis view. The typical set of measurements consists of the length of the intraventricular septum (IVS), left ventricular internal dimension (LVID), and left ventricular posterior wall (LVPW) at both the end-diastole (ED) and end-systole (ES) phases of the cardiac cycle. Several examples of these measurements are shown in Fig. 2. Because LV dimension measurements are performed frequently, automated measurement tools could provide tremendous time savings for clinical use.

Despite its widespread use, there is a high variability in LV dimension measurements due to variations in training and the difficulty of precisely detecting relevant structures. The 2010 HUNT study [11] measured inter-analyser (difference between experts reading the same exam) and intra-analyser (difference between the same expert reading the same exam several weeks apart) for several standard echocardiographic measurements. The intra-analyser mean percent error (MPE) for IVS, LVID, and LVPW measurements was 10%, 4%, and 10% respectively and inter-analyser results were similar. For IVS and LVPW measurements this corresponds to about half of the standard deviation of normal ranges [3] so a patient on the borderline could easily be put in a different diagnostic group. The high variability highlights the difficulty of the task at hand, but effective automation is one promising approach to reduce this variability and implement a more reproducible diagnostic pipeline.

Previous work on 2D ultrasound measurements has focused on individual measurements. Snare et al. used deformable models with Kalman filtering to outline the septum shape [9], achieving bias and standard deviation of 0.14 ± 1.36 mm for automated IVS measurements compared to manual measurements. Baracho et al. used perceptron style neural networks and filtering to generate a septum segmentation [1]. They achieved results of $0.5477\text{mm} \pm 0.5277\text{mm}$ for IVS measurements but failed to validate directly against measurements from an expert cardiologist. Finally, Sofka et al. developed an automated method for detecting LVID measurements using convolutional neural networks (CNNs) [10]. Sofka et al. introduce a center of mass layer to regress keypoint locations and achieved a 50th percentile error of 4.9% and a 95th percentile error of 18.3%.We extend the work of Sofka et al. by targeting the IVS and LVPW measurements in addition to LVID. Including more measurements increases the difficulty of the task because the network should not only achieve high accuracy on all measurements but also find measurement vectors that have a logical relationship to each other (i.e. all measurement vectors should be parallel to follow clinical guidelines). Additionally, the upper IVS and lower LVPW endpoints do not fall at distinct gradient boundaries within the image making them more difficult to find, even for an expert.

As with Sofka et al., we frame the task as a landmark detection problem, where the goal is to identify 6 key points (the 2 endpoints of IVS, LVID, and LVPW measurements) from an input image. A landmark based approach was chosen to increase user-interpretability and allow editing of the found points by users in a clinical workflow. Many architecture variants have been applied in previous work on landmark detection problems, but the most common approach

is to generate a heatmap of likely locations for each key point of interest [6, 7, 12]. The heatmap is directly compared to a reference heatmap generated from the key point's known location, or the coordinates of the key points are regressed from the heatmap and compared to known coordinates.

We propose several modifications to the general landmark detection strategy above because, in contrast to facial recognition, there is no defined local appearance of these landmarks. Instead, their location is determined from local appearance and global structural information. For example, while the septum typically extends through a large part of the image, ASE guidelines recommend measuring at the level of the mitral valve leaflets [4] which means an algorithm needs to be aware of structural information to find the correct IVS endpoints.

The novelty of our approach lies in it's ability to handle these challenges and achieve high accuracy. First, we generate anatomically meaningful ground truth heatmaps which follow the expected spatial distribution of the point. Second, we propose the integration of coordinate convolution layers [5] within feature detection networks for medical imaging. Third, we optimize network performance using a multi-component loss function which provides feedback to the network in multiple components including measurement endpoint coordinate locations, angle of measurement, and measurement distances. Including all these terms allows us to optimize for both measurement accuracy and a logical relationship between measurement vectors. Finally, we evaluate several different architectures within the constraints of our first two contributions to show the optimal architecture for the given task.

2 Methods

2.1 Network

The input to the proposed network is a single 2D frame. The accurate detection of ES and ED frames from a full cardiac loop is left for future work. The image is first passed through a CoordConv layer, which adds pixel-wise spatial location information to allow CNNs to more easily find objects [5]. The core of our approach is a U-Net [8]. A U-Net is a CNN with a sequence of down and up sampling paths with skip connections concatenating each down-sampling output to the corresponding up-sampling level. In each successive down-sampling layer, the number of filters doubles and the spatial resolution in each dimension is cut in half, while the reverse is true in up-sampling. We make several modifications in our implementation. The number of down-sampling levels and the number of filters are parameterized to tune the network. Padding is added on all layers to ensure output heatmap resolution matches the input. Batch normalization and spatial dropout layers are included between convolutional blocks for regularization, avoiding standard dropout since neighboring pixels are strongly correlated [12]. Each convolutional layer uses a kernel size of 3x3.

Our output is the same size as the original image but contains 6-channels, with each channel representing a heatmap corresponding to one landmark. Although

the top and bottom endpoints of LVID typically match the bottom of IVS and top of LVPW respectively, they can be different for some pathologies which is the reason they are independent points in our framework. Each channel is normalized to be a probability map and passed through a differential spatial-numerical transform block [7] to calculate the center of mass in x and y: the endpoints of the three measurement vectors. From the coordinate endpoint locations, we calculate the final distance measurements. The network architecture is shown in Fig. 1.

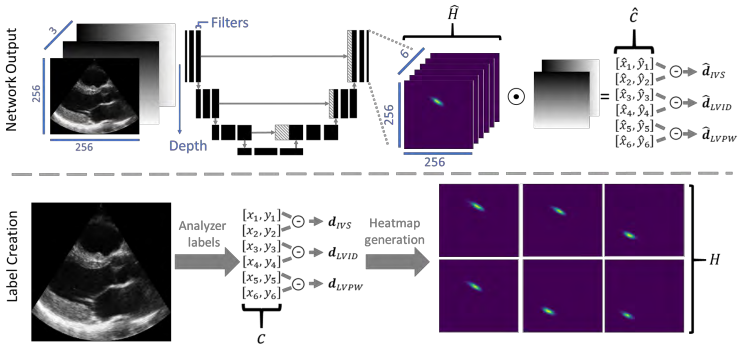


Fig. 1. Network architecture. The input image (256x256) is appended with x and y coordinate channels to create a 3 channel image and passed through a U-Net-based architecture. The output contains 6 heatmaps ($\hat{H}$), one for each detected landmark. The center of mass of each heatmap is extracted as the found coordinates ($\hat{c}$), and vectors for each measurement are obtained ($\hat{\boldsymbol{d}}$). Label distances ($\boldsymbol{d}$) and heatmaps (H) are generated from labeled endpoints (c) to compare to the network output.

2.2 Loss Function

Our labels are the coordinate locations of all caliper endpoints. We extrapolated these to match the network output including heatmaps of coordinate locations, and distances between coordinate pairs. For the label heatmaps, a 2D gaussian is centered at the location of the labeled coordinate. The gaussian is elongated in one dimension with a ratio of 20 to 1 between the variances of the long and short axes and rotated such that the long axis was orthogonal to the direction of measurement (see H in Fig. 1 for example). This both followed the expected spatial distribution of the points and gave the network feedback that a miss orthogonal to the direction of measurement was more acceptable than one parallel to the measurement, which would substantially affect measurement results. The variance of the gaussian in the long axis is 14 pixels (or 5% of the image size).

L2 loss is used for the six coordinate locations and three distance measurements, although the distance loss was divided by the relative actual distance ($\boldsymbol{d}$) to equally weight each measurement. The heatmap loss is the root mean squared error (RMSE) between the generated and output heatmaps, following Newell et al. [6]. The heatmap loss helps the network converge to a reasonable result quickly, because feedback is provided to the network at every pixel in the output, rather than just a single metric fed back to all pixels such as with the distance or coordinate measures. The difference in the relative angles of the measurement vectors is also included in the loss function as the cosine similarity between the two vector sets. Including the angle loss is critical because even if the network can correctly find point delineations across the relevant structure (e.g. septum), if the measurement vector is not orthogonal to that structure then the measurement will be overestimated. The angle and coordinate loss also help promote a logical relationship between measurement vectors.

3 Experiments

3.1 Datasets and Pre-processing

LV intraventricular septum (IVS), internal diameter (LVID), and posterior wall (LVPW) dimensions were annotated in parasternal long axis 2DE scans. To avoid overfitting to a single acquisition protocol, exams were collected from four sites. All measurements were performed by a single cardiologist experienced in 2DE measurements. Diagnostic information was stripped from the images, but a mix of normal patients and varied pathologies is typical for the chosen sites. Exams were labeled at ED and ES except for where image quality in one phase prohibited accurate measurements. A total of 585 images were gathered from 309 unique patients. To generate a comparison with intra-analyser variability, 32 recordings (mixed ED and ES) were labeled multiple times by the same expert. These 64 images were set aside to be used as the test set for the network leaving 521 images for training and validation. The training, validation, and test sets were split such that images from the same patient would always remain in the same set. The coordinates and image data from the relevant frames were extracted from the stored files and converted to 256x256 one-channel images.

During training, random brightness, contrast, and gamma transformations were applied to each image. Additionally, we used mean normalization and applied random translations of 0 to 40 pixels in each direction, while ensuring coordinate locations were never within 16 pixels of the image boundaries.

3.2 Implementation Details

The network was implemented using PyTorch 0.4.1 with Python 3.6 on an Ubuntu 18.04 machine with an NVIDIA Titan X GPU. The batch size was 16 images for training and 4 images for validation. We trained for 120 epochs and reduced the learning rate by a factor of 10 every 50 epochs. Using 10% of the training set for

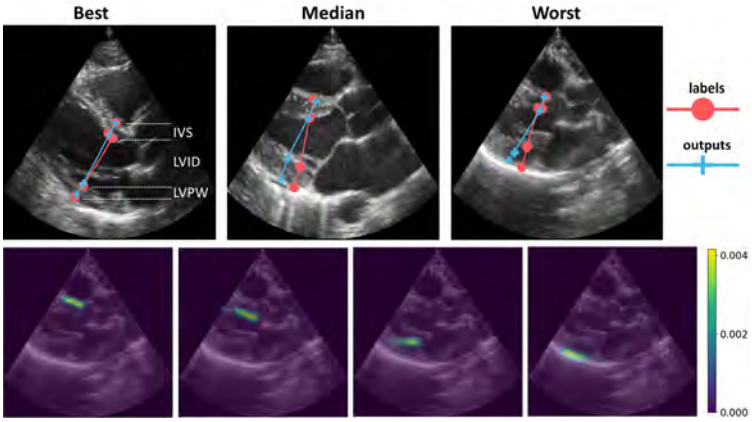


Fig. 2. Top row: Qualitative results on the best, median and worst images from the test set showing expert labels and network outputs for each measurement. Bottom row: Characteristic heatmaps showing how the network learns to prioritize a small distribution in the direction parallel to the measurement direction. Only four heatmaps are shown for simplicity since the top and bottom LVID endpoints overlap with the bottom of IVS and top of LVPW respectively and produce very similar heatmaps.

validation of hyperparameters, we found 4 levels was the optimal network depth and 2^6 was the optimal number of filters in the first layer.

The primary metric important for clinical use is the accuracy of the distances for each of the three measurements. The coordinate locations of the endpoints and angle of the measurement vectors are secondary metrics that are important to create a tool that accurately follows clinical guidelines. For clinical use, it is not important that the generated heatmap matches the artificial heatmap. However, we found that keeping the relative weighting of the heatmap loss high compared to the other metrics helped improve network accuracy on all metrics.

3.3 Evaluation and Comparison

The primary metric for evaluation was the mean percent error between the network output and ground truth distance measurements on IVS, LVID, and LVPW. The test set was composed of the 32 images that had been labeled multiple times. The median of the two labels was set as ground truth although comparing to a randomly chosen label yielded very similar results.

While much of the strategy revolved around pre- and post- processing, we implemented several other networks in addition to U-Net for comparison. Results were compared to a stacked hourglass network [6], which currently obtains state of the art results on the FLIC and MPII human pose estimation metrics as well as

ResNet18, ResNet34, and ResNet50 networks [2]. We tuned the number of stacks (4) and blocks (2) of the stacked hourglass network on the validation set. We implemented the ResNet networks following the strategy proposed by Nibali et al. [7], reducing the stride in several layers to increase output heatmap resolution, while using dilated convolutions to maintain receptive field sizes . The output heatmap size for the ResNet and stacked hourglass networks was 64x64 and we appended up-sampling layers to achieve 256x256 resolution. A CoordConv layer was added to the beginning of all networks and the same coordinate regression method and loss function were used. For a fair comparison to the other networks, results with default values of an out-of-the-box implementation of U-Net is included (no batch normalization or dropout, depth and number of filters set to 5 and 2^6 respectively).

4 Results

The best, median, and worst examples (in terms of RMSE) from the test set are shown in Fig. 2. The network achieves intra-analyser accuracy on LVPW and LVID measurements, and slightly worse than intra-analyser on IVS measurements. The algorithm's worse performance on IVS measurements possibly occurs because the upper septum is often not defined as a clear gradient boundary because the septum blurs together with trabeculae in this region (see median image in Fig. 2, although the network correctly found the location in this case). Expert labelers typically rely on scrolling back and forth between several frames to accurately find these points. In general, intra-analyser error is high on this task since boundaries are often blurred and lost in the noise (see the upper LVPW boundary in the worst image in Fig. 2 for example). The network's ability to approach intra-analyser error using only a single frame indicates that it is accurately detecting the important structures despite the high noise level. Full results on the final test set are summarized in Table 1. The proposed network compares favorably to the other networks implemented on this task, achieving lower error on most metrics. We hypothesize that the performance of the other deeper networks would improve if the training dataset size were increased. However, our network has fewer parameters (which translates to a smaller memory size) and faster inference time. It is encouraging that close to expert level performance was achieved with a small network since efficient and fast implementations are important for clinical implementations.

5 Conclusion

In this work we present an effective landmark detection network for 2D measurements of the LV. We demonstrate the application of these techniques in determining LV dimensions. Implementation of this network could reduce high clinical inter-/intra-analyser variability in these measurements and lead to a more repeatable diagnostic pipeline. Additionally, it enables rapid historical analysis of patients to provide robust long-term analysis. We expect that many

Model	Mean Percent Error (%)				Params	Time (ms)
	Total	IVS	LVID	LVPW		
ResNet18	12.8	12.7	11.7	14.2	1e7	21
ResNet34	13.0	**11.2**	12.1	15.8	2e7	38
ResNet50	11.6	13.7	8.8	12.3	2e7	43
Stacked Hourglass	11.3	12.1	7.4	14.4	3e7	79
U-Net	13.5	14.0	8.3	18.1	3e7	**10**
Modified U-Net	**10.0**	13.4	**6.0**	**10.8**	**7e6**	11
Intra-analyser	8.9	8.0	5.2	13.8	n/a	-

Table 1. Comparison of proposed network to implementations of state-of-the-art networks in landmark detection and intra-analyser results. Inference time is for a single image.

of the techniques presented here would be applicable to other landmark detection problems in 2D and 3D ultrasound. In the future we will increase the size of the datasets, apply cross-validation, automate the detection of ED and ES frames from a full cardiac cycle, and add a confidence metric for detecting outlier results to provide a fully automated measurement tool for clinical use.

References

1. Baracho, S., Pinheiro, D., De Melo, V., Coelho, R.: A hybrid neural system for the automatic segmentation of the interventricular septum in echocardiographic images. Proc. Int. Jt. Conf. Neural Networks 2016-Octob, 5072–5078 (2016)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: CVPR. pp. 770–778 (2016)
3. Kou, S., et al.: Echocardiographic reference ranges for normal cardiac chamber size: Results from the NORRE study. Eur. Heart J. Cardiovasc. Imaging 15(6) (2014)
4. Lang, R.M., Badano, L.P., 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 (2015)
5. Liu, R., Lehman, J., et al.: An Intriguing Failing of Convolutional Neural Networks and the CoordConv Solution (2018)
6. Newell, A., Yang, K., Deng, J.: Stacked Hourglass Networks for Human Pose Estimation. ECCV 9908, 483–499 (2016)
7. Nibali, A., He, Z., Morgan, S., Prendergast, L.: Numerical coordinate regression with convolutional neural networks. CoRR abs/1801.07372 (2018)
8. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: MICCAI (2015)
9. Snare, S.R., Mjølstad, O.C., , et al.: Automated septum thickness measurement-A Kalman filter approach. Comput. Methods Programs Biomed. 108(2), 477–486 (2012)
10. Sofka, M., Milletari, F., Jia, J., Rothberg, A.: Fully convolutional regression network for accurate detection of measurement points. In: DLMIA (2017)
11. Thorstensen, A., Dalen, H., Amundsen, B.H., Aase, S.A., Stoylen, A.: Reproducibility in echocardiographic assessment of the left ventricular global and regional function, the HUNT study. Eur. J. Echocardiogr. 11(2), 149–156 (2010)
12. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using Convolutional Networks. In: CVPR. vol. 07-12-June, pp. 648–656 (2015)

Paper III

Septal Curvature as a Robust and Reproducible Marker for Basal Septal Hypertrophy

Maciej Maciniak, Andrew Gilbert, Filip Loncaric, Joao Filipe Fernandes, Bart Bijnes, Marta Sitges, Andrew King, Fatima Crispi, Pablo Lamata

Published in *Journal of Hypertension*, 2021, volume 38, DOI: 10.1097/HJH.0000000000002813.

III

CE: Tripti; JH-D-20-01452; Total nos of Pages: 8;

JH-D-20-01452

Original Article

Septal curvature as a robust and reproducible marker for basal septal hypertrophy

Maciej Marciniak[a], Andrew Gilbert[b], Filip Loncaric[c], Joao Filipe Fernandes[a], Bart Bijnens[c,d], Marta Sitges[e], Andrew King[a], Fatima Crispi[c,e,f], and Pablo Lamata[a]

Background: Basal septal hypertrophy (BSH) is an asymmetric, localized thickening of the upper interventricular septum and constitutes a marker of an early remodelling in patients with hypertension. This morphological trait has been extensively researched because of its prevalence in hypertension, yet its clinical and prognostic value for individual patients remains undetermined. One of the reasons is the lack of a reliable and reproducible metric to quantify the presence and the extent of BSH. This article proposes the use of the curvature of the left ventricular endocardium as a robust feature for BSH characterization, and as an objective criterion to quantify current subjective 'visual assessment' of the presence of sigmoidal septum. The proposed marker, called average septal curvature, is defined as the inverse of the radius adjacent to each point of the endocardial contour along the basal and mid inferoseptal segments of the left ventricle.

Method: Robustness and reproducibility were assessed on a cohort of 220 patients, including 161 hypertensive patients (32 with BSH) and 59 healthy controls.

Results: The results show that compared with the conventionally used wall thickness metrics, the new marker is more reproducible (relative standard deviation of errors of 7 vs. 13%, and 8 vs. 38% for intra-observer and inter-observer variability, respectively) and better correlates to the functional parameters related to BSH, with main difference (absolute rank correlation 0.417 vs. 0.341) in local deformation changes assessed by longitudinal strain.

Conclusion: Average septal curvature is a more precisely defined and reproducible metric than thickness ratios, it can be fully automated, and better infers the functional remodelling related to hypertension.

Keywords: basic research, curvature, diastolic function, hypertension, hypertrophy, systolic function

Abbreviations: 2D, two-dimensional; ASC, average septal curvature; BSH, basal septal hypertrophy; IVS, interventricular septum; WTR, wall thickness ratio

INTRODUCTION

The increase in cardiac wall thickness in hypertensive heart disease is gradual and not uniform [1]. In early stages of hypertension, some patients demonstrate localized hypertrophy of the basal septum. The significance of this finding is not definitive; however, some studies indicate that it serves as a marker of a more advanced impact of afterload on cardiac function in patients with hypertension [2]. If a patient presents with localized thickening, there is reasonable concern that they have hypertension influencing the cardiac function and might benefit from more stringent blood pressure control and a detailed follow-up.

The link between this localized thickening and elevated blood pressure was demonstrated when volunteers with basal septal hypertrophy (BSH) and no known history of hypertension were diagnosed with masked hypertension using 24-h blood pressure monitoring [3]. In the general population, the Framingham Heart study found the prevalence of BSH to be at 1.5%, reaching up to 17.8% in older patients [4]. A number of other studies investigated the prevalence of BSH in hypertensive cohorts, finding it to be around 20% [5,6].

This morphological trait is thus an early sign of structural and functional remodelling [7], and has been extensively researched, in spite of a lack of consensus on nomenclature, with the finding being termed septal bulge [1,8,9], isolated interventricular septum thickening [5,10], discreet upper septal thickening [4,11,12], and basal septal hypertrophy [2,13,14], among others. Beyond the nomenclature, different criteria exist to define BSH, the majority of which are based on the myocardial thickness in the basal interventricular septum (IVS), and its relation to the thickness at the mid-point of the septum or the deviation from a normal value. For example, the basal septal wall thickness [10–13], basal septal and mid-septal wall thickness ratios (WTR) [1,5], various ratios of the septal thickness and other echocardiography parameters [9,12], or a combination of the

Journal of Hypertension 2021, 38:000–000

[a]School of Biomedical Engineering and Imaging Sciences, Kings College London, London, UK, [b]Cardiovascular Ultrasound, GE Vingmed, Oslo, Norway, [c]Institute of Biomedical Research August Pi Sunyer (IDIBAPS), Barcelona, [d]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, [e]Hospital Clinic de Barcelona and [f]Barcelona Center for Maternal Fetal and Neonatal Medicine, Hospital Sant Joan de Déu, Barcelona, Spain

Correspondence to Maciej Marciniak, MSc. Eng., School of Biomedical Engineering and Imaging Sciences, King's College London, 1 Lambeth Palace Rd, South Bank, London SE1 7EU, UK. Tel: +44 7543200961; e-mail: maciej.marciniak@kcl.ac.uk

Received 17 December 2020 **Revised** 11 January 2021 **Accepted** 12 January 2021

DOI:10.1097/HJH.0000000000002813

above-mentioned [4,15] have all been used to define BSH. The reproducibility of these criteria is often limited as arbitrary cut-off values are employed by different groups (given the lack of gold standard), and the fact that the measurement of the thickness often relies on the M-mode. Moreover, the discrepancies in the metrics may occur for several reasons: inconsistency in finding the epicardial border, IVS measurements not exactly perpendicular to the septal wall and/or different opinions on the positioning of the basal and mid segments among the experts.

Apart from the quantitative wall thickness measurements, an equally important diagnostic criterion is the visual assessment of the septal geometry as the shape of the BSH is also distinctive. This diagnostic marker was termed a sigmoidal septum or an upper septal 'knuckle' [2,4,8,15,16]. The visual determination of the BSH is based on its location in the basal segment of the IVS, the fact that it is asymmetric and the visible myocardial thickening is more abrupt and convex than in the cases of concentric LV hypertrophy. Given the lack of clear guidelines, this assessment is subjective, and suffers from low reproducibility.

The problem addressed in this article is this lack of a clear criterion and reproducibility in the assessment of BSH. Together with the variability in the characteristics of the examined cohorts, this problem causes current difficulties in determining the prevalence and the associated risks of BSH. The solution proposed to distinguish between patients with and without BSH in a robust and reproducible manner is a combination of a semiautomatic segmentation of the left ventricle (LV) and the computation of curvature along the LV endocardium of the septal wall. The rationale is that curvature is a metric that objectively quantifies the subjective visual assessment of the presence of the sigmoidal septum, and thus removes human variability. The motivation is also based on the hypothesis that the curvature of the endocardial contour is more reproducible than metrics relying on highly variable thickness measurements.

A retrospective study of hypertensive patients is presented here in order to evaluate the performance of the septal curvature compared with the traditional methods. We analyse the intra-observer and inter-observer variability of the metrics and validate the hypothesis of the improved reproducibility of curvature. We then investigate the relationship of the anatomical BSH metrics with early markers of functional remodelling associated with hypertension, finding that curvature is a better descriptor of the deformation changes. We finally explore the relationship between thickness and curvature, both by their correlation and by the diagnostic agreement based on defined thresholds, and discuss the clinical relevance of our findings.

METHODS

Hypertensive patients ($N = 161$) with well-controlled blood pressure, treated with antihypertensive drugs for a minimum of 3 years, were included. Patients were recruited from the outpatient clinic and general practitioner referrals. Exclusion criteria included history of heart failure or previously known target organ damage. Volunteers from the local community ($N = 59$) who were presumed healthy, without prior history of hypertension, diabetes or other significant cardiac or noncardiac diseases represented healthy controls. Full description of the cohorts is included in the supplementary material, http://links.lww.com/HJH/B583.

Examinations were performed by qualified sonographers according to the current recommendations of the European Association of Cardiovascular Imaging [17] on a commercially available GE Vivid 9 system equipped with a M5S transthoracic transducer. Machine settings, including gain, time gain compensation, and compression, were adjusted for optimal visualization. Images were analysed using GE EchoPAC software (v.202.41.0, GE Vingmed, Horten, Norway).

BSH is defined based on the basal-to-mid septal thickness ratio of at least 1.4 in either the four-chamber (4CH) or parasternal long-axis (PLAX) view [7]. In order to classify the cohort, thicknesses in the anteroseptum and inferoseptum were manually measured during end-diastole at basal-level and mid-level in PLAX and 4CH views, respectively. The measurements were obtained perpendicular to the LV long axis, at the interface between the myocardial wall and cavity to the transition of the LV to the RV septal myocardium [17]. On the basis of these criteria, 32 (20%) patients were diagnosed with BSH.

Curvature is the amount by which a line deviates from being straight; therefore, it is expected to be highest around the apex, and low at the mid and basal wall segments. Along a contour, three consecutive points in space define a circumference, and the smaller the radius of this circumference, the larger the curvature. With these intuitive ideas, the LV endocardial curvature is defined as the inverse of the radius of the circle that can be defined by triplets of adjacent points of the contour delineated on the endocardium. We consider it being concave if the adjacent circle has its centre outside of the left ventricular cavity and convex otherwise (see Fig. 1a). The curvature in the LV septum provides information about the convexity of the hypertrophy directly quantifying the visual assessment performed by clinicians (Fig. 1b).

In practice, curvature of a 2D curve is calculated with its first and second derivatives and is quantified with inverse length units (dm^{-1}) [18]. Let $c(t) = (x(t), y(t))$ be a representation of a curve with existing, continuous first and second derivatives, defined in 2D space. For each point t of the trace defined on the 2D plane, the curvature $\boldsymbol{\kappa}$ is calculated as (Eq. 1):

$$\boldsymbol{\kappa}(\boldsymbol{t}) = \frac{\chi(t)''y(t)' - \chi(t)'y(t)''}{\left(x(t)'^2 + y(t)'^2\right)^{3/2}}$$

where $x(t)'$, $x(t)''$ are the first and second derivatives in point t along one of the spatial directions.

Although the curvature value must always be positive [i.e. the absolute value of (Eq. 1)], the LV endocardial contour may contain both convex and concave curvatures, and to distinguish between the two, we consider the curvature convex if the calculated value is positive and concave otherwise.

The curvature indexes are calculated for all contour points using a semi-automatic pipeline (see Fig. 2). In a

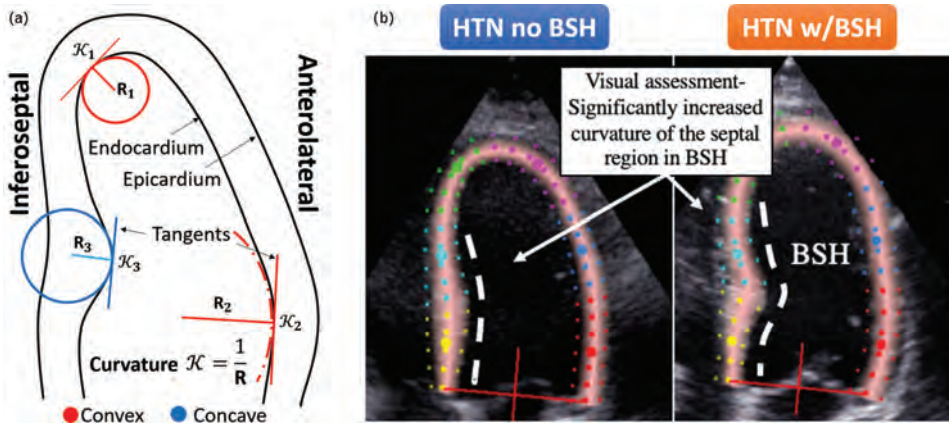


FIGURE 1 (a) An intuitive explanation of convex and concave curvature of the left ventricular endocardium, as seen in the four-chamber echocardiography view. The curvature is a reciprocal of the radius of circle adjacent to the endocardium. (b) The comparison of the septal wall in hypertensive patients with and without the basal septal hypertrophy (BSH). The visual assessment differentiating the two can be quantitatively compared with the curvature metric.

4CH view, BSH is located within the basal and sometimes mid inferoseptal segments. In order to create a useful metric, the calculated curvature indices from these two segments are aggregated into a single average value, which we named the average septal curvature (ASC). We hypothesize that the ASC of patients with BSH is far more concave in these segments than among hypertensive patients without BSH (conf. Fig. 1b).

In order to measure the curvature of the septal wall, the two-dimensional (2D) LV endocardial contour of the end-diastolic frame (used in the diagnosis of BSH [14]) is required. In an attempt to maximize clinical translation, we use a widely available semi-automatic segmentation tool (EchoPAC, GE Ultrasound), designed to measure the longitudinal 2D strain and re-purposed in this work to extract LV endocardial contours through the entire cardiac cycle. A similar pipeline could be applied to other modalities, such as cardiac MRI or computed tomography.

Using GE EchoPAC strain module, the operator sets a number of points on the endocardium of the LV in the end-systolic phase. Then, the algorithm interpolates the given set, to create additional points, which resemble a smooth curve. The operator then adjusts the segmentation shape by re-positioning the original points. In the next step, the contours are propagated thorough the entire cardiac cycle with the speckle-tracking algorithm.

The contours are exported from the module for post-processing as further interpolation is required to avoid

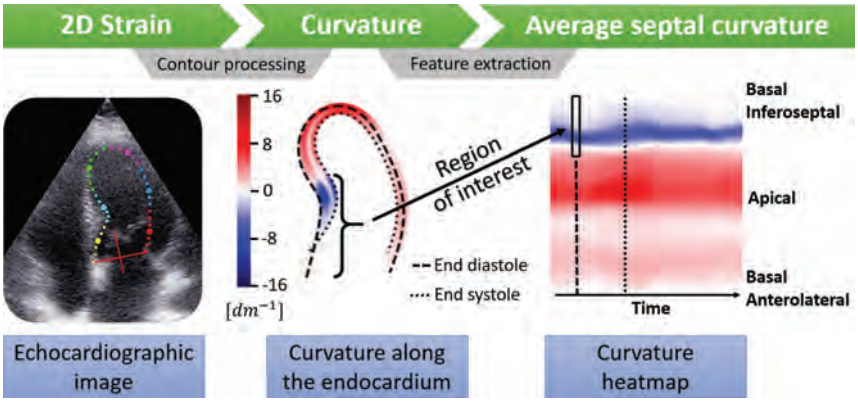


FIGURE 2 A proposed pipeline to go from ultrasound images to curvature analysis. The images from a full cardiac cycle are semi-automatically segmented with a 2D strain tool. Then, the contours of the endocardium are extracted, smoothed and interpolated to create a continuous contour from which the curvature is calculated. Finally, the curvature indexes within the region of interest (30% of the contour) in end-diastole are aggregated to compute the average septal curvature.

underestimating the curvature in critical points. A radial basis function interpolator is chosen, with the smoothness parameter set to the original number of points in the trace as this algorithm provides smooth first and second spatial derivative of the interpolated contour (required in Eq. 1). As a result, each contour is up-sampled to 500 points, a resolution that was empirically tested to lead to robust and smooth curvature values (details on the implementation can be found in [19]). The end-diastolic frame is found based on the maximum area enclosed by the contour. The curvature is calculated along the entire contour in that frame. Finally, the 30% of the curvature indices from the top of the septal base through basal and mid segments are aggregated to compute the ASC.

To test the variability of measurements, an inter-observer/intra-observer analysis was conducted on a group of 20 patients: 10 healthy controls and 10 hypertensive patients, including two cases with BSH. The measurements were performed by clinical experts. The metrics were compared on the basis of the average absolute difference between the measurements relative to the range of measurements, to account for different metric units.

First, to evaluate intra-observer variability, the original observer (O1) remeasured the basal and mid-wall IVS dimensions in end-diastole and performed segmentation on all 20 patients at least 2 months after the original measurement (O1∗). Second, to evaluate inter-observer variability, a different observer (O2) measured the same patients using the same 4CH and PLAX images that O1 used. Finally, another observer (O3) measured the same patients using a different 4CH image from the one used by O1 and O2. For most patients, only a single high-quality PLAX image was available, so the IVS was measured in the same PLAX image by O3 as O2 and O1. Similarly, the semi-automatic 4CH segmentation was performed by two other observers on the same (O2) and different (O3) images. This last study gives a measure of inter-observer inter-image variability on 4CH images, which is an important measure of robustness as an ideal method should give consistent results across any image from a patient given reasonable image quality.

To capture the deterioration caused by hypertension, functional metrics proven to be affected by BSH [7] were studied, including local longitudinal LV septal strain, longitudinal conduit and contractile left atrial strain (as well as ratio of the two), diastolic function marker (E/A), and mitral annulus septal and lateral velocities. Moreover, the metrics were studied against the anatomical markers of BSH, namely LV mass, and end-diastolic and end-systolic volumes indexed to BSA. The ability of both metrics, thickness and curvature, to infer the functional impairment is assessed by the regression between the variables within the hypertensive group ($N = 161$). The agreement between WTR (measured in two views) and ASC to identify BSH among the hypertensive population is studied by their correlation.

The processing pipeline, the statistical analysis and the variability analysis are implemented using the Python programming language, v.3.6.5. To test for normality of the distribution, the Pearson-D'Agostino test is used [20]. Two sample t-test is employed to test hypotheses on normally distributed samples, otherwise the Mann–Whitney U test is used. To calculate the correlation, Spearman's ρ and Pearson's Rs are computed.

RESULTS

As ASC is a novel metric, the distribution analysis is provided as a reference. The distributions of the three metrics under study, the ASC and WTR in 4CH and PLAX, in both the 161 hypertensive and 59 normotensive patients are depicted in Fig. 3. Moreover, the relevant percentiles and distribution information is shown in Table 1. The hypertensive population displays with a sharp transition around the ratio of 1.4 in a WTR in 4CH view, in concordance with the threshold for diagnosis of BSH. This transition is less clear in the WTR of the PLAX view, and no separation can be seen in the ASC distribution. None of the hypertensive distributions are normal (conf. Table 1). ASC is negatively skewed, with median value equal to concave curvature of $-0.362\ \mathrm{dm}^{-1}$, and 75% of the cases have ASC below $-0.88\,\mathrm{dm}^{-1}$. Conversely, WTR distributions are positively skewed, and the 85th percentile denotes the threshold for the diagnosis of BSH with respect to the WTR in either view.

In contrast to the hypertensive population, the distribution of the curvature among the healthy controls was mainly convex (median $= 0.44\,\mathrm{dm}^{-1}$, 78% cases with convex curvature index). This agrees with the understanding that the concave ('sigmoid') septum is a sign of a specific pattern of remodelling [14,21]. Representative examples with the ASC values are provided in Fig. 3.

The intra-observer and inter-observer differences relative to the range of measurements were significantly lower for the curvature index when compared with WTR (conf. Table 2). In fact, the two cases originally identified as HTN w/BSH were mismatched with the control cases when compared among observers using WTR. There was no systematic bias within and between observers in ASC and $\mathrm{WTR}_{\mathrm{PLAX}}$ measurements. However, in the metric obtained from 4CH view, there was a significant inter-observer discrepancy, where WTR measurements were on average 0.28 and 0.2 higher than those obtained by O2 and O3, respectively. The measurements and Bland–Altman plots are provided in the supplementary material (Figure S1, Tables S1–S3, http://links.lww.com/HJH/B583).

Assuming current clinical standard the ground truth to identify the presence of BSH in hypertensive patients, we studied how each of the three individual metrics agreed with it. Given that WTR was used as the traditional metric to diagnose BSH, it was not surprising to find a better separation between groups with the WTR indexes than with ASC (left panel in Fig. 4). Nevertheless, the difference between the two groups was significant in all metrics ($P < 0.001$ in all cases).

A more detailed inspection of the metrics in a scatter plot (see right panel in Fig. 4) revealed the disagreements of the WTR acquired in different views. These discrepancies were more pronounced in the higher end of the WTR values, above the threshold value. The correlation between the two measurements was moderate ($R = 0.36$, $P < 0.001$). ASC was found to also be moderately correlated with the WTR in 4CH view ($R = -0.4$, $P < 0.001$) and weakly correlated to

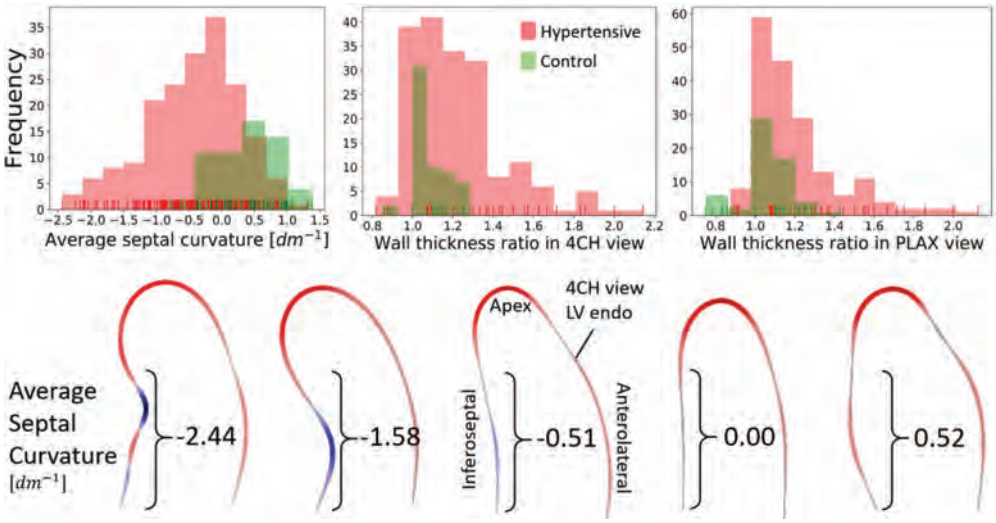


FIGURE 3 Upper: histograms of values of average septal curvature and wall thickness ratios measured in parasternal long-axis and four-chamber views among patients with and without hypertension. The negative values of the curvature represent concave curvature. The distribution of the curvature index resembles a normal distribution, whereas the wall thickness ratios are strongly skewed. Lower: 4CH view endocardial contours and their corresponding ASC values. The metric penalizes the sharp changes in the septal profile, regardless of its position. The two leftmost cases were diagnosed with BSH. 4CH, four-chamber; ASC, average septal curvature; BSH, basal septal hypertrophy.

TABLE 1. The description of the distributions of average septal curvature and wall thickness ratio indexes

Distribution description	Average septal curvature (dm^{-1})	Wall thickness ratio	
		4CH	PLAX
Minimum	−2.440	0.818	0.875
P_{15}	−1.163	1	1
P_{25}	−0.88	1.091	1
P_{50}	−0.362	1.167	1.125
P_{75}	0.041	1.333	1.286
P_{85}	0.335	1.429	1.413
Maximum	1.310	2.143	2.125
P_{normal}	0.035	<0.001	<0.001

P_i values denote the percentiles. P_{85} of WTR in either of the views is the threshold for the diagnosis of BSH. P_{normal} is the P value of the test for normality. None of the distributions can be qualified as normal. 4CH, four-chamber; PLAX, parasternal long-axis.

the WTR in PLAX view ($R = -0.21$, $P = 0.004$). Stronger correlation between WTR_{4CH} and ASC measurements was expected as the curvature indexes were calculated from trace delineated in the 4CH view.

Although the new method does not classify the cases in the same manner, the relation between the metrics and cardiac deformation revealed that ASC was more correlated with the functional remodelling than with the anatomical markers: Table 3 shows that whenever the markers are significantly correlated with a functional parameter, the correlation with ASC is the strongest, except for medial a'. ASC is stronger correlated with all the strains, both ventricular and atrial, and the E/A ratio, with the most apparent difference in the basal longitudinal strain (shown in Fig. 5, $\rho_{ASC} = -0.417$ vs. $\rho_{4CH} = 0.341$ and $\rho_{PLAX} = 0.24$), which is also the most pronounced marker of BSH. In case of the mid-septal strain and left atrial conduit strain, ASC is

TABLE 2. Results of intra-observer and inter-observer variability

Observer	Metric	Range	Absolute values			Relative values		
			Max AD	AAD	SDAD	Max AD	AAD	SDAD
O1 and O1*	ASC	2.78	0.64	0.17	0.18	**23%**	**6%**	**7%**
	WTR_{4CH}	0.88	0.52	0.11	0.12	59%	12%	13%
	WTR_{PLAX}	0.94	0.45	0.16	0.12	47%	17%	13%
O1 and O2	ASC	2.85	0.67	0.23	0.20	**23%**	**8%**	**7%**
	WTR_{4CH}	0.86	0.77	0.28	0.22	89%	33%	25%
	WTR_{PLAX}	0.89	0.56	0.26	0.14	63%	30%	16%
O1 and O3	ASC	2.39	0.74	0.32	0.22	**31%**	**13%**	**9%**
	WTR_{4CH}	0.51	0.98	0.26	0.26	190%	51%	51%
	WTR_{PLAX}	0.67	0.67	0.18	0.17	100%	26%	26%

Differences in absolute and relative (to the range of the index) terms. Bold indicates most relevant values. 4CH, four-chamber; AD, absolute difference; ASC, average septal curvature; AAD, average AD; BSH, basal septal hypertrophy; SDAD, standard deviation of AD.

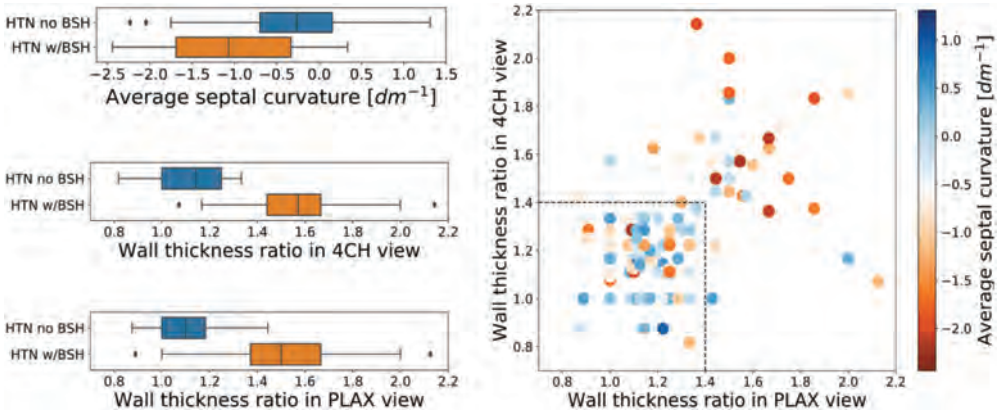


FIGURE 4 Left: box plots of values of average septal curvature and wall thickness ratios measured in parasternal long-axis and four-chamber views among patients diagnosed and not diagnosed with BSH. Right: WTR values measured in two ultrasound views: PLAX and 4CH. The colours signify the values of ASC. The discrepancies between two metrics increase with the calculated values. The majority of the cases with WTRs below the BSH threshold also hold a low ASC value. 4CH, four-chamber; ASC, average septal curvature; BSH, basal septal hypertrophy.

TABLE 3. Rank-correlation between average septal curvature and wall thickness ratios markers and anatomical and functional parameters

	Spearman ρ		
Variables	**ASC**	**WTR_{4CH}**	**WTR_{PLAX}**
LV mass index	0.018	0.162*	0.173*
LV end-diastolic volume index	0.062	0.079	0.157*
LV end-systolic volume index	0.113	0.060	0.122
2D left atrial conduit volume	0.051	0.057	0.100
E/A ratio	0.187*	−0.163*	−0.131
Mitral annulus *e'* medial velocity	0.234*	−0.156*	−0.158*
Mitral annulus *a'* medial velocity	−0.192*	0.227*	0.163*
Mitral annulus *e'* lateral velocity	0.062	−0.116	−0.085
Average mitral annulus *e'* velocity	0.129	0.148	0.137
Basal septal strain	−0.417*	0.341*	0.24*
Mid septal strain	−0.164*	0.100	0.109
LA contractile strain	−0.219*	0.158*	0.152
LA conduit strain	0.159*	−0.137	−0.060
LA conduit/contractile strain ratio	−0.232*	0.203*	0.140

ASC, average septal curvature; BSH, basal septal hypertrophy; LA, left atrium; LV, left ventricle; WTR, wall thickness ratio.
*P value less than 0.05 for correlation.

the only significantly correlated metric. The function of mitral annulus in lateral location is not significantly correlated with any marker.

DISCUSSION

Average curvature of the septal segment is proposed as an anatomical feature to identify the presence of BSH. It is a more precisely defined and reproducible metric than thickness ratios, and better infers the functional remodelling of the basal septal segment.

The problem addressed in this work is the high variability between multiple WTR measurements, as described in the literature [22,23] and confirmed in our inter-observer and intra-observer analysis that indicated that 2D measurements of the septum thicknesses are not consistent (13 and 38% for intra-observer and inter-observer variability, see Table 2). This large variability led to discrepancies in diagnosis among the observers, making multiple control cases considered as having BSH and vice versa. The proposed solution is a quantitative definition of the so far qualitative visual assessment of the presence of the sigmoid

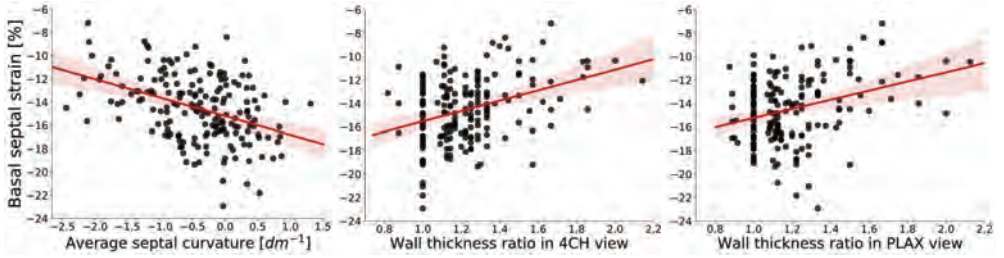


FIGURE 5 Linear regression between functional (basal septal strain index) and anatomical markers of basal septal hypertrophy: average septal curvature and wall thickness ratio indexes.

septum currently done in clinical practice. ASC indices are more user-independent (7 and 8% for intra-observer and inter-observer variability, Table 2) and show better diagnosis agreement across observers.

ASC is more reproducible than WTR for two reasons. First, as it has a precise definition, on the contrary to WTR where the locations of the minimum and maximum thickness measurements are ambiguous. ASC is computed along the basal and mid inferoseptal segments of the left ventricle in a 4CH view, objectively derived as 30% of the LV endocardial contour. The second reason is that ASC only requires segmentation on the bright LV endocardial border: curvature defined on the LV endocardial border removes the source of variability in the delineation of the epicardium, which is usually more challenging and required in WTR.

There are still two potential sources of variability in ASC. First, there could be inconsistency in placing the first basal points of the endocardial contour, but its impact should be small because of its rigorous description in guidelines [17]. Second, the variability during image acquisition in positioning the basal segment in face of the outflow tract is an unsolved problem [24]. As an example, a few outliers exhibiting strong concave curvature among the non-BSH patients (conf. Fig. 4) were caused by low acquisition quality, where the 4CH view image almost entirely captured the left ventricular outflow tract and was thus closer to an apical long-axis view. The segmentation in the presence of the outflow track (e.g. on the apical long-axis view) lacks the consensus on the reproducible border between the ventricle and outflow, that would allow for the usage of curvature in this context. Despite these two potential sources of variability, ASC has shown to lead to a reproducible and easily interpretable metric, and could be further refined for other echocardiographic views, as well as other imaging modalities and 3D images. In addition, curvature could be defined in other echo views, such as the PLAX, provided that the guidelines were established.

Beyond reproducibility, and in the absence of outcome metrics and disease onset, this work searched for the BSH metric that correlated with those functional markers that indicate early signs of degradation. Changes in left atrial and ventricular function have been shown to be present in hypertensive patients with BSH – decreased regional LV systolic contraction was related with impaired LV relaxation, a higher level of indeterminate diastolic dysfunction and LA functional impairment [7]. Therefore, BSH has been linked to both morphological and functional cardiac remodelling, potentially increasing the patient's risk for atrial fibrillation and heart failure. In our study, ASC shows a better ability to infer function impairment, and with the additional strength of being independent from LV mass and volume indexes making it a complementary anatomical marker, as opposed to the existing WTR markers (conf. Table 3). In addition, ASC is a metric that changes in patients with hypertension and is strongly amplified in BSH patients (conf. Fig. 3), suggesting its potential value in tracking the adaptation of the heart to the hypertensive insult. Further research is needed to investigate the ASC utility in risk quantification and prediction of disease onset or clinical outcomes. In detail, a longitudinal study showcasing the speed and pattern of local remodelling and changing curvature would shed light on the BSH development. In addition, ASC in BSH cases should be compared with other diseases related to thickening of the septal segments, such as hypertrophic cardiomyopathy.

Previous studies have explained the mechanistic cause of BSH by the extra stress that this region holds compared with the rest of the left ventricle when pressure increases [2,3,5]. Given a certain pressure in a chamber, the stress that a wall holds depends on its curvature accordingly to Laplace's law: the flatter the surface, the larger the stress. The septum near the outflow track is the flattest anatomical part of the left ventricle, and thus becomes the early beacon of the presence of cardiac remodelling caused by hypertension. Our study of curvature provides additional evidence for this mechanistic explanation: the curvature in the septal region is on average flat (ASC = 0), or even convex (positive ASC), in many control patients (see Fig. 3). One interesting hypothesis is the existence of morphological features in the septum and outflow track that make the LV more prone to the occurrence of BSH: patients that have null or negative ASC values at the onset of the hypertension would be more sensitive to BSH in the progression of the disease.

The future methodological research will focus on the automation of the metric acquisition and unveiling the remodelling patterns, to classify the adaptive and maladaptive responses. It has recently been shown that deep neural networks can accurately segment the left ventricle, myocardium and the left atrium directly from ultrasound B-mode images [25]. Employing the networks could speed up generating the contours and create an objective metric independent of the observer, provided that the network could accurately delineate the septal profile. Furthermore, the statistical shape analysis has shown promise in aiding diagnosis and risk stratification [26,27]. This type of analysis could be directly applied to the segmented LV, to provide z scores describing the progress of the BSH in hypertensive patients. With such models and follow-up data, both maladaptive (increasing afterload) and adaptive (response to treatment) remodelling predictions could be made, to ensure appropriate treatment.

In conclusion, a novel shape biomarker, the LV endocardial septal curvature, was proposed to detect and quantify the presence of BSH. The distribution of this biomarker in control and hypertensive cohorts was described. It was found to be a more reproducible metric than thickness ratios and it better infers the functional deterioration related to hypertension. The tool to compute it [19], and the data to verify its implementation [28], are released to enable an easy adoption by the community.

ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under the Marie Skłodowska-Curie grant agreement No 764738. P.L. holds a Wellcome Trust Senior Research Fellowship (209450/Z/17/Z).

Conflicts of interest

There are no conflicts of interest.

REFERENCES

1. Baltabaeva A, Marciniak M, Bijnens B, Moggridge J, HE F, Antonios T, *et al*. Regional left ventricular deformation and geometry analysis provides insights in myocardial remodelling in mild to moderate hypertension. *J Echocardiogr* 2007; 9:501–508.
2. Pearson AC. The evolution of basal septal hypertrophy: from benign and age-related normal variant to potentially obstructive and symptomatic cardiomyopathy. *Echocardiography* 2017; 34:1062–1072.
3. Gaudron PD, Liu D, Scholz F, Hu K, Florescu C, Herrmann S, *et al*. The septal bulge - an early echocardiographic sign in hypertensive heart disease. *Am J Hypertens* 2016; 10:70–80.
4. Diaz T, Pencina MJ, Benjamin EJ, Aragam J, Fuller DL, Pencina KM, *et al*. Prevalence, clinical correlates, and prognosis of discrete upper septal thickening on echocardiography: the Framingham Heart Study. *Echocardiography* 2009; 26:247–253.
5. Verdecchia P, Porcellati C, Zampi I, Schillaci G, Gatteschi C, Battistelli M, *et al*. Asymmetric left ventricular remodeling due to isolated septal thickening in patients with systemic hypertension and normal left ventricular masses. *Am J Cardiol* 1994; 73:247–252.
6. Lewis JF, Maron BJ. Diversity of patterns of hypertrophy in patients with systemic hypertension and marked left ventricular wall thickening. *Am J Cardiol* 1990; 65:874–881.
7. Loncaric F, Nunno L, Mimbrero M, Marciniak M, Fernandes JF, Tirapu L, *et al*. Basal ventricular septal hypertrophy in systemic hypertension. *Am J Cardiol* 2020; 125:1339–1346.
8. Swinne CJ, Shapiro EP, Jamart J, Fleg JL. Age-associated changes in left ventricular outflow tract geometry in normal subjects. *Am J Cardiol* 1996; 78:1070–1073.
9. Canepa M, Pozios I, Vianello PF, Ameri P, Brunelli C, Ferrucci L, Abraham TP. Distinguishing ventricular septal bulge versus hypertrophic cardiomyopathy in the elderly. *Heart* 2016; 102:1087–1094.
10. Ranasinghe I, Ayoub C, Cheruvu C, Freedman SB, Yiannikas J. Isolated hypertrophy of the basal ventricular septum: characteristics of patients with and without outflow tract obstruction. *Int J Cardiol* 2014; 173:487–493.
11. Chen-Tournoux A, Fifer MA, Picard MH, Hung J. Use of tissue Doppler to distinguish discrete upper ventricular septal hypertrophy from obstructive hypertrophic cardiomyopathy. *Am J Cardiol* 2008; 101:1498–1503.
12. Nagaraj U, King M, Shah S, Ghosh S. Evaluation of discrete upper septal thickening on 64-slice coronary computed tomographic angiography. *J Thorac Imaging* 2012; 27:359–365.
13. Yalçin F, Yigit F, Erol T, Baltali M, Korkmaz ME, Müderrisoglu H. Effect of dobutamine stress on basal septal tissue dynamics in hypertensive patients with basal septal hypertrophy. *J Hum Hypertens* 2006; 20:628–630.
14. Kelshiker MA, Mayet J, Unsworth B, Okonko DO. Basal septal hypertrophy. *Curr Cardiol Rev* 2013; 9:325–330.
15. Shapiro LM, Howat AP, Crean PA, Westgate CJ. An echocardiographic study of localized subaortic hypertrophy. *Eur Heart J* 1986; 7:127–132.
16. Belenkie I, MacDonald RP, Smith ER. Localized septal hypertrophy: part of the spectrum of hypertrophic cardiomyopathy or an incidental echocardiographic finding? *Am Heart J* 1988; 115:385–390.
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 Imaging* 2015; 16:233–271.
18. Strang G, Herman E. *Calculus, vol. 3*. Ann Arbor: XanEdu Publishing Inc; 2016.
19. Marciniak M. github.com/MaciejPMarciniak/curvature [Internet]. London: King's College London; [updated 7 June 2020]. Available at: https://github.com/MaciejPMarciniak/curvature. [Accessed 15 January 2021]
20. D'Agostino R, Pearson ES. Tests for departure from normality. Empirical results for the distributions of b2 and $\sqrt{}$b1. *Biometrika* 1973; 60:613–622.
21. Ranasinghe I, Yeoh T, Yiannikas J. Negative ionotropic agents for the treatment of left ventricular outflow tract obstruction due to sigmoid septum and concentric left ventricular hypertrophy. *Heart Lung Circul* 2011; 20:579–586.
22. Schoenmaker NJ, van der Lee JH, Groothoff JW, van Iperen GG, Frohn-Mulder IM, Tanke RB, *et al*. Low agreement between cardiologists diagnosing left ventricular hypertrophy in children with end-stage renal disease. *BMC Nephrol* 2013; 14:170.
23. Pietro DA, Voelkel AG, Ray BJ, Parisi AF. Reproducibility of echocardiography: a study evaluating the variability of serial echocardiographic measurements. *Chest* 1981; 79:29–32.
24. Johnson C, Kuyt K, Oxborough D, Stout M. Practical tips and tricks in measuring strain, strain rate and twist for the left and right ventricles. *Echo Res Pract* 2019; 6:R87–R98.
25. Leclerc S, Smistad E, Pedrosa J, Ostvik A, Cervenansky F, Espinosa F, *et al*. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. *IEEE Trans Med Imaging* 2019; 38:2198–2210.
26. Warriner DR, Jackson T, Zacur E, Sammut E, Sheridan P, Hose DR, *et al*. An asymmetric wall-thickening pattern predicts response to cardiac resynchronization therapy. *JACC Cardiovasc Imaging* 2018; 11:1545–1546.
27. Marciniak M, Arevalo H, Tfelt-Hansen J, Ahtarovski KA, Jespersen T, Jabbari R, *et al*. A multiple kernel learning framework to investigate the relationship between ventricular fibrillation and first myocardial infarction. In: Pop M, Wright G, editors. *Functional imaging and mModelling of the heart. Lecture Notes in Computer Science*. Cham: Springer; 2017. pp. 161–171.
28. [dataset] Marciniak M, Lamata P, 2020, Average Septal Curvature in Basal Septal Hypertrophy, figshare, 10.6084/m9.figshare.12443399.

Paper IV

Generating Synthetic Labeled Data from Existing Anatomical Models: An Example with Echocardiography Segmentation

Andrew Gilbert, Maciej Marciniak, Cristobal Rodero, Pablo Lamata, Eigil Samset, Kristin McLeod

Published in *Transactions on Medical Imaging*, 2021, DOI: 10.1109/TMI.2021.3051806.

Generating Synthetic Labeled Data from Existing Anatomical Models: An Example with Echocardiography Segmentation

Andrew Gilbert, Maciej Marciniak, Cristobal Rodero, Pablo Lamata, Eigil Samset, and Kristin McLeod

***Abstract*— Deep learning can bring time savings and increased reproducibility to medical image analysis. However, acquiring training data is challenging due to the time-intensive nature of labeling and high inter-observer variability in annotations. Rather than labeling images, in this work we propose an alternative pipeline where images are generated from existing high-quality annotations using generative adversarial networks (GANs). Annotations are derived automatically from previously built anatomical models and are transformed into realistic synthetic ultrasound images with paired labels using a CycleGAN. We demonstrate the pipeline by generating synthetic 2D echocardiography images to compare with existing deep learning ultrasound segmentation datasets. A convolutional neural network is trained to segment the left ventricle and left atrium using only synthetic images. Networks trained with synthetic images were extensively tested on four different unseen datasets of real images with median Dice scores of 91, 90, 88, and 87 for left ventricle segmentation. These results match or are better than inter-observer results measured on real ultrasound datasets and are comparable to a network trained on a separate set of real images. Results demonstrate the images produced can effectively be used in place of real data for training. The proposed pipeline opens the door for automatic generation of training data for many tasks in medical imaging as the same process can be applied to other segmentation or landmark detection tasks in any modality. The source code and anatomical models are available to other researchers [1].**

***Index Terms*— Data Generation, Echocardiography, Generative Adversarial Networks, Segmentation, Synthesis**

I. Introduction

Submitted on August 1st, 2020. This project has received funding from the European Union's Horizon 2020 research and Innovation program under the Marie Sklodowska-Curie grant agreement No 764738. P. Lamata holds a Wellcome Trust Senior Research Fellowship (209450/Z/17/Z).

A. Gilbert, E. Samset, and K. McLeod are with GE Vingmed Ultrasound, GE Healthcare, Horten, NO. A. Gilbert and E. Samset are also with the Department of Informatics at the University of Oslo, Oslo, NO (email: andrew.gilbert@ge.com; eigil.samset@ge.com; kristin.mcleod@ge.com).

M. Marciniak, C. Rodero, and P. Lamata are with the Biomedical Engineering Department at King's College London, London, UK (email: maciej.marciniak@kcl.ac.uk, cristobal.rodero_gomez@kcl.ac.uk, pablo.lamata@kcl.ac.uk

[1] https://adgilbert.github.io/data-generation/

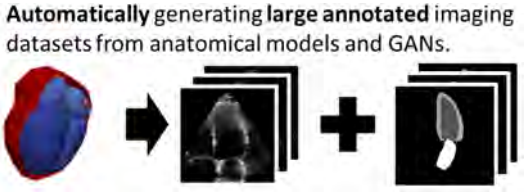


Fig. 1. Using anatomical models as high quality ground truth annotations, we propose a pipeline to generate large synthetic datasets for training convolutional neural networks.

MEDICAL imaging provides a window to capture the structure and function of internal anatomies. Imaging modalities such as ultrasound, computed tomography (CT) or magnetic resonance imaging (MRI) can be used to measure physical and physiological parameters. Accurate automation of these measurements would provide significant time-savings for clinical practitioners.

Convolutional neural networks (CNNs), have become the candidates of choice for measurement automation because they can accurately learn complex relevant features. However, CNNs require large sets of labeled data to learn and they are limited in accuracy by the quality of labels used in training. Inter-observer errors can be high in medical imaging tasks, especially when there is noise or other artifacts in the image. For example, in cardiovascular ultrasound (echocardiography or 'echo'), inter-observer errors for labeling common measurements can range from 4-22% even for experienced cardiologists [1], [2]. The variability in measurements is due to (a) the difficulty of accurately interpreting signals delineating structures amid image noise, and (b) differences in implementation between different acquisition machines and between practitioners in different institutions. A second problem when building datasets to automate tasks in medical imaging is labeling is time-consuming and expensive since quality annotations require experienced medical professionals. Finally, manual labels are inflexible and adapting them based on new insights requires a significant amount of time.

While CNNs have been at the forefront of automating diagnostic measurements, anatomical models are progressing the personalization of treatments. Simulations from "digital twins" (models with patient-specific parameters) are increas-

ingly being used to guide therapies and develop new treatments [3]. As with the revolution in statistical inferencing led by deep learning, larger computational resources have allowed the growth in complexity and realism of these anatomical models [4], [5]. While originally developed for personalized simulation of mechanics and biophysics, anatomical models are also a valuable source of high-quality shape information. We propose a method to solve the labeling challenges for medical deep learning by harnessing the information contained in anatomical models. Instead of labeling images, we let these models represent ground-truth anatomical shapes and generate task-specific paired realistic images as summarized in Fig. 1.

In particular, we demonstrate the usefulness of this pipeline for the task of segmenting parts of the left heart in echo images and thus make use of a set of cardiac models developed for electromechanical simulations of the heart. Similar anatomical models have been developed for a wide range of anatomies and most are free for academic use [4], [6]. The pipeline described here could readily be applied to those models as well with some application specific modifications. Section V-E provides more details on extensions to new anatomies.

A. Contributions

The proposed pipeline shifts the focus from annotating images to ensuring a CNN trained on synthetic images will generalize to real images. We test our pipeline by generating synthetic data for echo segmentation. Our main contributions are three-fold:

1) We present a pipeline to generate realistic synthetic images with paired labels using anatomical models and a CycleGAN [7]. The pipeline can generate datasets of arbitrary size and include labels from any region included in the original anatomical models.
2) We demonstrate the utility of the pipeline by building annotated synthetic 2D echo images from cardiac models. We show these synthetic images can be used for training deep learning algorithms, specifically by demonstrating accurate segmentation without any real labeled images. We present extensive validation of the proposed pipeline by testing on multiple datasets of real images from different clinical sites and annotators that were completely unseen during development.
3) We present an analysis of the sources of error in the segmentation including differences in image texture, tissue shape, and annotator style. We show that differences in the segmentations primarily come from differences in annotator bias, highlighting the need for standardized annotations.

B. Related Work

Because there are often only a few accurate anatomical models available, we first experiment with using shape analysis techniques to expand the available set of ground truth models. Shape analysis has previously been used in medical imaging for improving segmentations as well as for pathology detection and registration [8]–[11].

The proposed work translates labels from a source domain (slices from anatomical models) to a target domain (echo in the example application). Domain adaptation is a similar task, but uses labels from a different imaging modality instead of models. Recently, CycleGANs have facilitated domain adaptation with unpaired images by using two sets of generative and discriminative networks, one for each transformation direction [7]. Kazeminia et al. [12] and Taghanaki et al. [13] provide overviews of CycleGANs in medical imaging. So far CycleGANs have primarily been used for realistic cross-modality translation to CT or MRI images whereas this work focuses on echo. Generating echo images is challenging because of the complex noise patterns. These patterns change dramatically between images and even within a single image following the acquisition settings of the user and the stretching/squeezing of the scan-conversion process. Compared to echo, the well-defined boundaries in MRI or CT represent a more similar domain to the anatomical model images. The cone in echo images is also a consistent defining feature in the image which degrades the translational invariance of convolutional networks. CycleGANs have been applied in echo for segmentation with image quality improvement [14] and view conversion [15], but these works used two real datasets of echo images and thus did not have to address the above challenges of translating from a different modality to echo.

Others have developed alternative strategies for surmounting limited datasets in medical imaging and Tajbakhsh et al. provide an overview of different strategies for segmentation with unlabeled or limited data [16]. Specifically relevant to this work, several groups have proposed strategies using GANs to generate synthetic data. Eschweiler et al. proposed a CycleGAN strategy for synthesizing a microscopy cell image and location dataset [17]. However, their labels are randomly generated, which loses the key advantage of ground truth anatomical models and is not applicable to most other applications in medical imaging where anatomies cannot be randomly generated from scratch. Huo et al. proposed SynSeg-Net, a similar pipeline using unpaired labels from MRI to train networks on CT images using a CycleGAN [18]. While some of the methodologies are similar, the central difference is that our ground truth annotations come from 3D anatomical models rather than unpaired images from another modality. Because detailed 3D annotations are an intrinsic part of each anatomical model, *our pipeline is applicable to any segmentation or landmark detection task in any modality with no additional labeling required*. Our approach is focused on image synthesis rather than domain adaptation.

Previous works generating echocardiography images have primarily used physics simulators to exactly replicate speckle creation from a set of reflectors. In general, these approaches have focused on generating a few specific images rather than large datasets. For example, Alessandrini et al. demonstrated a full pipeline for generation of 3D echo video loops that were realistic enough to trick some human observers [19]. While useful for providing a ground truth of myocardial motion for strain estimation, this pipeline and similar approaches [20]–[22], are ill-suited for generating training data for deep learning algorithms because it does not scale well to larger

datasets. Each new generated image requires manual initialization and computationally heavy simulation. Other groups have used generative adversarial networks (GANs) for echo image synthesis. For example, Abdi et al. sampled new echo images from labels after conditioning a GAN on a paired dataset [23] an approach also demonstrated for skin lesions [24]. The key drawback is this approach can only be used to augment existing, already annotated datasets.

II. Methods

The proposed pipeline consists of two primary steps as shown in Fig. 2. First, pseudo images and paired labels are generated from 3D anatomical models as described in Sec. II-A. Second, the pseudo images are transformed into realistic synthetic ultrasound images using a CycleGAN and a set of real images as described in Sec. II-B. Afterwards, in Sec. II-C, we test the utility of the generated datasets by comparing segmentation networks trained with synthetic images to those trained with real images when testing on real images. The proposed pipeline is general to any medical imaging application, although models for the relevant anatomy are needed and application specific parameters are required in the extraction. Sec. III-A describes the models used for this application and Sec. III-B provides details on the example application (echo segmentation) and parameters.We demonstrate the method using several datasets as described in Sec. III-D.

A. Extraction

The input to the pipeline is a small set of anatomical models (”Original Models”) which contain labels for the attributes that will be segmented or detected.

1) New Shape Models: While anatomical models provide excellent ground truth, there are often few available which may not provide sufficient anatomical variability to build a heterogeneous dataset. We experiment with building additional anatomically realistic models using statistical shape analysis. The primary modes of variation are deconstructed from the original models using principle component analysis. New models are generated by randomly sampling from the first 9 modes of variation (capturing 90% of the total variation) within two standard deviations of the mean model. We repeat this procedure to generate 99 new models in total. Each of the synthetic models is still an anatomically plausible shape, but adds a heterogeneous example to our dataset. Full details of the construction are given in Appendix A.

2) Pseudo images and model labels: A CycleGAN learns to transform the appearance from one imaging set to another. To generate the input for the CycleGAN, a ”pseudo” database is generated where the anatomical shapes present in the pseudo database generally match the shape distribution found in an equivalent database of real images. Therefore the necessary functions here are application specific and are discussed in detail in Sec. III-B. The output of this step is both a pseudo image and a label image which contains ground truth for the learning step. Synthetic labels are generated from the original model to match the chosen task and selecting the task simply involves choosing the relevant regions in the model.

B. Transform

A CycleGAN [7] is trained to transform the pseudo images into synthetic ultrasound images using an unlabeled set of real ultrasound images. The default CycleGAN architecture and hyper-parameters are used except the generator network is replaced with a U-Net with 8 down-sampling levels [25] because it trains faster and gives equivalent results. The CycleGAN is trained for 200 epochs. Network weights are saved every 5 epochs. We select the best epoch by manually reviewing a sample result from each epoch (typically around epoch 180) but the exact epoch chosen did not have a significant impact on results in preliminary experiments. The selected network is used to save a synthetic image and paired label for each pseudo image.

C. Learn

A segmentation network is trained from the set of generated synthetic images and labels. The same U-Net architecture from the transform step is used. The network is trained for 30 epochs using cross-entropy loss. While the segmentation network can be included within the CycleGAN for end-to-end training [18], [26], we found the segmentation network was able to consistently achieve very good results on the synthetic images in preliminary results and did not find value in including this as a loss term within the transformation process. Additionally, splitting these two steps allowed us to develop an equal comparison between the synthetic and real data.

III. Example application: Echo Segmentation

The feasibility of the pipeline is proven by building synthetic datasets for 2D echo segmentation. This application was chosen to enable comparison against existing real datasets. Two task variants are tested. First, matching all overlapping constraints of the synthetic and real datasets presented in Sec. III-D, a network was trained to segment the left ventricle endocardial border (LV_{endo}) from apical four chamber images taken from the end diastole phase of the cardiac cycle. Second, the task was extended to include the left ventricle epicardial border (LV_{epi}) and left atrium (LA) border from both four chamber and apical two chamber views and both end diastole and end systole phases. Fig. 3 shows examples of apical four/two chamber images extracted from the anatomical model as well as examples of performing the relevant annotations in real ultrasound images.

A. Original Models

The original models for this application were a set of 19 3D heart models derived semi-automatically from CT images. CT images have high contrast and spatial resolution which enables accurate delineations of structure boundaries. These models were built for electromechanical simulations and contain a complete set of tissue labels. Each model contains labels for both ventricles, both atria, aorta, pulmonary artery and veins, and both venae cavae. Additional details on the model creation process are given in Appendix A and by Rodero et al. [27] (currently under review, the model construction matches that described in [28]).

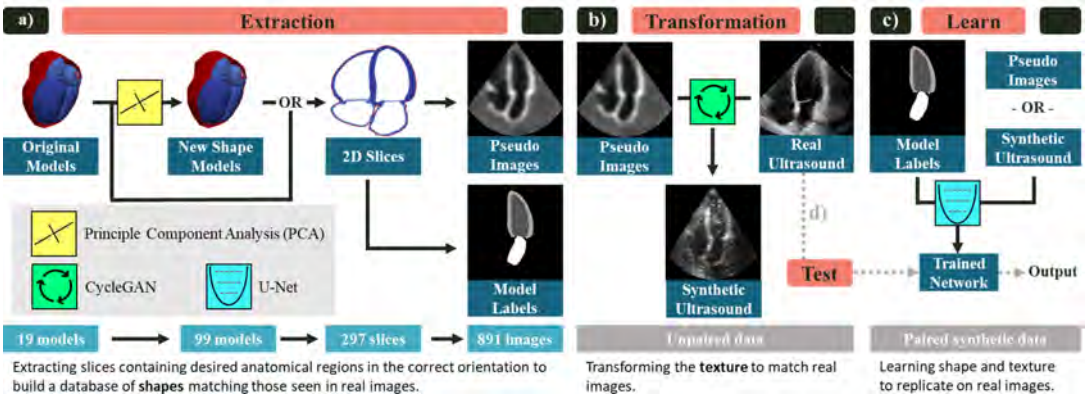


Fig. 2. **Overview of proposed pipeline implemented for echocardiography segmentation.** a) Extraction: pseudo images and ground truth labels are built from the 3D anatomical models. First, a larger cohort of shapes is generated by building a statistical shape model from the original anatomies and sampling new 3D instances using principle component analysis (PCA). Next, 2D slices of the desired view (apical four chamber shown) are sampled. Finally, pseudo-ultrasound images and the corresponding labels are built. Each step expands the size of the dataset. b) Transformation: The pseudo images and a dataset of unlabeled real echo images are used to train a CycleGAN to transform the pseudo images into synthetic ultrasound images. c) Learn: The paired synthetic ultrasound images and model labels are used to train a U-Net segmentation network. d) Test: The network trained on synthetic images is tested on real images to evaluate the utility of the pipeline. The creation of new shape models as well as the transformation module are optional extensions. The slicing can be performed on the original models and the segmentation network can be trained using pseudo images instead. We evaluate the effectiveness of these components in Sec. IV.

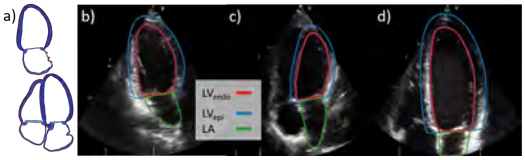


Fig. 3. **Example application: echo segmentation.** a) Apical two chamber (top) and apical four chamber (bottom) views as shown in an anatomical model. The right images show example real apical two chamber (b) and apical four chamber (c, d) echo images with task labels. (b, c) show the full heart while d) is zoomed to focus on the left ventricle.

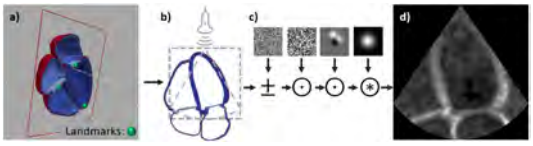


Fig. 4. **Extraction details for echocardiography.** To extract pseudo images a) a 2D plane is defined from a set of landmarks, b) the plane is rotated and cropped to match standard acquisition parameters and positioned to match standard positioning in real images, c) random noise and shadows are added and the slice is blurred yielding d) the final pseudo image. Additional details are given in Appendix A.

B. Task-Specific Data Generation

To generate a dataset for this task, 2D slices were sampled from the anatomical models and masked to mimic ultrasound images. A perfect 2D apical four chamber image is defined as the plane intersecting the apex, mitral valve center, and aortic valve center [29]. These landmarks were extracted from each model to define the optimal plane. Apical two chamber images were extracted by performing a 70 degree rotation counter-clockwise around the apical long axis from the four chamber landmarks (see Fig. 3). Although clinical guidelines suggest rotating the probe 60 degrees [29], using 70 degrees gave a better cut plane for the models from qualitative evaluation. To mimic natural variation in acquisition, random rotations of the cut planes around the long and short axes of the LV were sampled so that some slices are foreshortened or off-plane.

The 2D slices were transformed into pseudo images which mimic the appearance of ultrasound images. One of the most distinguishing features of an ultrasound image is the 'cone' marking the boundaries of imaging data. This is a consistent strong feature in all images and we found that the translational invariance of CNNs is degraded because the network could learn relationships between the cone boundary and structures. In other words, the CycleGAN discriminators could find difference between real and synthetic images from differences in structure location. In response, the generators would hallucinate structures in random locations. For the CycleGAN to properly transform structures as well as appearance, it is important that the distributions of locations of different anatomical structures are equally represented in both datasets.

To match this constraint, a series of affine transformations were applied to mimic the different LV orientations found in real images. This primarily consisted of masking the image with a cone and randomly cropping to either the entire heart ('whole heart' image) or the LV ('LV focused' image). The different crops are shown in Fig. 3 and match the image types suggested in clinical guidelines [29]. After cropping, other affine transforms such as rotations and squeezing were applied to ensure the region of interest remains inside the cone and add variance to the dataset (see Appendix A). Hard edges also decreased the realism of the generated images (see Appendix F) so random uniform noise and shadowing was added and

the images were blurred by convolving with a Gaussian kernel. This process is shown in Fig. 4. To introduce additional variety, the slicing and pseudo extraction processes were repeated 3 times each for a total of 891 images (99 × 3 × 3). The entire process is fully automated.

C. Segmentation Evaluation and Network Selection

Several metrics were used to evaluate the accuracy of the trained segmentation networks. First, the Dice score was measured where $\boldsymbol{D} = 200 * (S_{pred} \cap S_{ref})/(S_{pred} + S_{ref})$ and measures the overlap between a predicted segmentation S_{pred} and a reference segmentation S_{ref}. Second, following [30], we analyzed the convexity and simplicity of the output as criteria which identify successful annotations. Because we found these two metrics vary together, only simplicity is reported. Simplicity is defined as $\boldsymbol{S_p} = \sqrt{4 * \pi * Area(S_{pred})}/Perimeter(S_{pred})$ [31]. Note that simplicity relies only on the segmentation mask output from the network S_{pred} and not the label mask S_{ref}.

For the task of LV_{endo} segmentation, differences between annotators are often because of differing placements of the endocardial border within the myocardial tissue rather than differing ventricular shapes. According to guidelines [29], the LV_{endo} border falls at the interface between the non-compacted and compacted myocardium. If this border cannot be determined then the border falls at the blood-tissue interface. In noisy ultrasound images it can be difficult to accurately label this border, and there may be disagreement about which criteria should be used. There are no clear guidelines established for labeling LV_{epi} and LA borders for segmentation [30] which can lead to differences between annotators for those tasks as well.

To capture these potential disagreements, we calculated several additional metrics: mean distance between the contours and Bias. Mean distance (d_m) is the distance between two contours C_{ref} and C_{pred} averaged across their length. C_x indicates the border of S_x. Bias is the percentage error between the segmentation areas and is defined as:

$$\boldsymbol{B} = 200 * \frac{Area(S_{pred}) - Area(S_{ref})}{Area(S_{pred}) + Area(S_{ref})} \tag{1}$$

A high average Bias (positive or negative) across a dataset indicates a systematic difference in the labeling since the predicted results are consistently larger/smaller than the reference. Mean distance is calculated in pixels since we do not have access to image sizes in mm for all datasets. All other metrics are unit-less.

All networks were able to achieve high Dice scores on the synthetic data in preliminary experiments so selecting the network based on best Dice on a synthetic validation set lead to over-fitting to the synthetic data. Simplicity is a marker of the annotation quality that relies only on the network output and does not require a label. Therefore simplicity was tracked on an unlabeled set of real images (separate from the test set) through the course of training and the network with the highest simplicity was selected for final testing. This choice encouraged networks that generalized well to real data without requiring labels.

Median metrics were calculated in all cases since the distribution of scores was not normal. Therefore median absolute deviation was used as a measure of variance where $\boldsymbol{MAD} = Median(|X_i - \tilde{X}|)$. The Wilcoxon signed-rank test was used to calculate statistical significance between different results [32].

D. Real Datasets

Validating a dataset on a single source can lead to implicit bias in the developed methods [33]. For example, Degel et al. showed a decrease from 0.75 to 0.10 in Dice score for a CNN trained on one machine and tested on another for 3D left atrial segmentation. To account for this we validated the pipeline using a selection of real datasets. The characteristics of each dataset are described below and full details are listed in Appendix A.

1) Camus: The Camus dataset was introduced by Leclerc et al. [30]. It consists of apical four and two chamber images with segmentation labels for LV_{endo}, LV_{epi}, and LA at end diastole and end systole time points in the cardiac cycle. The images also include quality labels, and following the authors we limit our analysis to images of good or medium quality, leaving 1,600 images. The images are divided into training, validation and test splits of 80%, 10%, and 10% respectively, keeping images from the same patient in the same split.

2) EchoNet: The EchoNet dataset was introduced by Ouyang et al. [34]. It consists of 10,024 apical four chamber video loops with LV_{endo} segmentation labels for end diastole and end systole The images were divided into training, validation and test splits of 80%, 10%, and 10% respectively, keeping images from the same patient in the same split.

3) Additional real datasets: Since EchoNet contains only LV_{endo} annotations in apical four chamber images, additional real images were labeled with a full set of annotations, views and cardiac phases. Mixed apical four and two chamber videos from two different clinical sites were annotated by two experienced cardiologists (O_1 and O_2). Both cardiologists use echo as a part of their daily practice. To annotate the images they used the whole loop to check myocardial movement to find the correct structures and annotated LV_{endo}, LV_{epi}, and LA labels at end diastole and end systole ensuring that the labels between phases matched. The datasets were split by institution, $\textbf{Site}_{\textbf{A}}$ contains 336 images and was further divided into training and validation splits of 80% and 20% respectively. $\textbf{Site}_{\textbf{B}}$ contains 229 images and was left exclusively as a test set. Site_{A} was labeled by O_1 and Site_{B} was labeled by O_2.

4) Pathological dataset: The anatomical models were derived from asymptomatic patients and the aforementioned datasets contain no information on patient diagnosis. Therefore a set of pathological images was also gathered to test how well the networks trained on real and synthetic images would be able to adapt to pathological cases. 61 exams were gathered from patients diagnosed with severe functional mitral regurgitation, which is correlated with significant changes in LV shape [35], [36]. A severe diagnosis corresponds to a rating of 4 on a 4 point scale of severity. A random apical four chamber image was selected for each patient and O_2 labeled

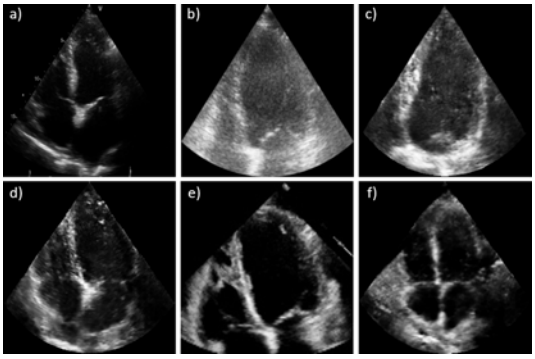


Fig. 5. **Synthetic images closely match real images.** Can you guess which images are synthetic? Answers below [3]

LV_{endo}, LV_{epi}, and LA areas at end diastole and end systole (yielding 122 images total) using the same criteria as above. All images were used exclusively for testing.

E. Synthetic Datasets

Synthetic versions of the Camus, EchoNet, and $Site_A$ datasets were generated using the pipeline in Fig. 2. No synthetic dataset was generated for the $Site_B$ or Pathological datasets since both were used for testing. Extraction and transformation were performed individually for each dataset and separately for each view. We predicted that using separate CycleGANs for each dataset and view would enhance the quality of the generated views and would allow the learned image features to be specific to the relevant dataset/view. Although customization of datasets and views could likely be combined into a single transformation process (using for instance an additional conditional input to the network), the focus of this work was on evaluating the feasibility of the pipeline rather than optimizing the generation process for multiple views and datasets. In most use cases all available datasets could be combined, however they were left separate here for evaluation purposes. Since $Site_A$ contained fewer images, the CycleGAN for that dataset was initialized from the final trained CycleGAN from the Camus dataset and it was trained for only 100 epochs. The models were built for only a single time step so the synthetic datasets contain only end diastole images.

To test the impact of the new shape models, a synthetic EchoNet dataset was created without using the additional models generated in the shape extension described in Sec. II-A.1. To maintain dataset size, the extraction part was modified to extract 5 2D slices per anatomical model and 9 pseudo images per slice (for a total of 855 images). This set is denoted with an * in the experiments in Sec. IV.

[3] a) Real $Site_A$ b) Real Camus c) Synthetic Camus d) Synthetic $Site_A$ e) Real EchoNet f) Synthetic EchoNet.

F. Inter-Observer Study

To analyze label variability, an inter-observer study was conducted for a subset of each dataset. 20 random images were selected from the test set (or validation if no test set was created) for the Camus, EchoNet, $Site_A$, Synthetic Camus, Synthetic EchoNet and Synthetic $Site_A$ datasets. To minimize the possible sources of variability, and match the overlapping constraints of the datasets, only apical four chamber end diastole images were selected. O_1 and O_2 annotated all images (except only O_1 labeled the EchoNet sets) with LV_{endo}, LV_{epi} and LA labels. The second round of labeling was conducted at least 2 months after the first round for $Site_A$.

G. Implementation Details

Hyperparameters for the segmentation such as the learning rate and loss function were tuned on the synthetic validation sets. All approaches were evaluated on the Camus validation set to ensure proper convergence and several different validation runs were run in the course of building the extraction and transformation steps. In general, the goal of this work was to evaluate the synthetic dataset construction using standard segmentation approaches rather than tuning an optimal segmentation network for the given application. The unlabeled EchoNet and $Site_A$ validation datasets were used only for network selection (see II-C) so the labels and metrics for these sets were never seen (and thus cannot influence design choices). This allows us to detect implicit bias in the design choices or training datasets. The test sets (Camus, EchoNet, and $Site_B$) were used only once during final testing for the results presented below. Additional details on implementation and hyperparameters can be found in the supplementary material.

IV. Results

The pipeline is evaluated first in Sec. IV-A by comparing expert cardiologist's annotations to those produced by the proposed pipeline. Next, since the aim of this pipeline is primarily to generate image/label pairs that are suitable for deep learning training, we check if a CNN can effectively learn from synthetic images in Sec. IV-B and compare to networks trained on real data. Finally, various versions of the synthetic dataset are analyzed in Sec. IV-C to determine which factors contributed to accurate segmentations.

A. Generated Images and Annotations

Images from the randomly selected inter-observer set are shown in Fig. 5 to demonstrate the realistic output of the generation pipeline. The synthetic images closely match their real counterparts in appearance. The GAN generates this appearance while maintaining the ground truth cardiac structures from the anatomical models. Generating a single ultrasound image from the prepared slice takes 81 ms.

Next, we checked if experts agreed with the pipeline-generated annotations. Metrics from the inter-observer study are shown at the top of Table I. O_2 had higher Dice scores on synthetic images than real images on LV_{endo} segmentations,

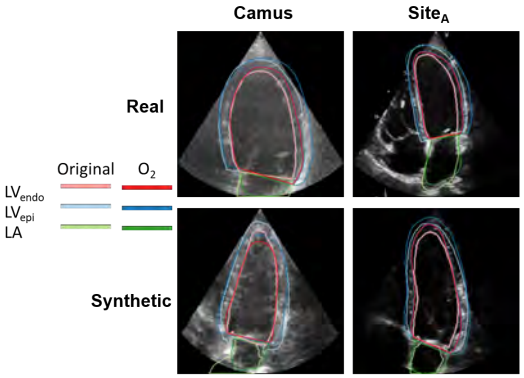


Fig. 6. **Expert annotations on synthetic images match the anatomical model annotations at a level equal to inter-observer error on real images**: A sample image with included labels from the Camus and Site$_A$ images used for the inter-observer study, chosen by taking the median Dice score between $\boldsymbol{O_2}$ and the original labeler. For the real datasets the original labeler was [30] and $\boldsymbol{O_1}$ for Camus and Site$_A$ respectively. For the synthetic datasets the original label comes from the anatomical models.

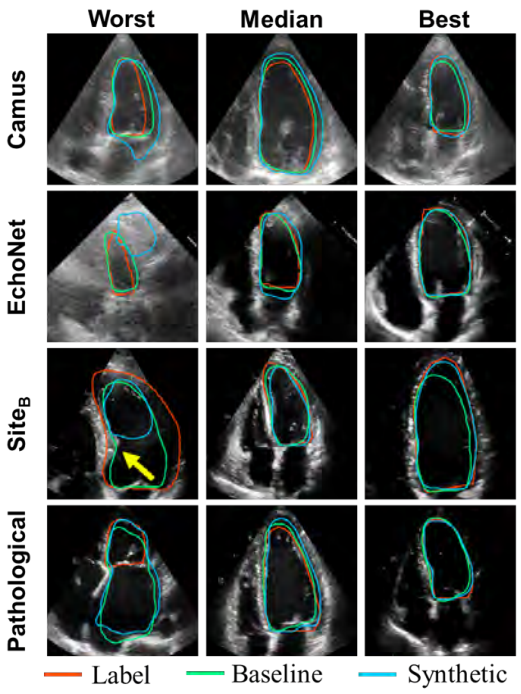


Fig. 7. **Networks trained on synthetic data produce accurate segmentations in most cases:** Worst, median, and best LV$_{endo}$ segmentation results on the Camus, EchoNet, Site$_B$, and Pathological test sets for the network trained on the baseline real data and the synthetic data. The task for all networks was LV$_{endo}$ segmentation in apical four chamber end diastole images. Images were ranked by Dice score for the network trained on synthetic data. The baseline and synthetic networks are always specific to the dataset (so for Camus the baseline network was trained on Camus and the synthetic network was trained on synthetic Camus). The yellow arrow points to a bulging septum in that image (see text). The baseline for Site$_B$ and Pathological was Site$_A$.

was comparable for LV$_{epi}$ and had higher scores on real images on LA segmentations. The median image in LV$_{endo}$ Dice score between O_2 and the original annotator is shown in Fig. 6. Overall, O_2 closely matched the pipeline-generated labels although there was some disagreement in the apical region. Fig. 6 also shows that while structure consistency between pseudo and synthetic images was not explicitly forced in the CycleGAN, the synthetic structures remain true to the original annotation mask. Only the results from O_2 are used for comparison here for simplicity and because there was a large intra-observer bias in the results for O_1. The results from O_1 are presented in Appendix C and showed the same patterns as O_2 between synthetic and real. Finally, Fig. 5 and Fig. 6 shows the difference in appearance between the different datasets for both synthetic and real images. The Camus images are typically cloudier in appearance while the EchoNet/Site$_A$ images usually have a higher gain setting and are thus clearer.

B. Learning from Synthetic Data

Networks were trained on Camus, EchoNet, Site$_A$, and each of the synthetic datasets for the task of LV$_{endo}$ segmentation in apical four chamber end diastole images. Networks were then tested on the EchoNet, Camus, Site$_B$, and Pathological test sets. Results for EchoNet are shown in Table I and for the other three sets in Appendix D. On the EchoNet test set the networks trained on real EchoNet data unsurprisingly achieved the best results, but the network trained on synthetic data was comparable to both the networks trained on separate real datasets (Camus and Site$_A$).

Qualitative results are shown in Fig. 7. In some cases the networks trained on synthetic data performed poorly. For example, in the worst case for Camus the network did not find the correct mitral valve cut-off plane. In the worst case for EchoNet the network found the wrong chamber, likely fooled by the strong reflective signal just beneath that resembles a valve. This image is also poor quality. In the worst case for Site$_B$ the network misread the bulging septum (yellow arrow) as the mitral valve and cut off the segmentation there. The anatomical models were originally built from CT scans of asymptomatic patients and thus the segmentation network from synthetic images was not exposed to pathological cases (such as those with a bulging septum) during training. This was shown explicitly on the Pathological test set where the network failed to identify the LV given an enlarged LA (although the baseline network also failed in this case). However, these results were outliers. In most cases the network trained on synthetic data performed well with annotations that are similar to the manual labels and baseline.

Next, the robustness of the synthetic data was tested by extending the task to all annotations, phases, and views. We evaluated end diastole and end systole although the synthetic datasets do not contain end systole images. Results testing on the Camus, Site$_B$, and Pathological test sets are shown in Table

TABLE I

SYNTHETIC DATA CAN EFFECTIVELY BE USED IN PLACE OF REAL DATA: MEDIAN METRICS COMPARING TRAINING WITH REAL DATASETS TO TRAINING WITH SYNTHETIC DATASETS. THE FIRST SECTION COMPARES INTER-OBSERVER RESULTS FOR O_2 ON REAL AND SYNTHETIC DATA. THE NEXT SECTION SHOWS NETWORKS TRAINED ON REAL AND SYNTHETIC DATA FOR LV_{ENDO} SEGMENTATION IN A4C ED IMAGES AND TESTED ON ECHONET. THE FINAL SECTION COMPARES NETWORKS TRAINED ON REAL AND SYNTHETIC DATA FOR ALL ANNOTATIONS/VIEWS/PHASES. ALL DICE RESULTS ARE STATISTICALLY DIFFERENT WITH A P-VALUE < 0.05 (COMPUTED WITH A WILCOXON SIGNED-RANK TEST) EXCEPT FOR THE MARKED COMPARISON (†). RESULTS ARE ORDERED BY DICE SCORE AND BOLD SHOWS THE BEST RESULT IN EACH SECTION.

Task	Training Data	Testing Data	D (%) [MAD]			B (%)			S_p	d_m
	OR inter-observer comparison		LV_{endo}	LV_{epi}	LA	LV_{endo}	LV_{epi}	LA		LV_{endo}
Inter-Observer	O_2 *vs. Camus/Site$_A$*		87.9 [4.4]	**93.4 [2.2]**	**87.5 [6.1]**	24	8	**-3**	**0.83**	5.0
	O_2 *vs. proposed pipeline*		**90.8 [2.3]**	92.0 [2.1]	83.0 [8.0]	**9**	**4**	14	0.80	**3.9**
LV_{endo} in A4C ED	EchoNet (base)	EchoNet	**94.0 [1.6]**			**0**			0.85	**2.7**
	Camus		89.6 [2.6]	n/a	n/a	-2	n/a	n/a	**0.87**	4.8
	Site$_A$		87.7 [3.8]			14			0.86	5.6
	Synth EchoNet		87.1 [3.7]			15			0.85	5.9
All	Camus (base)	Camus	**93.3 [2.3]**	**95.7 [1.0]**	**92.0 [3.2]**	**1**	**0**	1	0.84	**2.6**
	Site $_A$		88.9 [3.4]	92.3 [2.1]	85.4 [5.3]	10	-2	14	**0.85**	4.8
	Synth Camus		81.2 [5.4]	90.3 [2.8]	79.6 [8.3]	33	7	**-5**	0.83	7.5
	Synth Site$_A$	Site$_B$	**88.4 [3.8]**	90.7 [3.4]	83.1 [9.4]	**-9**	**-1**	-11	0.83	**5.4**
	Site$_A$ (base)		85.1 [4.0]	91.3 [2.0]†	84.4 [4.7]	-27	-8	**8**	**0.84**	7.5
	Camus		81.8 [4.9]	**91.9 [1.8]**†	**86.6 [4.4]**	-35	-10	-10	0.83	8.7
	Camus	Pathological	**89.8 [2.6]**	**92.5 [1.8]**	**92.2 [2.0]**	-9	**-2**	8	**0.86**	**4.1**
	Site$_A$ (base)		89.0 [3.3]	92.0 [2.9]	87.5 [3.1]	**-4**	5	19	0.85	4.2
	Synth Site$_A$		88.3 [4.9]	87.6 [4.1]	84.3 [6.1]	9	5	**-2**	0.84	4.6

LV_{endo} = left ventricle endocardium, LV_{epi} = left ventricle epicardium, LA = left atrium, A4C = apical four chamber, A2C = apical two chamber, ED = end diastole, ES = end systole. All refers to all annotations (LV_{endo}, LV_{epi}, and LA), views (A4C and A2C), and phases (ED and ES). D = Dice score, MAD = median absolute deviation, B = Bias percentage, S_p = simplicity, and d_m = mean average distance. For inter-observer S_p is listed for the second round annotations and B is calculated as O_2-Original. †: Not statistically different with a P-value < 0.05.

I. The network trained on the synthetic data performed worse in both cases on LA segmentation and for LV_{endo} segmentation in the Camus dataset. However, on Site$_B$ the synthetic network outperformed all real datasets in LV_{endo} Dice and distance scores. There was a high positive Bias for the synthetically trained networks on Camus and a strong negative Bias for Site$_A$ and Camus on Site$_B$. The network trained on synthetic data was able to achieve similar performance to the networks trained on real datasets on the Pathological dataset for LV_{endo} segmentation, although LV_{epi} and LA Dice scores were slightly lower.

C. Variability Analysis

To test the impact of parameters in the pipeline, synthetic datasets with tweaked parameters were generated and a segmentation network was trained for each. To test the effect of the transformation process, the pseudo dataset (before transformation with the CycleGAN) was compared to the synthetic dataset (after transformation with the CycleGAN). The Camus pseudo and synthetic datasets were compared to the EchoNet pseudo and synthetic dataset to analyze the effect of different parameters in the extraction process and different real datasets in the transformation process respectively. To test whether including additional variability helped, datasets extracted from just the 19 original anatomical models (Pseudo EchoNet* and Synth EchoNet*) were compared to datasets extracted from the set of 99 new shape models (Pseudo EchoNet and Synth EchoNet). To simplify results, all networks were trained for LV_{endo} segmentation only and were tested on the EchoNet test set since it was the largest.

TABLE II

EVALUATING DATA GENERATION VARIABILITY: MEDIAN RESULTS ON THE ECHONET TEST SET FOR LV_{ENDO} SEGMENTATION WHILE CHANGING VARIOUS PARTS OF THE GENERATION PIPELINE. ALL DICE RESULTS EXCEPT PSEUDO VS. PSEUDO* ARE STATISTICALLY DIFFERENT WITH A P-VALUE < 0.05 (COMPUTED WITH A WILCOXON SIGNED-RANK TEST). RESULTS ARE ORDERED BY DICE SCORE AND BOLD SHOWS THE BEST RESULT.

Train	Test	D(%) [MAD]	B (%)	d_m
Synth EchoNet	EchoNet	**87.1 [3.7]**	15	**5.9**
Synth Site$_A$		86.8 [4.0]	**14**	6.2
Synth EchoNet*		86.5 [4.0]	17	6.1
Pseudo EchoNet*		84.4 [5.3]†	14	6.7
Pseudo EchoNet		84.1 [4.5]†	18	6.7
Pseudo Camus		83.3 [4.7]	9	7.0
Synth Camus		81.9 [6.9]	13	7.3

*: these datasets were extracted from only the original 19 anatomical models rather than the extended set including the new shape models from PCA. †: Not statistically different with a P-value < 0.05.

Results are shown in Table II. Using the pseudo images provided a good baseline result even without the transformation process. Extending the anatomical model set as well as using dataset specific extraction processes slightly helped, but did not make a large difference. The transformation process did increase performance in the case that the correct dataset or a similar dataset was used (EchoNet/Site$_A$). However, using the Camus dataset actually significantly degraded the results.

V. Discussion

We developed a fully automated[4] pipeline for generating large annotated datasets for training CNNs from anatomical models. The generated synthetic images look realistic and expert annotations on the synthetic images closely matched those from the pipeline. Moreover, segmentation networks trained from the synthetic datasets produced accurate segmentations on real images in most cases. Dice scores from the synthetically trained networks were comparable to inter-observer errors and networks trained on a separate set of real data.

A. Generated Images and Annotations

We found that the expert annotations on synthetic images closely matched the ones generated by the pipeline. This indicates the paired synthetic images and labels are accurately delineating the LV in a manner consistent with expert expectations. Dice scores between experts and the anatomical model were lower (although still comparable) for the LA. To explain this, Fig. 8 shows samples from the first several modes of the shape analysis described in Sec. II-A.1. The anatomical models show complex LA shapes as well high variability in shapes between different models. However, the LA is typically still annotated as a half-ellipsoid shape by the annotators (similar to the LV - see Fig. 3) in images and we hypothesize the lower scores were due to this difference in annotation complexity. Apical images are typically optimized for image quality in the LV rather than the LA, which may hinder accurate labeling of detailed LA shapes.

The inter-observer LV_{endo} Dice scores for real images presented here are lower than those presented by Leclerc et al. on the same tasks [30]. There are two likely contributors to this. First, as discussed in Sec. III-C, a lack of explicit guidelines can cause differences in standard practice at different clinics and our results measure experts practicing in different sites. Second, in our inter-observer study the annotators were only given access to a single frame during the second round. This was necessary since the current pipeline only generates a single frame, but the lack of myocardial movement inhibits accurate detection of the compacted myocardium and other features. While more difficult, it also matches the task of the segmentation network, which is given a single frame only, and thus represents a better comparison for the pipeline.

B. Learning from Synthetic Data

We evaluated segmentation networks trained from synthetic data. First, we tested LV_{endo} segmentation in apical four chamber end diastole images and then extended the task to LV_{epi} and LA segmentation in apical four chamber and two chamber views and end diastole and end systole phases. Since there are numerous examples of deep learning methods failing once deployed due to implicit bias in the training dataset, we extensively validated our approach using five different datasets from various institutions and annotators. All hyperparameter tuning and initial tests were conducted using only a single dataset (Camus) and we then tested the same pipeline on additional unseen datasets. In some cases implicit bias towards the Camus dataset in the pseudo generation step were observed (see Appendix G), but the pipeline is still able to adapt and produce good results across datasets. This robustness is a strength of our work.

[4]Other than the manual step of selecting the CycleGAN epoch, which does not significantly impact results.

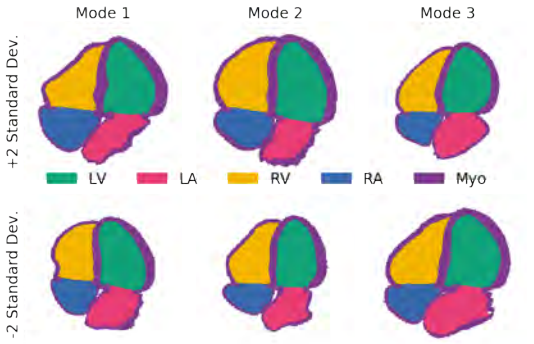


Fig. 8. **Shape variations are mainly seen in the LA, not the LV:** Four chamber slices showing $\pm$ 2 standard deviations from the mean model for the first three modes calculated using principal component analysis (see Sec. II-A.1). LV = left ventricle, LA = left atrium, RV = right ventricle, RA = right atrium, and Myo = myocardium.

The network was able to achieve comparable results to a network trained on a separate real dataset. In a review of the results, failure cases primarily occurred when the network struggled to properly identify the mitral valve plane in real images (such as the worst case in $Site_B$ of Fig. 7). Since the valve is included in the anatomical models as a flat disk, the synthetic images do not contain the same variation of valve appearances of real datasets. Including a variety of valve structures in the synthetic images is one way the proposed pipeline could be improved. The network trained on synthetic data was generally able to segment images from the Pathological dataset well, but could not properly identify the LV in cases with an enlarged LA (shown in Fig. 7 and in supplementary material). However, networks trained on real data also struggled on these images indicating that these cases would likely require expert review and adjustments regardless of the dataset used. If a known pathology should be handled, the models could also be adjusted to include this by including a single anatomical model exhibiting this pathology and using the PCA shape analysis to generate variations compared to a healthy normal model.

C. Clinical Applicability

LV_{endo} segmentation is used clinically for an estimation of volumes and ejection fraction which are important measures of the efficiency of heart function. Clinical measures are not presented here because metric pixel sizes are not given for the datasets. However, previous studies have shown a strong correlation between the accuracy of Dice scores and the accuracy of predictions of clinical parameters across mul-

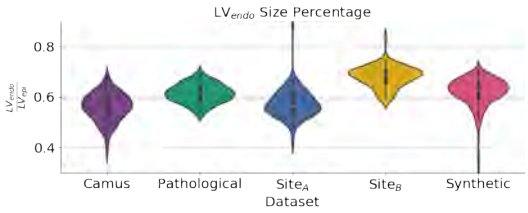


Fig. 9. **Annotation bias can yield large differences:** Violin plot showing the amount of LV_{epi} area comprised of the LV_{endo} area for the labels in each dataset. A lower value indicates a thicker myocardium. EchoNet does not have LV_{epi} labels.

tiple algorithms and inter-/intra-observer studies (correlation coefficient of -0.92 between Dice scores at end-diastole/end-systole and ejection fraction mean average error across 12 experiments) [30]. Thus, the small decrease in accuracy of Dice scores presented here would likely result in a small decrease in accuracy of clinical metrics. The Dice scores obtained with the synthetically trained networks are still within the range of inter-observer error, indicating the same would likely be true for clinical metrics. Annotators rely on visually tracking the same point across the cycle to ensure consistency between predictions at end-diastole and end-systole and ejection fraction prediction could also be improved by including this temporal coherence between the predictions at different phases in the segmentation networks (using recursive neural networks for example).

D. Variability Analysis

We also analyzed potential sources of error for the networks. When testing images trained on one dataset on a different dataset there are two primary elements that cause decreased performance:

1) **Texture differences**: In echo these are linked to acquisition changes such as varying ultrasound machines, gain, focus, resolution, and other imaging parameters. In the proposed pipeline, texture primarily comes from the transformation step.
2) **Shape differences**: Due to a) differences in the width/depth of the acquisition which change tissue shape in the produced image, b) changes in the underlying tissue shape, or c) differences in annotation style. In the proposed pipeline, shape changes comes from the extraction step. Annotation style is linked to the original anatomical models.

Texture and shape differences were previously explored in object recognition where Geirhos *et al.*, who showed that CNNs trained on ImageNet for classification were more biased towards changes in texture than shape [37]. We tested these differences in echo segmentation in Table II, using the generation pipeline to isolate the impact of each component.

Changes in shape due to imaging parameters were isolated by varying the width/depth/percentage of LV focused images in the two pseudo datasets and had a very small effect. Changes in underlying tissue shape were isolated by comparing the datasets built from the original models (Pseudo EchoNet*/Synth EchoNet*) to the set of models containing additional variability from the shape extension in Sec. II-A.1 (Pseudo EchoNet/Synth EchoNet). Changes in results were small and reversed between the pseudo and synthetic sets. This is likely because there were minimal variations in LV shape. As shown in Fig. 8, the largest changes in the LV are variations in size and width. Modifications to these parameters are already included in the pseudo image generation process, thus the shape extension did not add significant new variations of LV shape to the dataset. Pathological changes in the underlying shape (such as the bulging septum or enlarged left atrium in Fig. 7) do seem to reduce segmentation accuracy. To include these elements in the pipeline, new models could be built from pathological cases as discussed above.

Texture changes were isolated by comparing different synthetic datasets using CycleGANs tuned to different real datasets since the same underlying shape was used in all cases. Results showed that image appearance could make a significant difference as the Synth Camus network performed significantly worse than Synth EchoNet/Synth $Site_A$. This matches the qualitative appearance difference between EchoNet/$Site_A$ and Camus in Fig. 7. Results here also showed that solid performance could be obtained with only the pseudo network. This is an encouraging result indicating that applications without high accuracy needs could further simplify the pipeline by removing the transformation step.

Assuming that human observers are adept at adapting to differences in texture and shape, differences in annotator style can be isolated from the inter-observer study presented in Table I. Differences between observers were substantial both in terms of Dice score and Bias, indicating a systematic difference between annotators. Although there were various constraints in this study (as discussed above), this difference was also clearly present in the original datasets without those constraints. Fig. 9 shows the ratio of LV_{endo} area to LV_{epi} area for the labels of each dataset which generally corresponds to the thickness of the labeled myocardium. This percentage is much higher (indicating a thinner myocardium) for the synthetic datasets than all the other datasets excluding $Site_B$. While this difference could instead indicate the prevalence of pathologies (e.g. hypertension) in the dataset, we present additional validation in Appendix E that the differences in Fig. 9 are primarily due to changes in annotation style. Results also match previous studies showing echo measurements typically overestimate the thickness of the myocardium [38]–[40].

Our segmentation results also point towards annotation style as the critical factor in determining accuracy. Bias was high for LV_{endo} results for networks trained on synthetic data on all other datasets than $Site_B$. On the other hand, networks trained on real datasets had a high negative Bias when tested on $Site_B$. The increase in Bias was correlated with lower Dice sores and higher mean distances, but not with simplicity, showing that the segmentations were still well-formed. This Bias was not observed for LV_{epi} in Table I indicating that the variation comes purely from the differences in LV_{endo} annotation style. The high performance of the synthetic network on $Site_B$ matches both Fig. 9 and the low bias with O_2 in the inter-

observer study since O_2 labeled Site_B.

Therefore, the primary reason for decreased performance in our experiments (for networks trained on both synthetic and real data) was differences in annotation style, with texture differences playing a secondary role. Other than several outlier cases, the networks trained on synthetic data performed well and produced well-formed segmentations. One of the advantages of the pipeline proposed in this work is that the same annotation style can be applied to images from any dataset which will bring consistent performance for a network implemented in clinical practice. Given that the synthetic images are built from anatomical models derived from CT images, the synthetic images generated can be used to standardize annotation style.

E. Extensions and Future Applications

An abundance of augmentation techniques exist specifically for improving segmentation performance on limited datasets. For example, several authors introduced method based on statistical models to modify images following the deconstructed natural shape variation [8], [9]. Methods such as Jafari et al. [41] or Shin et al. [42] use GANs to expand the dataset with new natural images. This work focuses on the performance of the standard pipeline rather than one with augmentations tuned for a specific application, but these techniques, as well as any other task-specific augmentation techniques (or loss functions), could readily be applied here to improve results.

While we implemented the pipeline for 2D LV/LA echo segmentation to enable comparison against existing techniques, one of the strengths of our method is that the anatomical models are 3D and contain annotations for a variety of tissue types. Moreover, our method is not limited to ultrasound and a paired database of CT or MRI images could also be generated using this method. The pipeline is theoretically extensible to any segmentation or landmark detection task. Extension requires a) a small set of anatomical shape models similar to those described in Section III-A, b) a real dataset of unlabeled images from the relevant modality and view, and c) code to extract a slice from the anatomical models matching real images. Part c) can be accomplished through an analysis of important landmarks present in the relevant images that are also defined in the model. Additional unforeseen challenges likely exist for adapting to new anatomies and modalities, but we anticipate the ability to overcome these.

In addition to testing the pipeline on novel applications, future work will focus on adapting the pipeline to 3D, which is increasingly being used in clinical practice, but where manual labeling is even more difficult. The difficulty of manual labeling has thus far limited the development of benchmark datasets which is why the focus of this validation work is limited to 2D images. While challenging, other groups have previously shown the ability to adapt generative networks for 3D medical image synthesis (for example [42] and [43]). Due to GPU memory constraints these works required use of lower resolution volumes, a challenge for adapting the existing pipeline as well. The anatomical models could also be used as context for generation and/or segmentation as was proposed in [44]. Additionally, one of the strengths of echo is the high temporal resolution. Future work will also focus on extending image generation techniques to include labels and images across the cardiac cycle.

VI. Conclusion

Building large annotated datasets can be difficult and time-consuming. For cases where a small percentage of outliers are acceptable, or a confidence metric can be designed to catch outliers, we present a method to train a cardiac segmentation network with zero manual labeling required. The generated labels represent an accurate ground truth, can be rapidly built, and grant additional flexibility since the anatomical models providing the ground truth can be automatically adjusted as required. By eliminating or reducing labeling requirements, the proposed pipeline enables greatly accelerated deep learning algorithm development in cardiac imaging.

Acknowledgment

The authors thank Svein Arne Aase and Julia Schnabel for guidance and feedback throughout this work and also Daria Kulikova and Anna Novikova for assistance with annotating images.

References

[1] A. C. Armstrong, E. P. Ricketts, C. Cox, P. Adler, A. Arynchyn, K. Liu, E. Stengel, S. Sidney, C. E. Lewis, P. J. Schreiner, J. M. Shikany, K. Keck, J. Merlo, S. S. Gidding, and J. A. Lima, "Quality Control and Reproducibility in M-Mode, Two-Dimensional, and Speckle Tracking Echocardiography Acquisition and Analysis: The CARDIA Study, Year 25 Examination Experience," *Echocardiography*, vol. 32, no. 8, pp. 1233–1240, 2015.

[2] A. Thorstensen, H. Dalen, B. H. Amundsen, S. A. Aase, and A. Stoylen, "Reproducibility in echocardiographic assessment of the left ventricular global and regional function, the HUNT study," *Eur. J. Echocardiogr.*, vol. 11, no. 2, pp. 149–156, 2010.

[3] J. Corral-Acero, F. Margara, M. Marciniak, C. Rodero, F. Loncaric, Y. Feng, A. Gilbert, J. F. Fernandes, H. A. Bukhari, A. Wajdan, M. V. Martinez, M. S. Santos, M. Shamohammdi, H. Luo, P. Westphal, P. Leeson, P. DiAchille, V. Gurev, M. Mayr, L. Geris, P. Pathmanathan, T. Morrison, R. Cornelussen, F. Prinzen, T. Delhaas, A. Doltra, M. Sitges, E. J. Vigmond, E. Zacur, V. Grau, B. Rodriguez, E. W. Remme, S. Niederer, P. Mortier, K. McLeod, M. Potse, E. Pueyo, A. Bueno-Orovio, and P. Lamata, "The 'Digital Twin' to enable the vision of precision cardiology," *Eur. Heart J.*, pp. 1–11, 2020.

[4] W. Kainz, E. Neufeld, W. E. Bolch, C. G. Graff, C. H. Kim, N. Kuster, B. Lloyd, T. Morrison, P. Segars, T. S. Yeom, M. Zankl, X. G. Xu, and B. M. W. Tsui, "Advances in Computational Human Phantoms and Their Applications in Biomedical Engineering – A Topical Review," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 3, no. 1, 2019.

[5] X. G. Xu, "An exponential growth of computational phantom research in radiation protection, imaging, and radiotherapy: A review of the fifty-year history," *Phys. Med. Biol.*, vol. 59, no. 18, 2014.

[6] M. Caon, "Voxel-based computational models of real human anatomy: A review," *Radiat. Environ. Biophys.*, vol. 42, no. 4, pp. 229–235, 2004.

[7] F. Jay, J.-P. Renou, O. Voinnet, and L. Navarro, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks Jun-Yan," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 183–202.

[8] J. Corral Acero, E. Zacur, H. Xu, R. Ariga, A. Bueno-Orovio, P. Lamata, and V. Grau, "SMOD - Data Augmentation Based on Statistical Models of Deformation to Enhance Segmentation in 2D Cine Cardiac MRI," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11504 LNCS, pp. 361–369, 2019.

[9] R. Bhalodia, S. Y. Elhabian, L. Kavan, and R. T. Whitaker, "DeepSSM: A Deep Learning Framework for Statistical Shape Modeling from Raw Images," *Int. Conf. Med. Image Comput. Comput. Interv.*, vol. 11167 LNCS, pp. 244–257, 2018.

[10] V. Tavakoli and A. A. Amini, "A survey of shaped-based registration and segmentation techniques for cardiac images," *Comput. Vis. Image Underst.*, vol. 117, no. 9, pp. 966–989, 2013.

[11] G. Allan, S. Nouranian, T. Tsang, A. Seitel, M. Mirian, J. Jue, D. Hawley, S. Fleming, K. Gin, J. Swift, R. Rohling, and P. Abolmaesumi, "Simultaneous Analysis of 2D Echo Views for Left Atrial Segmentation and Disease Detection," *IEEE Trans. Med. Imaging*, vol. 36, no. 1, pp. 40–50, 2017.

[12] S. Kazeminia, C. Baur, A. Kuijper, B. Van Ginneken, N. Navab, S. Albarqouni, and A. Mukhopadhyay, "GANs for Medical Image Analysis," Tech. Rep.

[13] S. A. Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep Semantic Segmentation of Natural and Medical Images: A Review," 2019.

[14] M. H. Jafari, H. Girgis, N. Van Woudenberg, N. Moulson, C. Luong, A. Fung, S. Balthazaar, J. Jue, M. Tsang, P. Nair, K. Gin, R. Rohling, P. Abolmaesumi, and T. Tsang, "Cardiac point-of-care to cart-based ultrasound translation using constrained CycleGAN," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 15, no. 5, pp. 877–886, 2020.

[15] A. H. Abdi, M. H. Jafari, S. Fels, T. Tsang, and P. Abolmaesumi, "A Study into Echocardiography View Conversion," *arXiv Prepr.*, 2019.

[16] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Med. Image Anal.*, vol. 63, 2020.

[17] D. Eschweiler, T. Klose, F. N. Muller-Fouarge, M. Kopaczka, and J. Stegmaier, "Towards Annotation-Free Segmentation of Fluorescently Labeled Cell Membranes in Confocal Microscopy Images," in *Int. Work. Simul. Synth. Med. Imaging*, 2019, pp. 81–89.

[18] Y. Huo, Z. Xu, H. Moon, S. Bao, A. Assad, T. K. Moyo, M. R. Savona, R. G. Abramson, and B. A. Landman, "Synseg-net: Synthetic segmentation without target modality ground truth," *IEEE Trans. Med. Imaging*, vol. 38, no. 4, pp. 1016–1025, 2018.

[19] M. Alessandrini, M. De Craene, O. Bernard, S. Giffard-Roisin, P. Allain, I. Waechter-Stehle, J. Weese, E. Saloux, H. Delingette, M. Sermesant, and J. D'hooge, "A Pipeline for the Generation of Realistic 3D Synthetic Echocardiographic Sequences: Methodology and Open-Access Database," *IEEE Trans. Med. Imaging*, vol. 34, no. 7, pp. 1436–1451, 2015.

[20] M. De Craene, S. Marchesseau, B. Heyde, H. Gao, M. Alessandrini, O. Bernard, G. Piella, A. R. Porras, L. Tautz, A. Hennemuth, A. Prakosa, H. Liebgott, O. Somphone, P. Allain, S. Makram Ebeid, H. Delingette, M. Sermesant, J. D'Hooge, and E. Saloux, "3D strain assessment in ultrasound (Straus): A synthetic comparison of five tracking methodologies," *IEEE Trans. Med. Imaging*, vol. 32, no. 9, pp. 1632–1646, 2013.

[21] Y. Zhou, S. Giffard-Roisin, M. De Craene, S. Camarasu-Pop, J. D'Hooge, M. Alessandrini, D. Friboulet, M. Sermesant, and O. Bernard, "A Framework for the Generation of Realistic Synthetic Cardiac Ultrasound and Magnetic Resonance Imaging Sequences from the Same Virtual Patients," *IEEE Trans. Med. Imaging*, vol. 37, no. 3, pp. 741–754, 2018.

[22] Q. Duan, P. Moireau, E. D. Angelini, D. Chapelle, and A. F. Laine, "Simulation of 3D ultrasound with a realistic electro-mechanical model of the heart," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4466 LNCS, no. May 2014, pp. 463–473, 2007.

[23] A. H. Abdi, T. Tsang, and P. Abolmaesumi, "GAN-enhanced Conditional Echocardiogram Generation," 2019.

[24] K. Abhishek and G. Hamarneh, "Mask2Lesion: Mask-constrained adversarial skin lesion image synthesis," in *Int. Work. Simul. Synth. Med. Imaging*. Springer, 2019, pp. 71–80.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Int. Conf. Med. image Comput. Comput. Interv.*, 2015.

[26] M. H. Jafari, Z. Liao, H. Girgis, M. Pesteie, R. Rohling, K. Gin, T. Tsang, and P. Abolmaesumi, "Echocardiography Segmentation by Quality Translation Using Anatomically Constrained CycleGAN," in *Int. Conf. Med. Image Comput. Comput. Interv.* Springer, 2019, pp. 655–663.

[27] C. Rodero, M. Strocchi, M. Marciniak, J. Whitaker, D. O. Neill, K. Gillette, C. Augustin, G. Plank, E. Vigmond, P. Lamata, and S. A. Niederer, "Anatomical changes influences mechanics, electrophysiologyand haemodynamics in a complementary and localised wayin the healthy adult human heart," *PLOS Comput. Biol. (under Rev.)*.

[28] M. Strocchi, C. M. Augustin, M. A. Gsell, E. Karabelas, A. Neic, K. Gillette, O. Razeghi, A. J. Prassl, E. J. Vigmond, E. J. Vigmond, J. M. Behar, J. M. Behar, J. Gould, J. Gould, B. Sidhu, B. Sidhu, C. A. Rinaldi, C. A. Rinaldi, M. J. Bishop, G. Plank, and S. A. Niederer, "A publicly available virtual cohort of fourchamber heart meshes for cardiac electromechanics simulations," *PLoS One*, vol. 15, no. 6 June, pp. 1–26, 2020.

[29] C. Mitchell, P. S. Rahko, L. A. Blauwet, B. Canaday, J. A. Finstuen, M. C. Foster, K. Horton, K. O. Ogunyankin, R. A. Palma, and E. J. Velazquez, "Guidelines for Performing a Comprehensive Transthoracic Echocardiographic Examination in Adults: Recommendations from the American Society of Echocardiography," *J. Am. Soc. Echocardiogr.*, vol. 32, no. 1, pp. 1–64, 2019.

[30] S. Leclerc, E. Smistad, A. Ostvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, C. Lartizien, L. Lovstakken, and O. Bernard, "Deep Learning Segmentation in 2D echocardiography using the CAMUS dataset : Automatic Assessment of the Anatomical Shape Validity," in *Int. Conf. Med. Imaging with Deep Learn. – Ext. Abstr. Track*, 2019.

[31] Y. Zhu, Y. Tian, D. Metaxas, and P. Dollar, "Semantic amodal segmentation," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 3001–3009, 2017.

[32] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, 1945.

[33] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1521–1528, 2011.

[34] D. Ouyang, B. He, A. Ghorbani, C. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, J. Euan A Ashley, and a. Y. Zou, "Interpretable AI for beat-to-beat cardiac function assessment," *Nature*, 2020.

[35] H. N. Sabbah, T. Kono, H. Rosman, S. Jafri, P. D. Stein, and S. Goldstein, "Left ventricular shape: A factor in the etiology of functional mitral regurgitation in heart failure," *Am. Heart J.*, vol. 123, no. 4 PART 1, pp. 961–966, 1992.

[36] T. Kono, H. N. Sabbah, P. D. Stein, J. F. Brymer, and F. Khaja, "Left ventricular shape as a determinant of functional mitral regurgitation in patients with severe heart failure secondary to either coronary artery disease or idiopathic dilated cardiomyopathy," *Am. J. Cardiol.*, vol. 68, no. 4, pp. 355–359, 1991.

[37] R. Geirhos, C. Michaelis, F. A. Wichmann, P. Rubisch, M. Bethge, and W. Brendel, "Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," *7th Int. Conf. Learn. Represent. ICLR 2019*, no. c, pp. 1–22, 2019.

[38] R. B. Devereux, D. R. Alonso, E. M. Lutas, G. J. Gottlieb, I. Sachs, and N. Reichek, "Echocardiographic Assessment of Left Ventricular Hypertrophy: Comparison to Necropsy Findings," *Am. J. Cardiol.*, vol. 57, pp. 450–458, 1986.

[39] S. Malm, S. Frigstad, E. Sagberg, H. Larsson, and T. Skjaerpe, "Accurate and reproducible measurement of left ventricular volume and ejection fraction by contrast echocardiography: A comparison with magnetic resonance imaging," *J. Am. Coll. Cardiol.*, vol. 44, no. 5, pp. 1030–1035, 2004.

[40] V. Mor-Avi, C. Jenkins, H. P. Kühl, H.-J. Nesser, T. Marwick, A. Franke, C. Ebner, B. H. Freed, R. Steringer-Mascherbauer, H. Pollard, L. Weinert, J. Niel, L. Sugeng, and R. M. Lang, "Real-Time 3-Dimensional Echocardiographic Quantification of Left Ventricular Volumes," *JACC Cardiovasc. Imaging*, vol. 1, no. 4, pp. 413–423, 2008.

[41] M. H. Jafari, H. Girgis, A. H. Abdi, Z. Liao, M. Pesteie, R. Rohling, K. Gin, T. Tsang, and P. Abolmaesumi, "Semi-supervised learning for cardiac left ventricle segmentation using conditional deep generative models as prior," *Proc. - Int. Symp. Biomed. Imaging*, vol. 2019-April, no. Isbi, pp. 649–652, 2019.

[42] H. C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, and M. Michalski, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11037 LNCS, pp. 1–11, 2018.

[43] Z. Zhang, L. Yang, and Y. Zheng, "Translating and Segmenting Multimodal Medical Volumes with Cycle- and Shape-Consistency Generative Adversarial Network," *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp. 9242–9251, 2018.

[44] C. Wang and Ö. Smedby, "Automatic whole heart segmentation using deep learning and shape context," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10663 LNCS, no. March, pp. 242–249, 2018.

Supplementary Materials for Generating Synthetic Labeled Data from Existing Anatomical Models: An Example with Echocardiography Segmentation

TABLE I
ABBREVIATIONS USED IN THE APPENDICES.

ED	End diastole
ES	End systole
A4C	Apical four chamber
A2C	Apical two chamber
LV	Left ventricle
LA	Left atrium
LV_{endo}	LV endocardium
LV_{epi}	LV epicardium

APPENDIX A
ADDITIONAL IMPLEMENTATION DETAILS

A. Original Anatomical Models

The original anatomical models used were a set of 19 models developed from 3D CT data by Rodero *et al.* [1]. These models will be made publicly available with the publication of [1] (currently under review). The models were developed to enable electromechanical simulations of a virtual patient cohort for cardiac resynchronization therapy and other treatments targeting heart failure. The model construction process mimicked that of Strocchi *et al.* [3] and full details on the construction process are available there. In brief, the models were constructed using a semi-automatic segmentation of 3D CT scans. The models include labels for all major cardiac anatomical regions, although valve anatomies are simplified to a 2mm thick plane since valve anatomies are not visible in the CT scans. Fig. 1 provides an example of the construction process. Note the level of detail in the models, which is a product of the high resolution in CT images.

B. New Anatomical Model Construction

A statistical shape model was constructed from the 19 original models. The shape model captures the distribution of cardiac shapes of the cohort of original meshes. To construct the shape model the meshes were rigidly aligned in 3D space, using the barycenters of the tricuspid valve, mitral valve and the lowest point of the interventricular septum. An arbitrarily chosen subject served as a reference for computation of rotation and translation matrices for all other cases. The alignment was performed to focus on quantifying the variability in the anatomies and attenuate the bias in construction of the mean shape.

Registration on the rigidly aligned surface meshes was performed using the Deformetrica software. The program allows for omitting the search for point-to-point correspondences and allowed for comparison based on the geometrical features where the interrelationships are non-parametric. The anatomical mean shape and the variability around it is computed from the surface meshes, represented with mathematical currents [4], [5]. In this process, every model can be obtained by applying subject

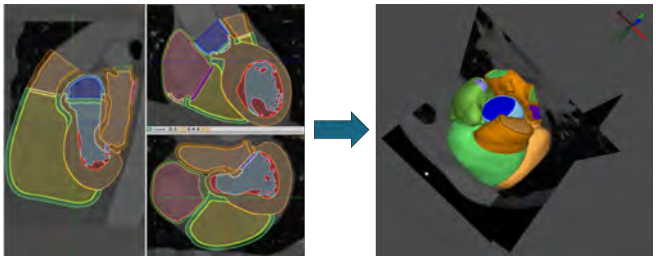


Fig. 1. Original model construction from Rodero *et al.* [1] using the Seg3D tool [2].

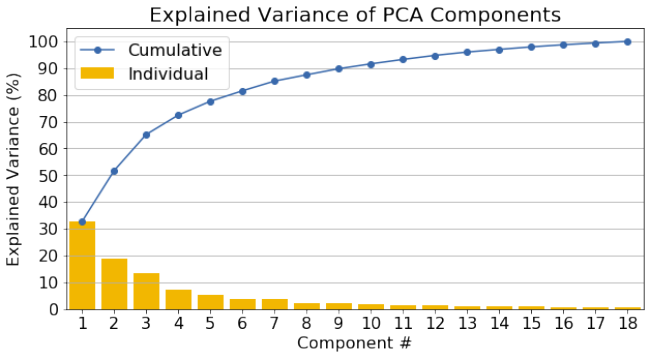


Fig. 2. The percentage of variance explained by each component after PCA.

specific transformation function to the template and adding residual information [6]. In this process, every model T_i from patient i can be obtained by applying subject specific transformation function ϕ_i to the template T^- and adding residual information as shown in Eq. (2) where $n \in \{0, 1\}$ and m is the shape mode [6].

$$H_j = (-1)^n \dot{\sum}_{k=0}^{8} m_k w_k^j \tag{1}$$

After the registration of the meshes, Principal Component Analysis (PCA) was applied to the deformation vectors characterised by functions ϕ_i to find the primary modes of shape variability within the population. The aim of PCA is to minimise the number of variables representing each sample while maximising the variance explained by these variables. The trade-off between model complexity and the amount of explained variability is controlled with the number of components chosen to represent the data set. The amount of variance explained by each component is shown in Fig. The 9 most prevalent shape modes computed with PCA captured over 90% of the variability and were chosen for the generative model.

The mesh k in the shape distribution of the original cohort can be approximated as a linear combination of the chosen PCA modes, weighted by certain weights w_k. We randomly sample each of these 9 modes within 2 standard deviations (square roots of the eigenvalues computed with PCA) to generate new synthetic meshes H_j as shown in

$$H_j = (-1)^n \dot{\sum}_{k=0}^{8} m_k w_k^j \tag{2}$$

where $n \in \{0, 1\}$ and m is the shape mode. Each of the synthetic models is still an anatomically plausible shape, but adds a heterogeneous example to our dataset. We repeat this procedure to generate 99 models in total. The generated surface meshes are transformed into volumetric meshes to allow for easy slice extraction.

The rigid transformation of the meshes, surface extraction, splitting the mesh into elements, labeling and merging was performed with Python programming language, using the VTK package [7]. Meshes were registered with Deformetrica software [8] and its atlas construction module. Transformation from surface to volumetric meshes was performed with gmsh [9].

C. 2D Slicing

Example two chamber and four chamber slices as well as their relationship are shown in Fig. 3. Valve center landmarks are defined as the center of mass of the respective tissue types while the apex is defined as the point in the LV myocardium which is farthest from the mitral valve center. The long axis is defined as the vector from the mitral valve center to the apex while the short axis is defined as the vector from mitral valve to tricuspid valve. These rotations were sampled from normal distributions with standard deviation of 16 and 9 degrees respectively. An example of extracting slices using varying rotations is shown in Fig. 3. Slicing was also implemented using the VTK package.

D. Pseudo Images

The generation of the pseudo images is application specific. It was performed based on an analysis of real A2C and A4C images. The first part of transforming the pseudo image is positioning the slice within the image. This consists of several steps:

1) An initial rough orientation of the slice. For A2C/A4C images this means rotating the slice such that the apex is at the top of the image and centering the LV within the image. Variations in the exact rotation and positioning are applied as an augmentation.

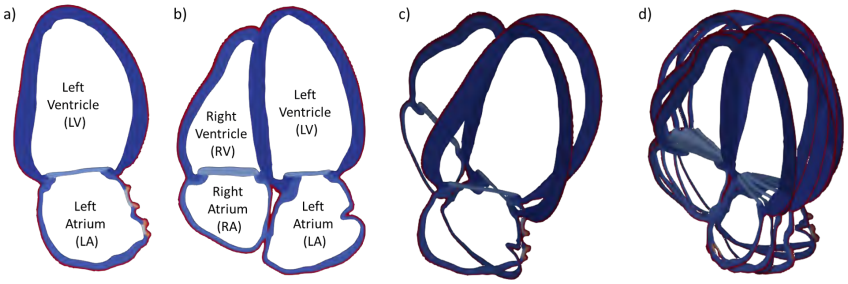


Fig. 3. Model slices as shown in Paraview [10]. a) Two chamber slice with labeled chambers in a model b) Apical four chamber slice with labeled chambers in a model c) Relationship between apical two chamber and apical four chamber slices in a model d) Example of how variable apical four chamber slices are extracted from the same model by varying rotation parameters.

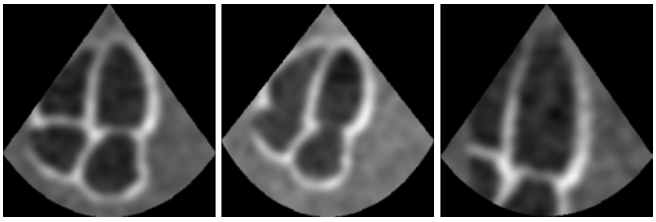


Fig. 4. Example of expanding the dataset using variable cropping parameters and affine transforms in the pseudo image construction step. All 3 pseudo images were constructed from the same slice. The first two images are full heart images while the last is LV-focused. The valve was removed from the middle image.

2) Following the real distribution, images were cropped to show either the entire heart or the left ventricle. Cropping was performed by cropping to $\pm$ 10% of the max/min locations of the relative attribute (pericardium for full heart images or LV myocardium for LV focused images). Fig. 4 shows an example of three different pseudo images created from the same slice where the first two are cropped to the pericardium and the last is cropped to the myocardium.
3) A2C/A4C images often are squeezed horizontally, particularly for full-heart images which much have a wider field of view to fit all chambers. When compressed to a 256 $\times$ 256 this has the effect of squeezing. Full heart images were squeezed by 30% and LV focused images were squeezed by 20%.
4) A cone of random width and depth was generated and used to mask the slice. The slice was shifted to ensure that the entire LV chamber was within the cone.

The second step of the pseudo transform is to modify the appearance to add noise and remove hard edges. The transformations applied to the slice consisted of several repetitions of a) adding /subtracting random uniform noise, b) multiplying by down-sampled random uniform noise, c) adding random shadowing or brightness in the form of 2D Gaussian kernels of varying size, d) shadowing the cone origin with the same method, and e) convolving with a Gaussian kernel to blur the image and remove hard edges. These transforms were done to add randomness to the pseudo images while also removing the hard edges (see Appendix F).

One of the primary problems for the segmentation network was correctly finding the mitral valve (see Appendix G). In order to try to increase the robustness of the network in finding the feature we randomly removed the model valve from 50% of the pseudo images (e.g. middle image of Fig. 4). We experimented to see if the CycleGAN would generate realistic valve features but it had no effect on preliminary results.

This step was performed individually for each dataset. The percentage of LV focused vs. whole heart images was set individually for each dataset based on qualitative analysis of a subset of 20 random images (60% for EchoNet and 50% for Camus/Site$_A$) and image widths/depths were modified accordingly.

E. Transformation

Image flipping was disabled in the CycleGAN training since A4C and A2C images are always oriented in the same direction. Network architectures for the generator and discriminator are shown in Tables II and III respectively. The generator is a UNet [11] with 8 downsampling/upsampling layers using 2$\times$2 strided convolutions and instance normalization [12]. The discriminator is a PatchGAN with a 70$\times$70 pixel field of view. Both networks were implemented following [13]. The UNet contains 5.4M

TABLE II
NETWORK ARCHITECTURE FOR THE CYCLEGAN GENERATOR AND SEGMENTATION. THE NETWORK IS COMPOSED OF INPUT AND OUTPUT LAYERS AND 8 RECURSIVE UNET MODULE BLOCKS WHERE A UNET MODULE IS SHOWN BELOW. FILTER SIZES ARE DOUBLED DURING EVERY INITIALIZATION (STARTING AT 128) UNTIL REACHING A MAX VALUE OF 512.

Name	Type	Filters	Kernel	Stride	Activation	Normalization	Skip Connection
Input	Conv 2D	64	4×4	2×2			
UNet Module(F = 128)$^{\diamond}$	Conv 2D	F=128	4×4	2×2	Leaky ReLU	Instance 2D†	*
	UNet Module(min(F×2, 512))$^{\diamond}$						*
	Conv Transpose 2D	F=128	4×4	2×2	ReLU	Instance 2D	
Output	Conv Transpose 2D	1/2/4‡	4×4	2×2	ReLU		

Activation layers are processed before the main layer in the same row, normalization layers are processed after. ∗ indicates a skip connection between the outputs of the marked rows. $^{\diamond}$: the UNet Module is recursively included 8 times with filter size F doubling in each iteration until reaching 512. †: the Instance Norm 2D is not included in the innermost layer. ‡: Number of Output channels depends on the task.

TABLE III
NETWORK ARCHITECTURE FOR THE CYCLEGAN DISCRIMINATOR.

Name	Type	Filters	Kernel	Stride	Activation	Normalization
Input	Conv 2D	64	4×4	2×2		
L1	Conv 2D	128	4×4	2×2	Leaky ReLU	Instance 2D
L2	Conv 2D	256	4×4	2×2	Leaky ReLU	Instance 2D
L3	Conv 2D	512	4×4	2×2	Leaky ReLU	Instance 2D
Output	Conv 2D	128	4×4	2×2	Leaky ReLU	

Activation layers are processed before the main layer in the same row, normalization layers are processed after.

parameters and the PatchGAN contained 276k parameters. Other transformation parameters followed the defaults in [13]. Because the network struggles to generate the smooth cone outline of the ultrasound image, the label cone was used to mask the generated image during the final inference.

F. Segmentation

The segmentation network was also a UNet with 8 downsampling layers with the same architecture as Table II except the output channels were modified based on the task. We tested a smaller version of this architecture as well as the UNet1 and UNet2 architectures from [14] but found no increase in results in initial validation tests on the synthetic validation set. Preprocessing consisted of random cropping (256×256 crops from 286×286 images), gamma modifications (coefficient between 0.5 and 1.2), and mean normalization (specific to each dataset) when training. At test time all images were loaded as 256×256 and mean normalization was applied. The segmentation masks were post-processed to extract only the largest contiguous activation region and holes in the segmentation mask were filled. Learning rate was set to 1e-4 and decreased by a factor of 10 after 10 epochs. We used an Adam optimizer [15] and a batch size of 16. These parameters were set from previous experience with similar problems and a rough sweep while validating on the synthetic data. Within standard ranges these hyperparameters did not have a large impact on the performance of the segmentation network. We also experimented with several other standard segmentation loss functions including Dice loss and weighted cross entropy loss but found no increase in performance. Because the synthetic images are derived from models, they are more consistent than manual annotations and it is easier for the network to achieve high Dice scores on training and validation, which decreases the value of sweeping hyperparameters. All network training and inference was conducted in PyTorch 1.5.0 using an NVIDIA Titan GPU.

G. Datasets

A complete breakdown of dataset sizes by view and phase is shown in Table IV. 891 pseudo images were originally generated in the pipeline (19 models × 3 slices per model × 3 pseudo images per slice) but several were eliminated because the LV and LA regions were not adjacent (due to an improper cut plane). An individual pseudo dataset was generated for each corresponding synthetic dataset, but only one is shown in the table since the characteristics matched. The usage columns show how each dataset was used in the pipeline. The Camus, EchoNet, and Site_A real datasets were used to train the CycleGAN as well as for the label-free network selection described in the text. The synthetic datasets were used for supervised training of the segmentation networks. Finally, the Camus, EchoNet, Site_B, and Pathological real datasets were used to evaluate the networks and generate the results presented in this work.

TABLE IV
SIZES OF EACH DATASET INCLUDING TRAINING/VALIDATION/TEST SPLITS AND BREAKDOWNS BY VIEW AND PHASE.

Dataset	Annotator	Total	Split	Annotations			Size	Views		Phases		Usage		
				LV_{endo}	LV_{epi}	LA		A4C	A2C	ED	ES	Transform	Learn	Test
Camus [14]	[14]	1,600	train val test	✔	✔	✔	1280 160 160	656 80 82	624 80 78	640 80 80	640 80 80	○	○	●
EchoNet [16]	[16]	20,048	train val test	✔			14920 2576 2552	14920 2576 2552	0 0 0	7460 1288 1276	7460 1288 1276	○	○	●
$Site_A$	O_1	336	train val	✔	✔	✔	281 55	156 40	125 15	138 27	143 28	○	○	
$Site_B$	O_2	229	test	✔	✔	✔	229	149	80	115	114			●
Pathological	O_2	122	test	✔	✔	✔	122	122	0	61	61			●
Pseudo	Models	1754	train val	✔	✔	✔	1415 339	713 168	702 171	1415 339	0 0		●	
Synth Camus	Models	1754	train val	✔	✔	✔	1415 339	713 168	702 171	1415 339	0 0		●	
Synth EchoNet	Models	879	train val	✔	✔	✔	711 168	711 168	0 0	711 168	0 0		●	
Synth $Site_A$	Models	1754	train val	✔	✔	✔	1414 340	712 169	702 171	1414 340	0 0		●	

LV_{endo} = left ventricle endocardium, LV_{epi} = left ventricle epicardium, LA = left atrium, A4C = apical four chamber, A2C = apical two chamber, ED = end diastole, ES = end systole. ✔ indicates that the indicated annotation/view/phase is present in the dataset. For usage, ○ indicates that only the images were used (not labels) while ● indicates both images and labels were used. Transform, Segment, and Test correspond to b)-d) of Fig. 2 in the manuscript. Note that the usage markers includes only the proposed pipeline not comparison experiments.

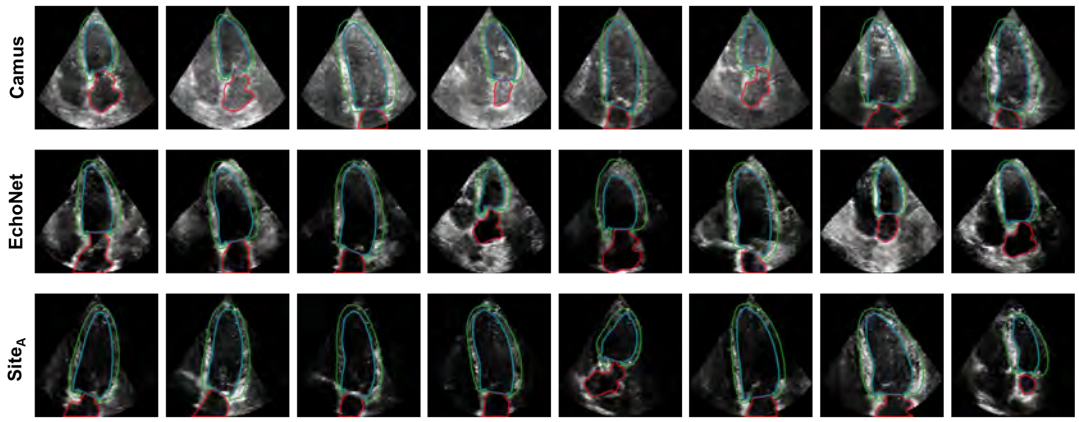


Fig. 5. 8 randomly selected synthetic images for each of the 3 synthetic datasets. Annotations are included, blue is LV_{endo}, green is LV_{epi}, and red is LA.

APPENDIX B
EXAMPLE SYNTHETIC IMAGES

Random example synthetic images from each of the 3 synthetic datasets are shown in Fig. 5. Note the difference in appearance of the different datasets which matches the difference seen in real images (see Appendix G for more example real images). While the images generally look realistic, the synthetic images can usually be identified because the noise pattern stays constant throughout the image. In real ultrasound the images are compressed at the top and stretched at the bottom because of the scan-conversion process, changing the noise characteristics throughout the image.

TABLE V
INTER-OBSERVER STUDY: MEDIAN RESULTS FROM A SECOND ANNOTATION ROUND ON 20 IMAGES FROM EACH DATASET. THE BOTTOM ROWS COMPARE ANNOTATIONS FROM O_1 AND O_2 FROM THE SECOND ANNOTATION ROUND.

Dataset	Version	Annotation Source		$\boldsymbol{D}$ (%) [$\boldsymbol{MAD}$]			$\boldsymbol{B}$ (%)			$\boldsymbol{S_p}$	$\boldsymbol{d_m}$
		Round 1	Round 2	LV_{endo}	LV_{epi}	LA	LV_{endo}	LV_{epi}	LA	LV_{endo}	
Camus	Real	[17]	O_1	78.7 [3.6]	90.6 [1.9]	85.5 [8.9]	42	17	3	0.83	10.0
			O_2	87.4 [4.4]	93.4 [1.7]	84.0 [6.9]	25	8	-5	0.83	5.7
	Synthetic	Models	O_1	83.8 [2.9]	88.6 [2.6]	78.0 [8.9]	30	16	26	0.80	6.7
			O_2	90.3 [2.3]	91.0 [2.3]	82.5 [9.3]	6	4	12	0.80	4.0
$Site_A$	Real	O_1	O_1	80.1 [5.1]	92.1 [2.2]	88.3 [3.5]	40	9	7	0.82	7.9
			O_2	88.9 [4.4]	93.3 [2.8]	90.2 [4.8]	21	8	-3	0.82	4.4
	Synthetic	Models	O_1	85.7 [4.1]	91.4 [2.5]	81.4 [4.9]	28	12	25	0.80	6.2
			O_2	91.1 [2.2]	92.5 [1.5]	83.3 [7.1]	11	4	15	0.80	3.9
EchoNet	Real	[16]	O_1	81.3 [5.4]	n/a	n/a	37	n/a	n/a	0.76	9.9
	Synthetic	Models	O_1	87.1 [2.3]	91.3 [1.7]	80.1 [5.6]	24	12	27	0.80	5.1
		Round 2	Round 2								
Camus & $Site_A$	Real	O_1	O_2	90.9 [3.2]	95.0 [2.0]	90.3 [4.5]	-17	-6	-10	0.85	4.3
	Synthetic	O_1	O_2	88.3 [4.1]	92.7 [3.4]	88.9 [3.2]	-21	-11	-10	0.84	4.9

LV_{endo} = left ventricle endocardium, LV_{epi} = left ventricle epicardium, LA = left atrium. $\boldsymbol{D}$ = Dice score, $\boldsymbol{B}$ = Bias percentage, $\boldsymbol{S_p}$ = simplicity, and $\boldsymbol{d_m}$ = mean average distance. $\boldsymbol{S_p}$ is listed for the dataset in the first column.

APPENDIX C
FULL INTER-OBSERVER RESULTS

Table V shows the full inter-observer results divided by dataset and observer. The bottom two rows show the comparison of the labels of O_1 and O_2 on the second round of labeling. The results show a significant difference between the first and second rounds of labeling by O_1. This is shown by the low Dice scores and high LV_{endo} biases in the intra-observer results on $Site_A$ as well as by the flip in biases between O_2 vs. O_1. In the first round O_2's LV_{endo} labels were significantly larger than O_1's (high positive bias) while in the second round they were significantly smaller. We estimate this difference comes from the differences in format of the second round of labeling (discussed in text) as well as the time difference (two months) in between the first and second rounds. It may also be due to other external factors.

APPENDIX D
ADDITIONAL SEGMENTATION RESULTS

A. Full results on A4C LV Segmentation Task

Quantitative results for testing the synthetically generated networks on the EchoNet, Camus, $Site_B$, and Pathological test sets for apical four chamber LV_{endo} segmentation in ED images only is shown in Table VI. The EchoNet results were already presented in the manuscript, but are shown again for comparison. Overall results show a similar pattern where the synthetic data approximately matches a network trained on a different real dataset. This varies by the test dataset with the synthetically trained network again performing well on the $Site_B$ data and Pathological data but worse on the Camus data.

B. Mean Results

The manuscript presents results using median and median absolute deviation because the results are not normally distributed, in which case the median is a better indicator of central tendency [18]. Mean and standard deviations for the results shown in Table I of the main manuscript are presented in Table VII for reference.

APPENDIX E
BREAKING DOWN OBSERVER BIAS

In this section we provide a more detailed analysis showing the most likely reason for LV_{endo} positioning differences (and corresponding differences in myocardial thickness) in the datasets is a bias between annotators rather than a difference in patient characteristics. Fig. 6 (O_1) and Fig. 7 (O_2) show violin plots for $\frac{LV_{endo}}{LV_{epi}}$ ratio on each of the different datasets O_1 and O_2 labeled in the inter-observer study. The results remain consistent across a given dataset. Fig. 8 shows the labeling ratio across different annotation rounds (i.e. between different annotators). In this case the ratio varies substantially between rounds. If the difference came from patient characteristics we would expect the ratios in Fig. 6 and Fig. 7 to vary between datasets. The converse indicates the primary difference in the ratio (and thus in LV_{endo} placement) instead comes from the labeler. These results also show that although O_1 was consistent within each labeling round there was a large bias between rounds as discussed above. This bias between labeling rounds was the reason the inter-observer results for O_1 were excluded from the main manuscript although results show the same pattern between real and synthetic data as O_2.

TABLE VI

TESTING ON REAL DATA: MEDIAN RESULTS FOR LV_{ENDO} SEGMENTATION NETWORKS TRAINED WITH SYNTHETIC AND REAL DATASETS AND EVALUATED ON REAL DATASETS OF APICAL FOUR CHAMBER IMAGES. RESULTS ARE ORDERED BY DICE SCORE. BOLD SHOWS THE BEST RESULT IN EACH SECTION.

Train	Test	D [MAD]	B (%)	S_p	d_m
EchoNet	EchoNet	**94.0 [1.6]**	**0**	0.85	**2.7**
Camus		89.6 [2.6]	-2	**0.87**	4.8
$Site_A$		87.7 [3.8]	14	0.86	5.6
Synth EchoNet		87.1 [3.7]	15	0.85	5.9
Camus	Camus	**94.8 [1.2]**	**0**	0.85	**2.9**
EchoNet		93.0 [1.5]	-7	0.84	3.1
$Site_A$		89.6 [2.8]	14	**0.86**	4.8
Synth Camus		88.3 [2.6]	18	0.84	4.7
Synth $Site_A$	$Site_B$	**90.2 [2.4]**†	**-17**	0.84	**5.6**
EchoNet		88.3 [2.9]†,‡	-23	0.84	6.2
$Site_A$		87.9 [4.0]‡	-23	**0.85**	7.1
Camus		83.5 [4.3]	-32	0.85	8.8
EchoNet	Pathological	**92.0 [2.3]**	-7	0.86	**3.2**
Synth $Site_A$		90.7 [2.6]†	**-1**	0.85	4.1
$Site_A$		90.3 [3.5]†	**1**	0.86	4.2
Camus		89.5 [2.9]†	-10	**0.87**	4.6

†, ‡: Rows in the same section marked with a matching symbol were not statistically different with a p-value < 0.05.

TABLE VII

MEAN METRICS COMPARING TRAINING WITH REAL DATASETS TO TRAINING WITH SYNTHETIC DATASETS. THE FIRST SECTION COMPARES INTER-OBSERVER RESULTS FOR O_2 ON REAL AND SYNTHETIC DATA. THE NEXT SECTION SHOWS NETWORKS TRAINED ON REAL AND SYNTHETIC DATA FOR LV_{ENDO} SEGMENTATION IN A4C ED IMAGES AND TESTED ON ECHONET. THE FINAL SECTION COMPARES NETWORKS TRAINED ON REAL AND SYNTHETIC DATA FOR ALL ANNOTATIONS/VIEWS/PHASES. ALL DICE RESULTS ARE STATISTICALLY DIFFERENT WITH A P-VALUE < 0.05 (COMPUTED WITH A WILCOXON SIGNED-RANK TEST). RESULTS ARE ORDERED BY DICE SCORE AND BOLD SHOWS THE BEST RESULT IN EACH SECTION.

Task	Training Data / *OR inter-observer comparison*	Testing Data	D (%) [STD] LV_{endo}	D (%) [STD] LV_{epi}	D (%) [STD] LA	B (%) LV_{endo}	B (%) LV_{epi}	B (%) LA	S_p	d_m LV_{endo}
Inter-Observer	*O_2 vs. Camus/$Site_A$*		86.9 [5.6]	**92.9 [2.8]**	**85.7 [9.6]**	24	7	**-4**	**0.85**	5.4
	O_2 vs. proposed pipeline		**90.2 [3.3]**	91.5 [3.0]	80.1 [14.5]	**8**	**2**	14	0.84	**4.1**
LV_{endo} in A4C ED	EchoNet (base)	EchoNet	**93.1 [3.5]**	n/a	n/a	**1**	n/a	n/a	0.85	**3.1**
	Camus		87.8 [7.4]			-2			**0.86**	6.2
	$Site_A$		86.1 [7.2]			15			**0.86**	6.2
	Synth EchoNet		85.6 [7.0]			16			0.84	6.7
All	Camus (base)	Camus	**92.5 [4.5]**	**95.2 [1.9]**	**88.3 [10.2]**	**2**	**-1**	**1**	0.83	**3.0**
	Site $_A$		87.5 [6.1]	91.0 [4.3]	79.8 [15.8]	10	-3	18	**0.84**	5.2
	Synth Camus		79.4 [9.9]	88.6 [5.4]	71.3 [23.7]	35	7	-4	0.82	8.4
	Synth $Site_A$	$Site_B$	**85.3 [8.8]**	88.1 [7.2]	73.7 [22.0]	**-4**	**3**	-11	0.82	**6.9**
	$Site_A$ (base)		85.1 [6.3]	**91.3 [4.6]**	**84.5 [12.5]**	-26	-9	16	**0.84**	7.9
	Camus		80.8 [12.0]	**91.3** [3.7]	83.0 [11.3]	-35	-9	**-4**	0.83	10
	Camus	Pathological	**88.2 [10.5]**	**91.1 [6.1]**	**88.6 [13.0]**	-8	**0**	**2**	**0.86**	**4.7**
	$Site_A$ (base)		86.5 [12.7]	89.6 [8.9]	83.8 [13.7]	**0**	9	14	0.85	5.3
	Synth $Site_A$		81.5 [17.8]	83.1 [12.7]	74.3 [23.5]	18	15	-20	0.82	6.5

LV_{endo} = left ventricle endocardium, LV_{epi} = left ventricle epicardium, LA = left atrium, A4C = apical four chamber, A2C = apical two chamber, ED = end diastole, ES = end systole. All refers to all annotations (LV_{endo}, LV_{epi}, and LA), views (A4C and A2C), and phases (ED and ES). D = Dice score, MAD = median absolute deviation, B = Bias percentage, S_p = simplicity, and d_m = mean average distance. For inter-observer S_p is listed for the second round annotations and B is calculated as $O_2 - \text{Original}$.

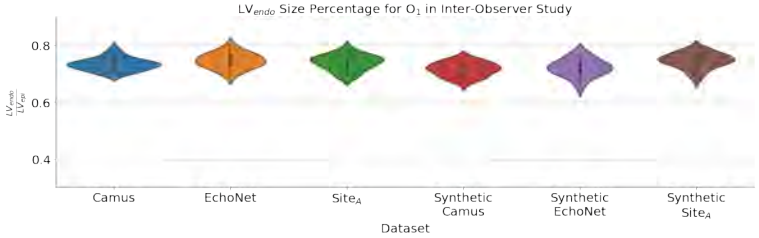


Fig. 6. LV_{endo} ratio for all datasets labeled by O_1 in the second round of labeling. The ratio stays consistent across the datasets.

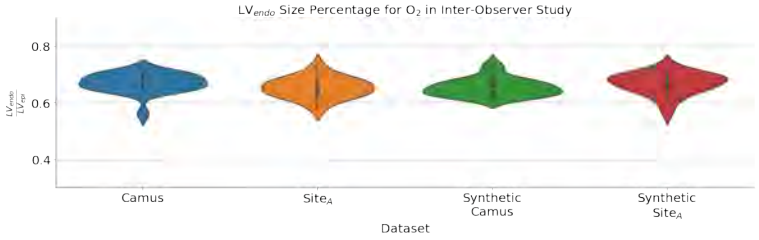


Fig. 7. LV$_{\text{endo}}$ ratio for all datasets labeled by O_2 in the second round of labeling. The ratio stays consistent across the datasets.

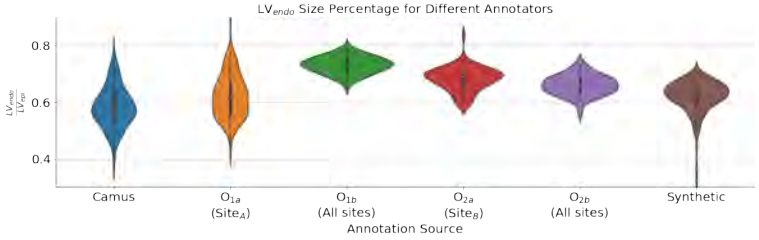


Fig. 8. LV$_{\text{endo}}$ ratio for each labeling round. The ratio varies substantially between rounds.

A qualitative evaluation of this phenomena is shown in Fig. 9. Example annotations are shown for each dataset in quantiles from 0% to 50% sorted by the $\frac{LV_{endo}}{LV_{epi}}$ ratio. The images show the much thicker myocardia in Camus and Site$_{\text{A}}$ even though the underlying tissue does not appear significantly different in many cases. Pseudo images are presented for the Synthetic dataset.

APPENDIX F
EFFECT OF PSEUDO TRANSFORM

To evaluate the effect of the pseudo transforms (Appendix A-D) several experiments were conducted. Segmentation networks were trained using images from the model after various versions of the pseudo and CycleGAN transformations as shown in Fig. 10. Networks were trained with images directly from the model ("slice images"), images with the pseudo appearance transforms ("pseudo images"), slice images transformed with a CycleGAN ("synthetic slice images"), and pseudo images transformed with a CycleGAN ("synthetic images"). Qualitatively, the synthetic slice images do not appear as realistic as the synthetic images. Many of the hard edges are maintained from the slice image making it easy to tell that the images are fake.

Segmentation results are shown in Table VIII. Results indicate that the synthetic images achieve the best results as shown in the manuscript. The slice image dataset achieves poor results as the network trained on this dataset was not able to translate learned features across the large texture differences. The networks trained on the synthetic slice dataset and the pseudo dataset were approximately the same and slightly worse than the network trained on the synthetic images.

TABLE VIII

COMPARING TRANSFORMATIONS: MEDIAN RESULTS FOR LV$_{\text{ENDO}}$ SEGMENTATION NETWORKS TRAINED ON THE DATASET VARIATIONS SHOWN IN FIG. 10. RESULTS ARE ORDERED BY DICE SCORE ($\boldsymbol{D}$). ALL DICE RESULTS EXCEPT THOSE MARKED (†) ARE STATISTICALLY DIFFERENT WITH A P-VALUE $<$ 0.05 (COMPUTED WITH A WILCOXON SIGNED-RANK TEST). BOLD SHOWS THE BEST RESULT FOR EACH METRIC.

Train	Test	$\boldsymbol{D}$ (%) [$\boldsymbol{MAD}$]	$\boldsymbol{B}$ (%)	$\boldsymbol{S_p}$	$\boldsymbol{d_m}$
Synthetic EchoNet	EchoNet	**87.1 [3.7]**	**15**	**0.85**	**5.9**
Synthetic Slice EchoNet		84.6 [4.9]†	**15**	0.84	6.8
Pseudo EchoNet		84.1 [4.5]†	18	0.84	6.7
Slice EchoNet		65.3 [11.8]	**15**	0.80	12.5

†: Not statistically different with a P-value $<$ 0.05.

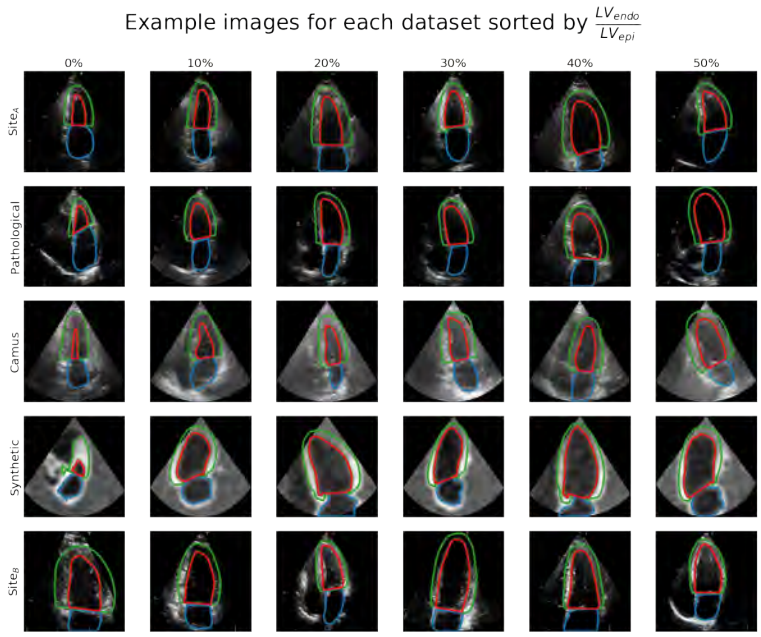


Fig. 9. Example annotation results for each dataset sorted by $\frac{LV_{endo}}{LV_{epi}}$ ratio within each dataset. Each column shows a quantile of the $\frac{LV_{endo}}{LV_{epi}}$ ratio from 0% to 50%. The 0% quantile represents the image with the lowest $\frac{LV_{endo}}{LV_{epi}}$ ratio (largest myocardium). As shown, the Camus dataset typically has much larger myocardia.

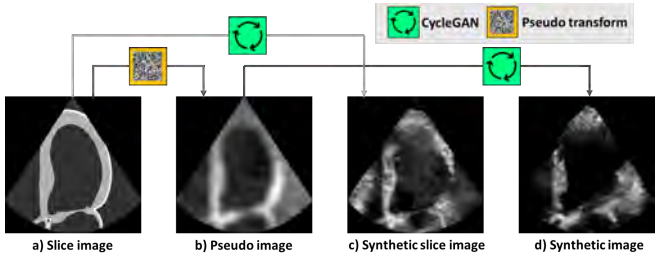


Fig. 10. **Four different image types were tested to determine the effects of the pseudo and CycleGAN transforms:** a) Slice images which are slices extracted from the model with only a cone mask applied, b) pseudo images which have undergone the pseudo transformations described in the manuscript, c) synthetic slice images which are the slice images passed through a CycleGAN, and d) synthetic images which are the pseudo images passed through the CycleGAN.

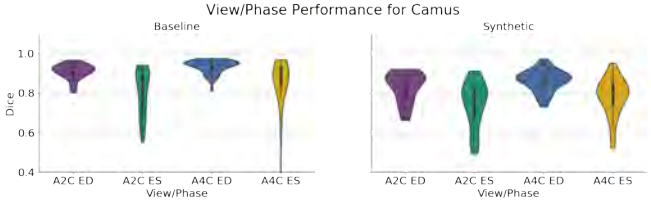


Fig. 11. **View/phase breakdown for Camus:** Performance split by echo views and phases for the Camus test set. The network trained on synthetic data shows much worse results for the end systole (ES) phase than end diastole (ED). However, this is also matched by the baseline network indicating that the ES images are more difficult in general for this dataset. This finding matches [14]. A2C = apical two chamber, A4C = apical four chamber, ED = end diastole, ES = end systole.

Fig. 12. **View/phase breakdown for Site$_B$:** Performance split by echo views and phases for the Site$_B$ test set. In this case the results are approximately consistent across views and phases for both the baseline and synthetic experiments. A2C = apical two chamber, A4C = apical four chamber, ED = end diastole, ES = end systole.

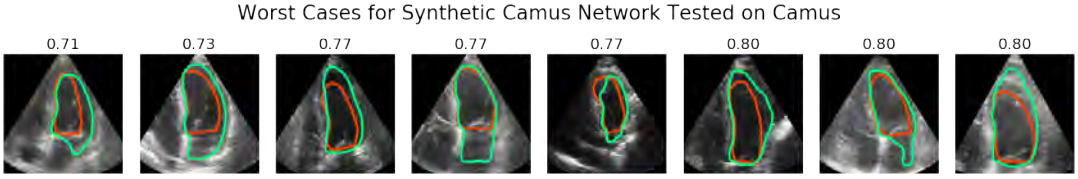


Fig. 13. The 8 worst images (sorted by Dice score) for the network trained on synthetic Camus data and tested on real Camus data. Titles show Dice score. Red shows the label while green shows the network output.

Appendix G
Failure Analysis

A. Breakdown by phase and view

Fig. 11 and Fig. 12 show the breakdown in performance for different echo views and phases for the Camus and Site$_B$ test respectively. Results are varied between the two sets with Camus showing a clear difference in performance between ED and ES for both the synthetic model and the baseline model. While we initially hypothesized the difference on synthetic data was because the model did not include ES images, the equivalent result on real images indicates the difference is likely an implicit bias with ES images in the dataset. Site$_B$ on the other hand shows approximately equivalent performance across the phases and views for both although ES is slightly higher than ED for the baselinereal-data model. This discrepancy could come from several sources, with the most likely being the selection criteria for determining the phase.

In both cases the synthetic four chamber images perform slightly better than the synthetic two chamber images. This is likely because the extraction process was optimized for four chamber images, meaning those images are of better quality.

B. Worst Cases for Segmentation

We focused this section on LV$_{endo}$ segmentation only to simplify the analysis, but the reasons for failure are similar when extending to additional annotations. Fig. 13, Fig. 14, Fig. 15, and Fig. 16 show the worst case images for each of the networks trained on the synthetic versions of Camus, Site$_A$, and EchoNet and tested on the corresponding real datasets. We analyzed these images to identify the primary sources of failure for the network:

1) In many cases the network is unable to properly detect the mitral valve and includes part of the LA in the LV. Because the valve is not visible in CT images it is modeled as a flat slab in the models. This means the synthetic images do not include the variety of valves that are included in the real images. Including accurate valve representations is a target for future work.

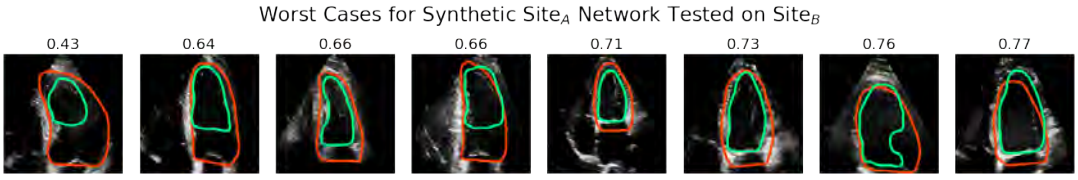


Fig. 14. The 8 worst images (sorted by Dice score) for the network trained on synthetic Site$_A$ data and tested on Site$_B$. Titles show Dice score. Red shows the label while green shows the network output.

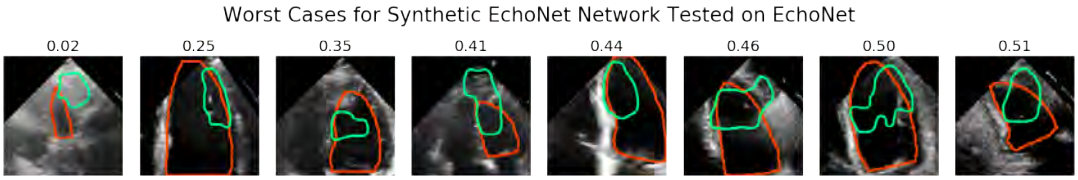


Fig. 15. The 8 worst images (sorted by Dice score) for the network trained on synthetic EchoNet data and tested on real EchoNet data. Titles show Dice score. Red shows the label while green shows the network output.

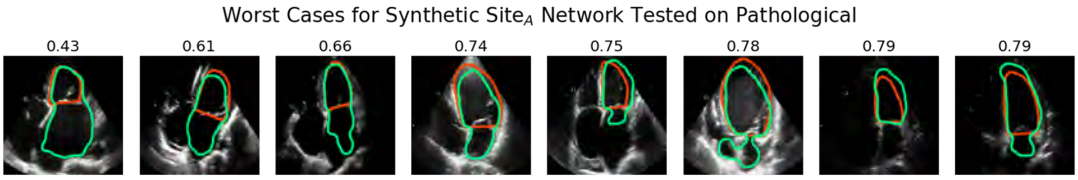


Fig. 16. The 8 worst images (sorted by Dice score) for the network trained on synthetic Site_A data and tested on the Pathological dataset. Titles show Dice score. Red shows the label while green shows the network output.

2) The network fails for some LV-focused images where the mitral valve is positioned at the bottom edge of the image. This is likely due to a difference in the pseudo dataset construction as all pseudo images were cropped such that there was at least a 10% border between the mitral valve and the bottom edge of the image. In other cases the walls were cropped out by the cone which was also avoided in the pseudo images.
3) In some cases the walls were dropped out of the image due to the high gain settings. With only a single frame it can be nearly impossible to detect the wall so giving additional frames to the segmentation network would be the best solution to address this. An approach such as random erasing augmentation [19] may also help resolve these cases.
4) In addition, the network appeared to struggle in cases where the image was wider or where the LV was slightly rotated. These factors of variability could be included in the pseudo image generation process.
5) The network also struggled in some cases where the image was wider or where the LV was slightly rotated.
6) In the pathological dataset the network struggled in some cases with high left atrial dilation where the left atrium is significantly larger than the left ventricle.

These items played a varying role in failure cases for the different datasets and illustrate the challenges of deploying an algorithm into the wild.

While we have presented the worst cases for the synthetic segmentation networks above, in general the networks perform very well. Fig. 17 shows a result from each quantile of the Dice score for a network trained on the baseline and synthetic versions of each dataset. As shown, the majority of the images have good segmentation results. For example, although there were many bad failures for EchoNet (Fig. 15), the test set was much larger in this case, providing more opportunities for failure. In general, the network performed well.

Appendix H
Pre-training With Synthetic Data

The primary goal of the proposed pipeline is to generate artificial data. However, real data can also be used to fine-tune the results from the synthetic data model. This may be especially useful in cases where only a small number of real images are available. Fig. 18 shows results of fine-tuning a model pretrained with synthetic data on varying amounts of real images compared to training from scratch with real images. Results show that for a small number of images, pre-training with synthetic data yields significantly higher Dice scores. This is particularly true for the LA which the model initially struggles to segment accurately (matching the trend reported in [17]). As more real images were added there are diminishing returns from pre-training with synthetic data, but in almost all cases the pre-trained model has better performance. Note that in the case of LV_endo segmentation trained on Site_A, the model trained on synthetic data initially outperformed the model trained on real data when tested on Site_B. In this case, as more real data images are added there is a small decrease in performance in LV_endo Dice scores. Fig. 19 shows the performance improvements from pretraining on the synthetic data on the Camus and Site_A datasets for each segmentation target.

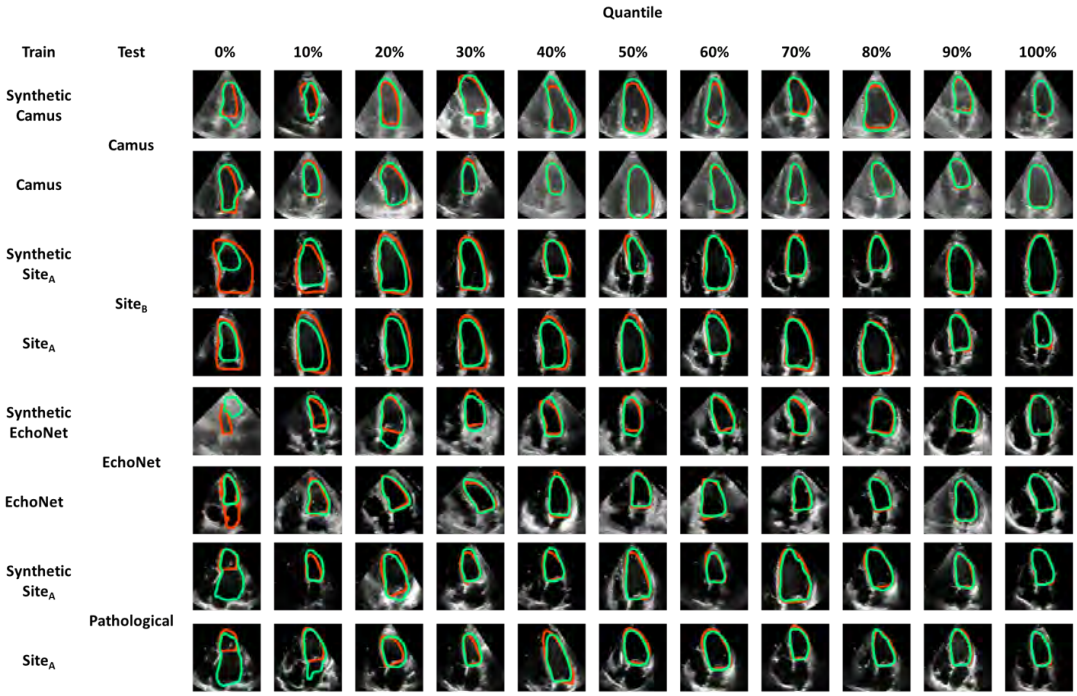


Fig. 17. An example image from each quantile of the Dice results for LV_{endo} segmentation in apical four chamber images. Each row shows results for a network trained on a different dataset. Red shows the label while green shows the network output.

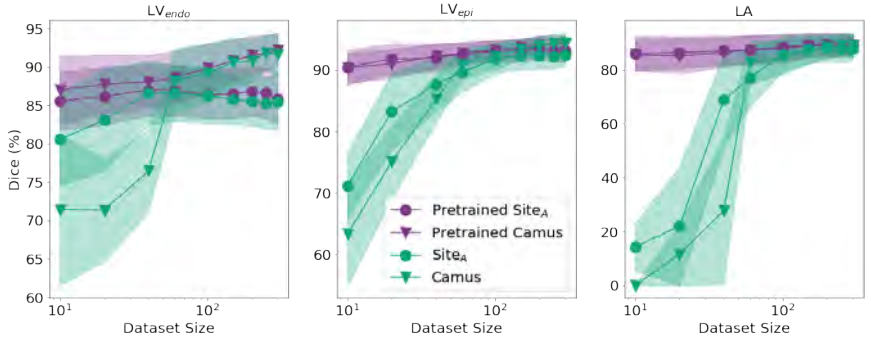


Fig. 18. Pretraining with synthetic data can improve results if only limited data is available. Median LV_{endo}, LV_{epi}, and LA Dice scores for networks pretrained on synthetic data vs. networks trained from scratch. Error bars show median absolute deviation. The networks trained on Camus were tested on Camus and the networks trained on $Site_A$ were tested on $Site_B$. X-axis is log scale.

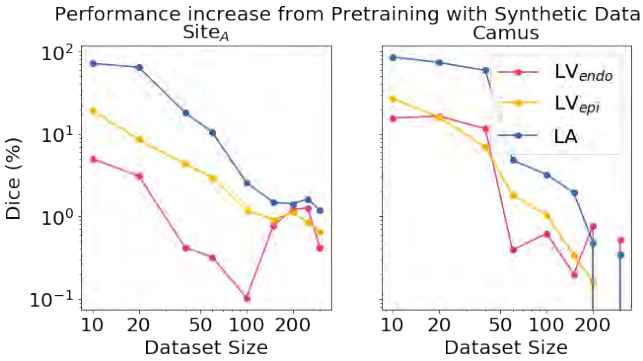


Fig. 19. The performance improvement from pretraining with synthetic data for Site$_A$ and Camus. Both axes are shown on a log scale.

REFERENCES

[1] C. Rodero, M. Strocchi, M. Marciniak, J. Whitaker, D. O. Neill, K. Gillette, C. Augustin, G. Plank, E. Vigmond, P. Lamata, and S. A. Niederer, "Anatomical changes influences mechanics, electrophysiologyand haemodynamics in a complementary and localised wayin the healthy adult human heart," *PLOS Comput. Biol. (under Rev.)*.

[2] CIBC, "Seg3D: Volumetric Image Segmentation and Visualization," 2016. [Online]. Available: http://www.sci.utah.edu/software/seg3d.html

[3] M. Strocchi, C. M. Augustin, M. A. Gsell, E. Karabelas, A. Neic, K. Gillette, O. Razeghi, A. J. Prassl, E. J. Vigmond, E. J. Vigmond, J. M. Behar, J. M. Behar, J. Gould, J. Gould, B. Sidhu, B. Sidhu, C. A. Rinaldi, C. A. Rinaldi, M. J. Bishop, G. Plank, and S. A. Niederer, "A publicly available virtual cohort of fourchamber heart meshes for cardiac electromechanics simulations," *PLoS One*, vol. 15, no. 6 June, pp. 1–26, 2020. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0235145

[4] M. Vaillant and J. Glaunès, "Surface Matching via Currents," in *Inf. Process. Med. Imaging*, G. E. Christensen and M. Sonka, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 381–392.

[5] T. Mansi, I. Voigt, B. Leonardi, X. Pennec, S. Durrleman, M. Sermesant, H. Delingette, A. M. Taylor, Y. Boudjemline, G. Pongiglione, and Others, "A statistical model for quantification and prediction of cardiac remodelling: Application to tetralogy of fallot," *IEEE Trans. Med. Imaging*, vol. 30, no. 9, pp. 1605–1616, 2011.

[6] S. Durrleman, X. Pennec, A. Trouvé, and N. Ayache, "Statistical models of sets of curves and surfaces based on currents," *Med. Image Anal.*, vol. 13, no. 5, pp. 793–808, 2009.

[7] W. Schroeder, K. Martin, and B. Lorensen, *The Visualization Toolkit*, 4th ed. Kitware, 2006.

[8] S. Durrleman, M. Prastawa, N. Charon, J. R. Korenberg, S. Joshi, G. Gerig, and A. Trouvé, "Morphometry of anatomical shape complexes with dense deformations and sparse parameters," *Neuroimage*, vol. 101, pp. 35–49, 2014. [Online]. Available: www.deformetrica.org.

[9] C. Geuzaine and J.-F. Remacle, "Gmsh: A 3-D finite element mesh generator with built-in pre-and post-processing facilities," *Int. J. Numer. Methods Eng.*, vol. 79, no. 11, pp. 1309–1331, 2009.

[10] J. Ahrens, B. Geveci, and C. Law, *ParaView: An End-User Tool for Large Data Visualization,*. Elsevier, 2005.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Int. Conf. Med. image Comput. Comput. Interv.*, 2015.

[12] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance Normalization: The Missing Ingredient for Fast Stylization," no. 2016, 2016. [Online]. Available: http://arxiv.org/abs/1607.08022

[13] F. Jay, J.-P. Renou, O. Voinnet, and L. Navarro, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks Jun-Yan," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 183–202.

[14] S. Leclerc, E. Smistad, A. Ostvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, C. Lartizien, L. Lovstakken, and O. Bernard, "Deep Learning Segmentation in 2D echocardiography using the CAMUS dataset : Automatic Assessment of the Anatomical Shape Validity," in *Int. Conf. Med. Imaging with Deep Learn. – Ext. Abstr. Track*, 2019.

[15] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–15, 2015.

[16] D. Ouyang, B. He, A. Ghorbani, C. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, J. Euan A Ashley, and a. Y. Zou, "Interpretable AI for beat-to-beat cardiac function assessment," *Nature*, 2020. [Online]. Available: https://doi.org/10.1101/19012419

[17] S. Leclerc, E. Smistad, J. Pedrosa, A. Ostvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P. M. Jodoin, T. Grenier, C. Lartizien, J. Dhooge, L. Lovstakken, and O. Bernard, "Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography," *IEEE Trans. Med. Imaging*, vol. 38, no. 9, pp. 2198–2210, 2019.

[18] M. Pagano and K. Gauvreau, *Principles of Biostatistics*, 2nd ed. CRC Press, 2018.

[19] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random Erasing Data Augmentation," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 07, pp. 13 001–13 008, 2020.

Appendices

Appendix A

The "Digital Twin" to enable the vision of precision cardiology

Jorge Corral-Acero, Francesca Margara, Maciej Marciniak, Cristobal Rodero, Filip Loncaric, Yingjing Feng, Andrew Gilbert, Joao F Fernandes, Hassaan A Bukhari, Ali Wajdan, Manuel Villegas Martinez, Mariana Sousa Santos, Mehrdad Shamohammdi, Hongxing Luo, Philip Westphal, Paul Leeson, Paolo DiAchille, Viatcheslav Gurev, Manuel Mayr, Liesbet Geris, Pras Pathmanathan, Tina Morrison, Richard Cornelussen Frits Prinzen, Tammo Delhaas, Ada Doltra, Marta Sitges, Edward J Vigmond, Ernesto Zacur, Vicente Grau, Blanca Rodriguez, Espen W Remme, Steven Niederer, Peter Mortier, Kristin McLeod, Mark Potse, Esther Pueyo, Alfonso Bueno-Orovio, Pablo Lamata

Published in *European Heart Journal*, 2020, DOI: 10.1093/eurheartj/ehaa159.

ESC European Society of Cardiology

European Heart Journal (2020) **41**, 4556–4564
doi:10.1093/eurheartj/ehaa159

STATE OF THE ART REVIEW
Disease management

The 'Digital Twin' to enable the vision of precision cardiology

Jorge Corral-Acero[1], Francesca Margara [2], Maciej Marciniak [3], Cristobal Rodero [3], Filip Loncaric [4], Yingjing Feng [5,6], Andrew Gilbert [7], Joao F. Fernandes [3], Hassaan A. Bukhari[6,8], Ali Wajdan[9], Manuel Villegas Martinez[9], Mariana Sousa Santos[10], Mehrdad Shamohammdi[11], Hongxing Luo [11], Philip Westphal[12], Paul Leeson [13], Paolo DiAchille [14], Viatcheslav Gurev [14], Manuel Mayr [15], Liesbet Geris [16], Pras Pathmanathan[17], Tina Morrison[17], Richard Cornelussen[12], Frits Prinzen[11], Tammo Delhaas [11], Ada Doltra [4], Marta Sitges [4,18], Edward J. Vigmond [5,6], Ernesto Zacur [1], Vicente Grau [1], Blanca Rodriguez [2], Espen W. Remme[9], Steven Niederer [3], Peter Mortier[10], Kristin McLeod [7], Mark Potse [5,6,19], Esther Pueyo [8,20], Alfonso Bueno-Orovio [2], and Pablo Lamata [3]*

[1]Department of Engineering Science, University of Oxford, Oxford, UK; [2]Department of Computer Science, British Heart Foundation Centre of Research Excellence, University of Oxford, Oxford, UK; [3]Department of Biomedical Engineering, Division of Imaging Sciences and Biomedical Engineering, King's College London, London, UK; [4]Institut Clínic Cardiovascular, Hospital Clínic, Universitat de Barcelona, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain; [5]IHU Liryc, Electrophysiology and Heart Modeling Institute, fondation Bordeaux Université, Pessac-Bordeaux F-33600, France; [6]IMB, UMR 5251, University of Bordeaux, Talence F-33400, France; [7]GE Vingmed Ultrasound AS, Horton, Norway; [8]Aragón Institute of Engineering Research, Universidad de Zaragoza, IIS Aragón, Zaragoza, Spain; [9]The Intervention Centre, Oslo University Hospital, Rikshospitalet, Oslo, Norway; [10]FEops NV, Ghent, Belgium; [11]CARIM School for Cardiovascular Diseases, Maastricht University, Maastricht, The Netherlands; [12]Medtronic PLC, Bakken Research Center, Maastricht, the Netherlands; [13]Radcliffe Department of Medicine, Division of Cardiovascular Medicine, Oxford Cardiovascular Clinical Research Facility, John Radcliffe Hospital, University of Oxford, Oxford, UK; [14]Healthcare and Life Sciences Research, IBM T.J. Watson Research Center, Yorktown Heights, NY, USA; [15]King's British Heart Foundation Centre, King's College London, London, UK; [16]Virtual Physiological Human Institute, Leuven, Belgium; [17]Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, MD, USA; [18]CIBERCV, Instituto de Salud Carlos III, (CB16/11/00354), CERCA Programme/Generalitat de, Catalunya, Spain; [19]Inria Bordeaux Sud-Ouest, CARMEN team, Talence F-33400, France; and [20]CIBER in Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Madrid, Spain

Received 6 September 2019; revised 29 November 2019; editorial decision 16 February 2020; accepted 24 February 2020; online publish-ahead-of-print 4 March 2020

Providing therapies tailored to each patient is the vision of precision medicine, enabled by the increasing ability to capture extensive data about individual patients. In this position paper, we argue that the second enabling pillar towards this vision is the increasing power of computers and algorithms to learn, reason, and build the 'digital twin' of a patient. Computational models are boosting the capacity to draw diagnosis and prognosis, and future treatments will be tailored not only to current health status and data, but also to an accurate projection of the pathways to restore health by model predictions. The early steps of the digital twin in the area of cardiovascular medicine are reviewed in this article, together with a discussion of the challenges and opportunities ahead. We emphasize the synergies between mechanistic and statistical models in accelerating cardiovascular research and enabling the vision of precision medicine.

Keywords Precision medicine • Digital twin • Computational modelling • Artificial intelligence

* Corresponding author. Tel: (+44) 20 784 89563, Email: pablo.lamata@kcl.ac.uk

Introduction

Providing therapies that are tailored to each patient, and that maximize the efficacy and efficiency of our healthcare system, is the broad goal of precision medicine. The main shift from current clinical practice is to take inter-individual variability into greater account. This exciting vision has been championed by the -omics revolution, i.e., the increasing ability to capture extensive data about the pathophysiology of the patient.[1,2] This -omics approach has already delivered great achievements, especially in the management of specific cancer conditions.[3] Nevertheless, the initial conception of precision medicine has already been criticized for being too centred in genomics and failing to address challenges of clinical management.[4] The concept is thus gradually widening, shifting from the original gene-centric perspective to the wide spectrum of lifestyle, environment, and biology data.[5,6]

In this context, we argue that the definition of optimal therapy options requires a mechanistic understanding that links all levels from genetic and molecular traces to the pathophysiology, lifestyle and environment of the patient, and back. Precision medicine requires, not only better and more detailed data, but also the increasing ability of computers to analyse, integrate, and exploit these data, and to construct the 'digital twin' of a patient. In health care, the 'digital twin' denotes the vision of a comprehensive, virtual tool that integrates coherently and dynamically the clinical data acquired over time for an individual using mechanistic and statistical models.[7] This borrows but expands the concept of 'digital twin' used in engineering industries, where *in silico* representations of a physical system, such as an engine or a wind farm, are used to optimize design or control processes, with a real-time connection between the physical system and the model.[8]

This position paper claims that precision cardiology will be delivered in a synergetic fashion that combines induction, by using statistical models learnt from data, and deduction, through mechanistic modelling and simulation integrating multiscale knowledge and data.[9] These are the two pillars of the digital twin (*Figure 1*). We review the state of the art of the interplay between such models that supports this vision, considering that there are already excellent independent review papers in the fields of statistical[14–16] and mechanistic[17,18] models for cardiovascular medicine.

Mechanistic models encapsulate our knowledge of physiology and the fundamental laws of physics and chemistry. They provide a framework to integrate and augment experimental and clinical data, enabling the identification of mechanisms and/or the prediction of outcomes, even under unseen scenarios without the need for retraining.[19] Examples of such mechanistic models are the bidomain equations for cardiac electrophysiology[20] or the Navier–Stokes equations for coronary blood flow.[21] In a complementary manner, statistical models encapsulate the knowledge and relations induced from the data. They allow the extraction and optimal combination of individualized biomarkers with mathematical rules. Examples of statistical models applied to computational cardiology are random forests for assessment of heart failure severity[22] or Gaussian processes to capture heart rate variability.[23]

There are clinical needs that can be solved with a single modelling approach. But both mechanistic and statistical models have limitations that can be addressed by combining them. Mechanistic models are constrained by their premises (assumptions and principles), while statistical models are constrained by the observations available (the amount and diversity of data). A mechanistic model may be a good choice when a good understanding of the system is available. A statistical model, on the other hand, can serve to find predictive relations even when the underlying mechanisms are poorly understood or are too complex to be modelled mechanistically. The rest of the article describes the synergies between mechanistic and statistic models (see *Figure 2* for an overview), motivated by actual clinical problems and needs, with specific representative components of the digital twin. Supplementary material online reviews the model synergies for exploiting and integrating clinical data.

Mechanistic and statistical model synergy for improving clinical decisions

Technical, ethical, and financial constraints limit the data acquisition needed to assist clinical decision-making.[14,15] Synergy between mechanistic and statistical models has shown value in aiding diagnosis, treatment, and prognosis evaluation. A fully developed digital twin will combine population and individual representations to optimally inform clinical decisions (*Figure 3*).

Model synergy in aiding diagnosis

Models can pinpoint the most valuable piece of diagnostic data. An example is the simulation study that revealed that fibrosis and other pulmonary vein properties may better characterize susceptibility to atrial fibrillation.[24] Models can also reliably infer biomarkers that cannot be directly measured or that require invasive procedures. For instance, the combination of cardiovascular imaging and computational fluid dynamics enables non-invasive characterizations of flow fields and the calculation of diagnostic metrics in the domains of coronary artery disease, aortic aneurysm, aortic dissection, valve prostheses, and stent design.[25–29]

The key to guide diagnosis is the personalization of a mechanistic model to the actual health status of the patient as captured in available clinical data. In this personalization process, statistical models enable robust and reproducible analysis of clinical data and infer missing parameters. An example of this synergy is the assessment of left ventricular myocardial stiffness and decaying diastolic active tension by fitting mechanical models to pressure data and images during diastole.[30,31] Another example is the non-invasive computation of pressure drops in flow obstructions,[32,33] such as aortic stenosis or aortic coarctation, which has been proven more accurate than methods recommended in clinical guidelines.[34] Models have also been used to derive fractional flow reserve from computed tomography (CT) to non-invasively identify ischaemia in patients with suspected coronary artery disease, avoiding invasive catheterized procedures.[29,35–37]

Some diagnostic medical devices based on personalized mechanistic models have already reached their industrial translation and clinical adoption. HeartFlow FFR_{CT} Analysis (HeartFlow, USA) and CardioInsight (Medtronic, USA) use patient-specific mechanistic

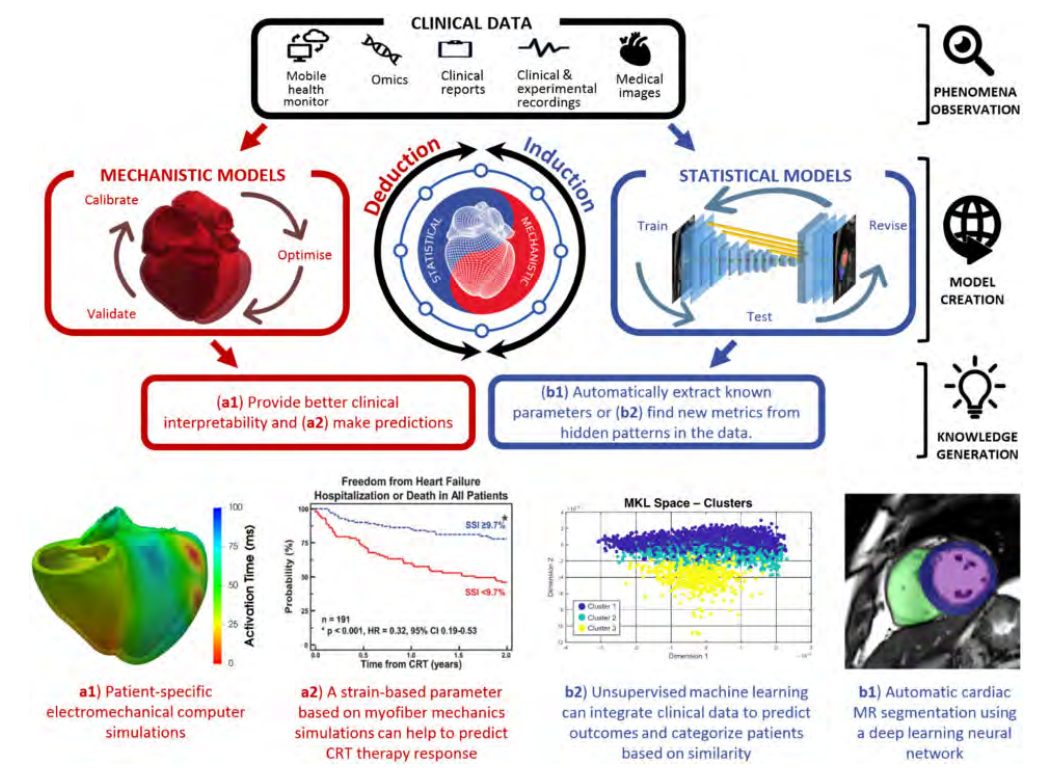


Figure 1 The two pillars of the digital twin, mechanistic and statistical models, illustrating its construction and four examples of use: **a1**,[10] **a2**,[11] **b1**,[12] **b2**.[13]

models to non-invasively calculate clinically relevant diagnostic indexes and have received clearance from the USA Food and Drug Administration (FDA).[38] HeartFlow predicts fractional flow reserve by means of a personalized 3D model of blood flow in the coronary arteries.[36] In the CardioInsight mapping system, the electrical activity on the heart surface is recovered from body surface potentials using a personalized model of the patient's heart and torso.[39]

Model synergy in guiding treatments

A digital twin may indicate whether a medical device or pharmaceutical treatment is appropriate for a patient by simulating device response or dosage effects before a specific therapy is selected.

The benefits of cardiac resynchronization therapy (CRT) have been demonstrated in patients with prolonged QRS duration. However, uncertainty remains in patients with more intermediate electrocardiogram (ECG) criteria.[40] To guide decision-making in this 'grey zone', approaches using mechanistic modelling have investigated the role of different aetiologies of mechanical discoordination in CRT response.[10] For example, a novel radial strain-based metric was defined based on simulations of the human heart and circulation to differentiate patterns of mechanical discoordination, suggesting that the response to CRT could be predicted from the presence of non-electrical substrates.[11] Statistical methods were used to verify these findings in a clinical cohort, and the novel index remained useful in predicting response in the clinical 'grey zone', creating the opportunity to improve patient selection in the group with intermediate ECG.

Another example is the improvements in ablation guidance of infarct-related ventricular tachycardia, where the accurate identification of patient-specific optimal targets is provided before the clinical procedure.[41] Mechanistic models can propose novel electro-anatomical mapping indices to locate critical sites of re-entry formation in scar-related arrhythmias, aid acquisition and quantitative interpretation of electrophysiological data, and optimize future clinical use.[42]

The industrial translation and clinical adoption of models for guiding treatment are exemplified by the optimal planning of valve prosthesis with the HEARTguide™ platform (FEops nv, Belgium), or by the platform to guide ventricular tachycardia ablations (inHeart, France).

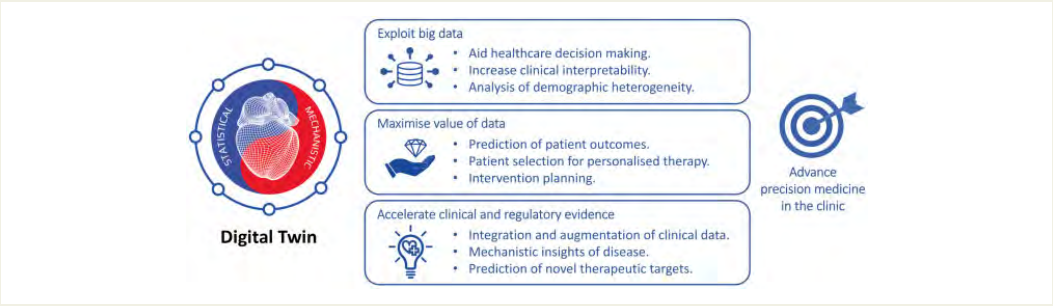


Figure 2 Conceptual summary of the main benefits of digital twin technologies.

Model synergy in evaluating prognosis

While statistical modelling allows for categorizing patients based on the probability of various outcomes, mechanistic modelling provides more insights to support or reject the categorization.

For example, model synergies represent an exciting approach to interpret structure–function relationships and improve risk prediction in inherited disease conditions, such as hypertrophic cardiomyopathy (HCM). Relationships among specific ECG changes, ventricle morphologies, and sudden cardiac death have been inferred from observations.[43,44] However, the complex process of translating underlying heterogeneous substrates in HCM to ECG findings is still poorly understood, and there exists a 'grey zone' of clinical decision-making in the low-risk patient subgroups, specifically when deciding on restriction of involvement in professional sports.[45] In this context, by using methods of statistical inference and mathematical modelling (see *Figure 4*), HCM patients were categorized into phenogroups based on ECG biomarkers extracted from 24-h ECG recordings,[48] and the aetiology of each ECG phenogroup linked with different underlying substrates, suggesting ion-channel and conduction system abnormalities.[46,47] The results directly highlighted the potential of personalized anti-arrhythmic approaches in the treatment of HCM patients, and addressed the low-risk patients, showing that a normal ECG might indeed be the discriminatory factor signalling minimal ionic remodelling, fibrosis, disarray, and ischaemia in these 'grey zone' patients.

Models have also been used in the prediction of arrhythmic events in post-myocardial infarction, outperforming existing clinical metrics including ejection fraction.[49] When the amount of data is not sufficient to inform state-of-the-art machine learning methods, statistical methods can still prove useful. An example is the use of principal component analysis to account for right ventricular motion in predicting survival in pulmonary hypertension,[50] or to identify signatures of anatomical remodelling that predict a patient's prognosis following CRT implantation.[51]

While statistical models allow predictions, mechanistic models provide the underlying explanations. Understanding the actual meaning of the selected features improves the plausibility of findings and increases their credibility. For both approaches, quantifying uncertainty of prediction can help identify cases that may require further review, while building trust in cases where models are shown to be robust.[52,53]

Mechanistic and statistical model synergy to accelerate evidence generation

While digital twin technologies in cardiology show promising research results, only a small number of models have reached clinical translation. The difficulties encountered include the need to increase validation, lack of clinical interpretability, and potentially obscure model failures.[54] Therefore, solid evidence for the generalization of preliminary findings and efficient testing strategies are needed. Even when these barriers are overcome, rigid assessment of algorithmic performance and quality control from regulatory bodies can slow down the adoption. In this context, model synergy can be used to accelerate the integration of novel technologies into clinical practice by increasing clinical interpretability, validating generality of findings, and accelerating regulatory decision-making.

Model validation towards generality of findings

The goal after validating an initial concept is to extend it to a more general patient cohort, with less controlled characteristics. The problem of sampling bias, based on both intrinsic (physiological) and extrinsic (environmental) demographic heterogeneity of the population, becomes relevant when implementing solutions for broader patient cohorts.[55,56] Consequently, models (as clinical guidelines) may need recalibrations when used on populations from different countries or ethnicities, or even from different centres in the same country. In recent years, only 6% of artificial intelligence algorithms had external evaluation performed (note this is beyond the minimum requirement of using the learning, validation, and testing partitions of the data), and none adopted the three design criteria of a robust validation: diagnostic cohort design, the inclusion of multiple institutions, and prospective data collection.[57] The quality of datasets also needs to be thoroughly validated to avoid possible biases before the models developed from them can be integrated in clinical decision-making.[58]

To address this issue, an increasing number of institutions are creating initiatives for data-sharing platforms, aiming at reusing existing

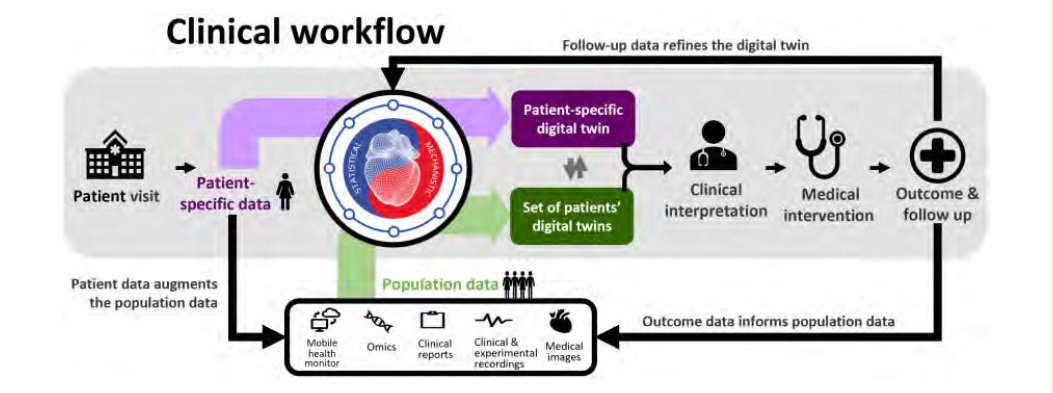


Figure 3 Envisioned clinical workflow using the fully developed digital twin concept. Population data, collected from preceding patients and study cohorts, are used to create and validate statistical and mechanistic models, as well as to create a population-based digital twin (green). Novel patient data are analysed with the help of the existing models and integrated to form the patient's digital twin (purple). The comparison and interaction between digital twins give valuable insight (phenotyping, risk assessment, prediction of disease development. . .) that is clinically interpreted and combined with traditional data to aid in the process of clinical decision-making. The digital twin develops in line with the patient's condition—adjusting and improving in accordance with the follow-up data. Resulting outcomes are supplemented to shape population data and refine the follow-up data.

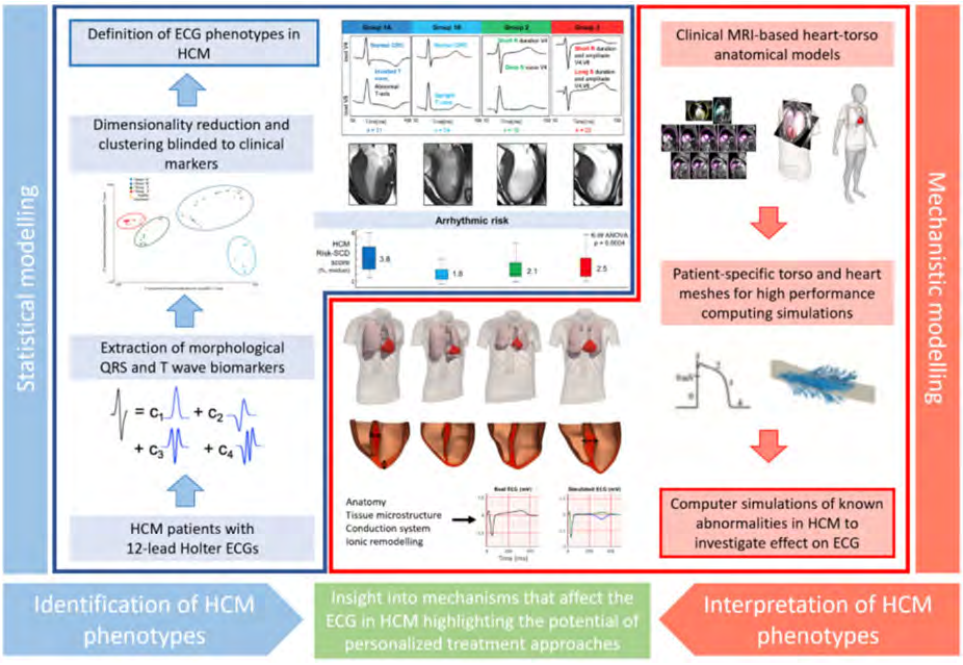


Figure 4 Synergy between mechanistic and statistical models in the definition of electrocardiogram (ECG) biomarkers for the management of hypertrophic cardiomyopathy.[47,48]

datasets and verifying published research works.[59] Governments, regulatory agencies, and philanthropic funders are promoting the open science culture, enforcing publishing patient-level data by means of compliance to product launching, funding application, and journal publishing.[60]

Another approach to improve the generality of data is the generation of synthetic cases of a representative wider population. The core idea is to expand the average mechanistic model to obtain populations of models, all of them parameterized within the range of physiological variability obtained by experimental protocols.[61,62] Such an approach, which allows investigating many more scenarios than possible experimental acquisitions, is not only able to evaluate the impact of physiological variability but to explain the mechanisms underpinning inter-individual variability in therapy response (e.g. adverse drug reactions), and to identify sub-populations at higher risk.[63,64] Statistical shape modelling techniques can represent inter-patient anatomical variability for a cohort, and be used in combination of mechanistic models for clinical decision support systems.[65]

As in traditional scientific research, mechanistic and statistical models are complementary tools to verify the findings derived from one another. Finding a mechanistic explanation of an inductive inference from statistical models increases its plausibility, such as the redistribution of work in the left bundle branch block to explain the remodelling pattern that predicts response to CRT.[51] Equivalently, data computed from mechanistic models need to be scrutinized quantitatively as it was done in the comparison of clinical and simulation groups to validate a model for acute normovolaemic haemodilution.[66]

An important final remark is that randomized control trials will always be needed to establish evidence that can never be obtained from large observational databases.[67]

Models as critical tools for accelerating regulatory decision-making

Clinical decisions are built on evidence from bench to bedside. Regulatory decisions, on the contrary, are often based on heterogeneous, limited, or completely absent human data, as in the case of approval for first-in-human clinical trials. In this regard, the results of computational models can now be accepted for some regulatory submissions.[68,69] Digital evidence obtained using computer simulations can be used for safety of therapy prior to first-in-human use, or under scenarios not ethically possible in human.[38] Computational models have an increasingly important role in the overall product life cycle management, proving useful in the processes of design optimization for development and testing, supplemental non-clinical testing, and post-market design changes and failure assessment.[27]

The development process for medical devices involves manufacturing and testing samples under a wide range of scenarios, which is often time-consuming and financially overwhelming. Moreover, preclinical testing conditions are often very simplified with respect to the actual patient environment. Statistical and mechanistic models synergistically offer to streamline this process, where statistical models can be used to collect a representative virtual patient cohort, and mechanistic models can then be used to simulate the device behaviour under defined scenarios. In this way, new devices can be tested in a representative virtual patient population, thereby decreasing the risk before moving to an actual clinical trial. An example is HEARTguide™ (FEops nv, Belgium), where device–patient interactions after transcatheter aortic valve implantation can be predicted.[25]

The augmentation of clinical trial design with virtual patients is also an evolving idea.[70–72] This would overcome limitations of current empirical trials, where patients burdened with comorbidities or complex treatment regimens are often excluded from the trials, and enrolled individuals are handled under reductionist approaches, assuming they share a common phenotype. Such approaches often fail to capture differences in response to treatment.[70] Alternatively, computational evidence can inform collection of novel evidence from clinical trials,[13,38] where models can improve patient selection by derived biomarkers and predictions. This offers an opportunity to answer questions traditionally restricted by financial or ethical considerations, and to investigate therapy efficacy in more clinically relevant cases. Computational modelling can also facilitate safe methods to explore treatment effects in sub-populations clinically more complex to address, such as patients with rare diseases or paediatric cohorts, and therefore may allow for insights not possible in the current clinical trial practice.

One of the first examples in which digital evidence (i.e. an *in silico* trial) replaced any additional clinical evidence was in the approval of the Advisa MRI SureScan pacemaker (Medtronic, Inc.).[73] Another powerful example is a computer simulator of type 1 diabetes mellitus,[74] which was accepted by the FDA as a substitute to animal trials for the pre-clinical testing of control strategies in artificial pancreas studies. Later, an investigational device exemption (i.e. the approval needed to initiate a clinical study), issued solely on the basis of modelling testing, was granted by the FDA for a closed-loop control clinical trial of the safety and effectiveness of the proposed artificial pancreas algorithm.

In the context of drug safety and efficacy assessment, an unmet need is filling the gaps between animal translation or *in vitro* preparations and prediction of the human response. Mechanistic models may assist in scaling observations into humans.[75] This is, for example, the goal of the CIPA initiative,[69] sponsored by the FDA among others, aiming at facilitating the adoption of a new paradigm for assessment of potential risk of clinical Torsades de Pointes, where mechanistic models of human electrophysiology will play a crucial role. This is reinforced by a recent study in which human *in silico* trials outperformed animal models in predicting clinical pro-arrhythmic cardiotoxicity, so they might be soon integrated into existing drug safety assessment pipelines.[63]

Finally, after a product is launched, mechanistic models can be still used for post-market re-evaluation and failure assessment in order to identify any potential underlying problems. This creates a valuable opportunity for simulations to evaluate any design changes planned for next-generation productions, ultimately closing the product life cycle loop, and demonstrating the ubiquitous presence and utility of statistical and mechanistic models in the future of medical product regulation.

Discussion

The digital twin, i.e., the dynamic integration and augmentation of patient data using mechanistic and statistical models, is the actual

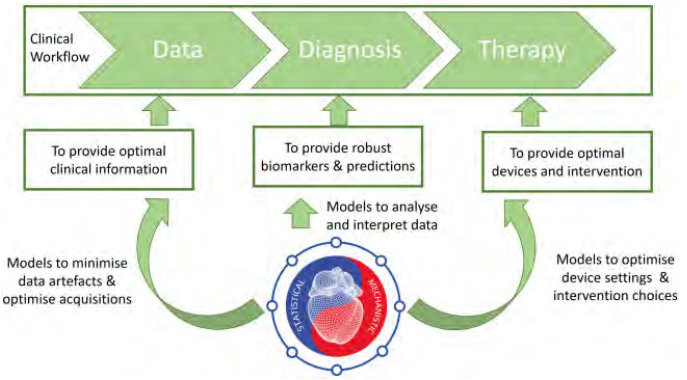


Figure 5 The vision of a personalized *in silico* cardiology, where the digital twin informs all the stages through the clinical workflow. Models are used (i) to optimize data acquisition and the information extracted from it, (ii) to evaluate current health status and inform diagnosis and risk stratification, and (iii) to optimize clinical devices and drug selection to deliver a personalized therapy.

pathway towards the vision of precision medicine. Simple and fragmented components of the digital twin are already used in clinical practice: a decision tree in a clinical guideline encapsulates the best-documented evidence that is based in statistical and mechanistic insights. The digital twin will gradually include tailored computer-enabled decision points, and create the transition from healthcare systems founded on describing disease to healthcare systems focused on predicting response, and thus shifting treatment selection from being based on the state of the patient today to optimizing the state of the patient tomorrow.

Envisioned impact and timeline

The digital twin provides a pathway to map current patient observations into a predictive framework, combining inductive and deductive reasoning. Early components of the digital twin are already making a clinical impact. In a generic clinical workflow divided in the stages of data acquisition, diagnosis, and therapy planning, computational models can provide value in the three stages, see *Figure 5*. To improve data acquisition techniques, there are already statistical models to automate the image analysis tasks.[16] To provide better diagnosis, a virtual fractional flow reserve can replace an invasive catheter,[29,37] or the body surface recordings can be mapped to the surface of the heart.[39] With regards to therapy planning, a virtual deployment of the valve replacement[25,76] or a roadmap to guide ablation procedures[77,78] represents existing techniques (statistical and mechanistic) that have been implemented into the clinical workflow. These solutions have thus met regulatory approval, where they are referred to as 'software as a medical device', and where guidelines from the International Medical Device Regulators Forum are accepted by the EU and the USA.

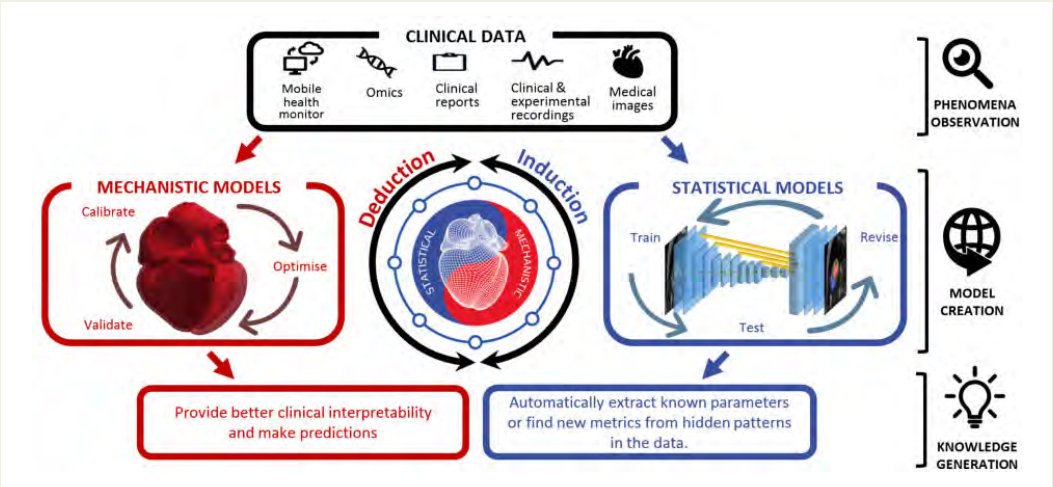
A digital twin will follow the life journey of each person and harness both data collected by wearable sensors and lifestyle information that patients may register, shifting the clinical approach towards preventive healthcare. A notable challenge is the integration of these data with healthcare organizations, where security and confidentiality of the sensitive information remain paramount.

The currently still fragmented and incipient concept of the digital twin will be gradually crystallized and adopted during the next 5–10 years. The holistic integration of a Digital Twin is the aspiration that will be reached through two complementary and synergetic pathways: the first is the refinement of key decision points in the management of cardiac disease, driven by personalized mechanistic models that are informed by key pieces of patient's data; and the second is the disease-centred optimization of the patient's lifetime journey through the healthcare system, driven by statistical models being informed by the electronic health record of a large population.

On the actual implementation of the digital twin, we envision that the evolution will be towards a gradually better inter-operability of current health information systems, leading to a distributed location of the information. Digital twin users will mainly be citizens and physicians, with different interfaces that retrieve the relevant data and trigger the analysis capabilities hosted in the local device or remote cloud resources. The analysis may also require specialized skills that may be delivered by industry, or even by *computational cardiologists* inside healthcare organizations.

Organizational and societal challenges ahead

Access to data is the main challenge in both the development and the clinical translation of the digital twin, caused by infrastructural, regulatory, and societal reasons. Information systems and electronic health records are fragmented, highly heterogeneous and difficult to interoperate. Information is often contained in unstructured format, and its extraction requires either manual work or further research efforts of automation through natural language processing technologies.[79] Simulations may also require specialized skills and supercomputers.



Take home figure The cardiovascular digital twin that will deliver the vision of precision medicine by the synergetic combination of computer-enhanced induction (using statistical models learnt from data) and deduction (mechanistic modelling and simulation integrating multi-scale knowledge).

In this context, provision of digital twin technologies may be enabled by cloud infrastructures (e.g. HeartFlow FFR_{CT} Analysis).

Consent and confidentiality are key ingredients to address the societal concerns when handling the personal data needed to develop and validate digital twin technologies. The EU General Data Protection Regulation (GDPR) has imposed new legal requirements, such as the right to withdraw consent and the right to be forgotten, causing controversy about the cost and feasibility of its enforcement.[80] Any digital twin solution that holds enough information to identify a patient needs to carefully watch these requirements, that also apply to retrospective data and safety backups.

Potential professional, cultural, and ethical issues

As more clinical tasks are performed by models, the fear of replacement of physicians by machines may arise. In some scenarios, machines may match or even outperform physicians.[81] In other scenarios, human experts, by not practising on the easy problems solved by the machine, may lose the skills that may still be needed when dealing with difficult cases.

The second professional barrier is the mistrust that originates from a 'black box', where predictions derived by algorithms are not matched with a plausible explanation. Generation of evidence is one clear way to generate trust. Another solution is to use methods to illustrate the logic inside the box, including clustering and association techniques,[82] which may help to identify the causes and mechanisms.

From the patient's perspective, personalization creates the opportunity of more involvement in healthcare decisions. Patients will be empowered to better manage their disease using the digital twin to gain information about their current and predicted state, and potentially to adopt optimized lifestyle suggestions. A well-informed patient shall have more efficient discussions with physicians, and consent and decide faster on diagnostic or treatment procedures.

Finally, on the ethical side, there is a risk of models to create or exacerbate existing racial or societal biases in healthcare systems: if a group is misrepresented in the data used to train models, that group may receive a sub-optimal treatment.[83]

Recommendations

The pathway to accelerate the clinical impact with digital twin technologies is to generate trust among researchers, clinicians, and society.

Research communities shall avoid inflating expectations. Claims about generality and potential impact should be based on rigorous methodology, with external cohorts to demonstrate the validity of inferences, and with the quantification of the uncertainty of predictions.[84] Any model is a simplified representation of the reality, with a limited scope and dependence on assumptions made. The opportunity is an adequate handling of these limitations, with models able to identify data inconsistencies, and with data used to constrain and verify the model assumptions.[85]

As an emerging field, the digital twin needs guidelines, gold-standards, and benchmark tests.[86,87] Scientific organizations and regulatory bodies have released guidelines that can be used to establish the level of rigour needed for computational modelling.[27] Such guidelines and standards are useful tools as they allow regulators to judge computational evidence and industry to understand regulatory requirements for computational models, leveraging a substantial part of the risk and uncertainty associated to the development of these new technologies. They can even increase and facilitate their translational impact, as the quality and robustness of the models and their reporting will

increase by adhering to such guidance during model development. Further effort is needed to widen the scope of these first multi-stakeholder consensuses involving industry, academia, and regulators. Current initiatives that develop visions, technologies, or infrastructure relevant to the 'digital twin' community are Elixir (https://elixir-europe.org/), FAIRDOM (https://fair-dom.org/), and EOSC (https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud).

The education of citizens, care providers, physicians, and researchers in the uses and possibilities of digital twin technologies is key for its adoption and acceptance. University education systems should also allow for the exchange of knowledge at the earliest stages of the career: medical students should have some computational training, just as engineers in biomedical industry should be trained in cardiology during their studies.[88] And postgraduate training programmes should bridge remaining cultural and language gaps between disciplines, such as our Personalised In-silico Cardiology EU funded Innovative Training Network (https://picnet.eu).

Conclusion

Precision cardiology will be delivered, not only by data, but also by the inductive and deductive reasoning built in the digital twin of each patient. Treatment and prevention of cardiovascular disease will be based on accurate predictions of both the underlying causes of disease and the pathways to sustain or restore health. These predictions will be provided and validated by the synergistic interplay between mechanistic and statistical models. The early steps towards this vision have been taken, and the next ones depend on the coordinated drive from scientific, clinical, industrial, and regulatory stakeholders in order to build the evidence and tackle the organizational and societal challenges ahead.

Supplementary material

Supplementary material is available at *European Heart Journal* online.

Conflict of interest: none declared.

Funding

This work was supported by the EU's Horizon 2020 Marie Sklodowska-Curie ITN Projects (g.a. 764738 and 766082), the EU's Horizon 2020 research and innovation programme (g.a. 675451 and 823712), the Wellcome/EPSRC Centre for Medical Engineering (WT 203148/Z/16/Z), the National Research Agency (ANR) (g.a. ANR-10-IAHU-04), the NC3RS (NC/P001076/1) and the British Heart Foundation (RE/13/2/30182, RE/13/1/30181, TG/17/3/33406, PG/16/75/32383, FS/17/22/32644, CH/16/3/21406, RG/16/14/32397). E.Pueyo holds an ERC Starting Grant (g.a. 638284). B. Rodriguez and P.Lamata hold Wellcome Trust Senior Research Fellowships (214290/Z/18/Z, 209450/Z/17/Z).

References

1. Antman EM, Loscalzo J. Precision medicine in cardiology. *Nat Rev Cardiol* 2016;**13**:591–602.
2. Trayanova N. From genetics to smart watches: developments in precision cardiology. *Nat Rev Cardiol* 2019;**16**:72–73.
3. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med* 2015;**372**:793–795.
4. Joyner MJ, Paneth N. Promises, promises, and precision medicine. *J Clin Invest* 2019;**129**:946–948.
5. Khoury MJ. Precision medicine vs preventive medicine. *JAMA* 2019;**321**:406.
6. Noble D. Evolution beyond neo-Darwinism: a new conceptual framework. *J Exp Biol* 2015;**218**:7–13.
7. Alber M, Buganza Tepole A, Cannon WR, De S, Dura-Bernal S, Garikipati K, Karniadakis G, Lytton WW, Perdikaris P, Petzold L, Kuhl E. Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *NPJ Digit Med* 2019;**2**:115.
8. Tao F, Cheng J, Qi Q, Zhang M, Zhang H, Sui F. Digital twin-driven product design, manufacturing and service with big data. *Int J Adv Manuf Technol* 2018;**94**:3563–3576.
9. Lamata P. Teaching cardiovascular medicine to machines. *Cardiovasc Res* 2018;**114**:e62–e64.
10. Niederer SA, Plank G, Chinchapatnam P, Ginks M, Lamata P, Rhode KS, Rinaldi CA, Razavi R, Smith NP. Length-dependent tension in the failing heart and the efficacy of cardiac resynchronization therapy. *Cardiovasc Res* 2011;**89**:336–343.
11. Lumens J, Tayal B, Walmsley J, Delgado-Montero A, Huntjens PR, Schwartzman D, Althouse AD, Delhaas T, Prinzen FW, Gorcsan J. Differentiating electromechanical from non-electrical substrates of mechanical discoordination to identify responders to cardiac resynchronization therapy. *Circ Cardiovasc Imaging* 2015;**8**:e003744.
12. Corral Acero J, Zacur E, Xu H, Ariga R, Bueno-Orovio A, Lamata P, Grau V. *SMOD - Data Augmentation Based on Statistical Models of Deformation to Enhance Segmentation in 2D Cine Cardiac MRI*. FIMH 2019: Functional Imaging and Modeling of the Heart - pp. 361–369. doi: 10.1007/978-3-030-21949-9_39.
13. Cikes M, Sanchez Martinez S, Claggett B, Solomon SD, Bijnens B. Machine-learning integration of complex echocardiographic patterns and clinical parameters from cohorts and trials. *Eur Heart J* 2019;**40**: doi: 10.1093/eurheartj/ehz745.0147.
14. Shameer K, Johnson KW, Glicksberg BS, Dudley JT, Sengupta PP. Machine learning in cardiovascular medicine: are we there yet? *Heart* 2018;**104**:1156–1164.
15. Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. *Nat Rev Cardiol* 2016;**13**:350–359.
16. Dey D, Slomka PJ, Leeson P, Comaniciu D, Shrestha S, Sengupta PP, Marwick TH. Artificial intelligence in cardiovascular imaging: JACC state-of-the-art review. *J Am Coll Cardiol* 2019;**73**:1317–1335.
17. Niederer SA, Lumens J, Trayanova NA. Computational models in cardiology. *Nat Rev Cardiol* 2019;**16**:100–111.
18. Johnson KW, Shameer K, Glicksberg BS, Readhead B, Sengupta PP, Björkegren JLM, Kovacic JC, Dudley JT. Enabling precision cardiology through multiscale biology and systems medicine. *JACC Basic Transl Sci* 2017;**2**:311–327.
19. Davies MR, Wang K, Mirams GR, Caruso A, Noble D, Walz A, Lavé T, Schuler F, Singer T, Polonchuk L. Recent developments in using mechanistic cardiac modelling for drug safety evaluation. *Drug Discov Today* 2016;**21**:924–938.
20. Tung L. A bi-domain model for describing ischemic myocardial D-C potentials. 1978. https://dspace.mit.edu/handle/1721.1/16177 (29 February 2020).
21. Sherwin SJ, Formaggia L, Peiró J, Franke V. Computational modelling of 1D blood flow with variable mechanical properties and its application to the simulation of wave propagation in the human arterial system. *Int J Numer Methods Fluids* 2003;**43**:673–700.
22. Guidi G, Pettenati MC, Miniati R, Iadanza E. Random forest for automatic assessment of heart failure severity in a telemonitoring scenario. In: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, pp. 3230–3233.
23. Stegle O, Fallert SV, MacKay DJ, Brage S. Gaussian process robust regression for noisy heart rate data. *IEEE Trans Biomed Eng* 2008;**55**:2143–2151.
24. Roney CH, Bayer JD, Cochet H, Meo M, Dubois R, Jaïs P, Vigmond EJ. Variability in pulmonary vein electrophysiology and fibrosis determines arrhythmia susceptibility and dynamics. *PLoS Comput Biol* 2018;**14**:e1006166.
25. de Jaegere P, De Santis G, Rodriguez-Olivares R, Bosmans J, Bruining N, Dezutter T, Rahhab Z, El Faquir N, Collas V, Bosmans B, Verhegghe B, Ren C, Geleinse M, Schultz C, van Mieghem N, De Beule M, Mortier P. Patient-specific computer modeling to predict aortic regurgitation after transcatheter aortic valve replacement. *JACC Cardiovasc Interv* 2016;**9**:508–512.
26. Gray RA, Pathmanathan P. Patient-specific cardiovascular computational modeling: diversity of personalization and challenges. *J Cardiovasc Transl Res* 2018;**11**:80–88.
27. Morrison TM, Dreher ML, Nagaraja S, Angelone LM, Kainz W. The role of computational modeling and simulation in the total product life cycle of peripheral vascular devices. *J Med Device* 2017;**11**:024503.
28. Dillon-Murphy D, Noorani A, Nordsletten D, Figueroa CA. Multi-modality image-based computational analysis of haemodynamics in aortic dissection. *Biomech Model Mechanobiol* 2016;**15**:857–876.
29. Morris PD, van de Vosse FN, Lawford PV, Hose DR, Gunn JP. "Virtual" (computed) fractional flow reserve. *JACC Cardiovasc Interv* 2015;**8**:1009–1017.

30. Xi J, Lamata P, Niederer S, Land S, Shi W, Zhuang X, Ourselin S, Duckett SG, Shetty AK, Rinaldi CA, Rueckert D, Razavi R, Smith NP. The estimation of patient-specific cardiac diastolic functions from clinical measurements. *Med Image Anal* 2013;**17**:133–146.
31. Wang ZJ, Wang VY, Bradley CP, Nash MP, Young AA, Cao JJ. Left ventricular diastolic myocardial stiffness and end-diastolic myofibre stress in human heart failure using personalised biomechanical analysis. *J Cardiovasc Transl Res* 2018;**11**: 346–356.
32. Krittian SBS, Lamata P, Michler C, Nordsletten DA, Bock J, Bradley CP, Pitcher A, Kilner PJ, Markl M, Smith NP. A finite-element approach to the direct computation of relative cardiovascular pressure from time-resolved MR velocity data. *Med Image Anal* 2012;**16**:1029–1037.
33. Donati F, Figueroa CA, Smith NP, Lamata P, Nordsletten DA. Non-invasive pressure difference estimation from PC-MRI using the work-energy equation. *Med Image Anal* 2015;**26**:159–172.
34. Donati F, Myerson S, Bissell MM, Smith NP, Neubauer S, Monaghan MJ, Nordsletten DA, Lamata P. Beyond Bernoulli: improving the accuracy and precision of non-invasive estimation of peak pressure drops. *Circ Cardiovasc Imaging* 2017;**10**:e005207.
35. Nørgaard BL, Leipsic J, Gaur S, Seneviratne S, Ko BS, Ito H, Jensen JM, Mauri L, De Bruyne B, Bezerra H, Osawa K, Marwan M, Naber C, Erglis A, Park S-J, Christiansen EH, Kaltoft A, Lassen JF, Bøtker HE, Achenbach S. Diagnostic performance of noninvasive fractional flow reserve derived from coronary computed tomography angiography in suspected coronary artery disease. *J Am Coll Cardiol* 2014;**63**:1145–1155.
36. Min JK, Taylor CA, Achenbach S, Koo BK, Leipsic J, Nørgaard BL, Pijls NJ, De Bruyne B. Noninvasive fractional flow reserve derived from coronary CT angiography. *JACC Cardiovasc Imaging* 2015;**8**:1209–1222.
37. Rajani R, Modi B, Ntalas I, Curzen N. Non-invasive fractional flow reserve using computed tomographic angiography: where are we now and where are we going? *Heart* 2017;**103**:1216–1222.
38. Morrison TM, Pathmanathan P, Adwan M, Margerrison E. Advancing regulatory science with computational modeling for medical devices at the FDA's Office of Science and Engineering Laboratories. *Front Med* 2018;**5**: doi: 10.3389/fmed.2018.00241.
39. Haissaguerre M, Hocini M, Shah AJ, Derval N, Sacher F, Jais P, Dubois R. Noninvasive panoramic mapping of human atrial fibrillation mechanisms: a feasibility report noninvasive panoramic mapping of human atrial fibrillation mechanisms. Introduction. *J Cardiovasc Electrophysiol* 2013;**24**:711–717.
40. Daubert C, Behar N, Martins RP, Mabo P, Leclercq C,. Avoiding non-responders to cardiac resynchronization therapy: a practical guide. *Eur Heart J* 2016;**38**: 1463–1472.
41. Prakosa A, Arevalo HJ, Deng D, Boyle PM, Nikolov PP, Ashikaga H, Blauer JJE, Ghafoori E, Park CJ, Blake RC, Han FT, MacLeod RS, Halperin HR, Callans DJ, Ranjan R, Chrispin J, Nazarian S, Trayanova NA. Personalized virtual-heart technology for guiding the ablation of infarct-related ventricular tachycardia. *Nat Biomed Eng* 2018;**2**:732–740.
42. Hill YR, Child N, Hanson B, Wallman M, Coronel R, Plank G, Rinaldi CA, Gill J, Smith NP, Taggart P, Bishop MJ. Investigating a novel activation-repolarisation time metric to predict localised Vulnerability to reentry using computational modelling. *PLoS One* 2016;**11**: doi: 10.1371/journal.pone.0149342.
43. Alfonso F, Nihoyannopoulos P, Stewart J, Dickie S, Lemery R, McKenna WJ. Clinical significance of giant negative T waves in hypertrophic cardiomyopathy. *J Am Coll Cardiol* 1990;**15**:965–971.
44. Pelliccia A, Di Paolo FM, Quattrini FM, Basso C, Culasso F, Popoli G, De Luca R, Spataro A, Biffi A, Thiene G, Maron BJ. Outcomes in athletes with marked ECG repolarization abnormalities. *N Engl J Med* 2008;358:152–161.
45. Pelliccia A, Corrado D, Bjørnstad HH, Panhuyzen-Goedkoop N, Urhausen A, Carre F, Anastasakis A, Vanhees L, Arbustini E, Priori S. Recommendations for participation in competitive sport and leisure-time physical activity in individuals with cardiomyopathies, myocarditis and pericarditis. *Eur J Prev Cardiol* 2006;**13**: 876–885.
46. Passini E, Mincholé A, Coppini R, Cerbai E, Rodriguez B, Severi S, Bueno-Orovio A. Mechanisms of pro-arrhythmic abnormalities in ventricular repolarisation and anti-arrhythmic therapies in human hypertrophic cardiomyopathy. *J Mol Cell Cardiol* 2016;**96**:72–81.
47. Lyon A, Bueno-Orovio A, Zacur E, Ariga R, Grau V, Neubauer S, Watkins H, Rodriguez B, Mincholé A. Electrocardiogram phenotypes in hypertrophic cardiomyopathy caused by distinct mechanisms: apico-basal repolarization gradients vs. Purkinje-myocardial coupling abnormalities. *Europace* 2018;**20**: III102–III112.
48. Lyon A, Ariga R, Mincholé A, Mahmod M, Ormondroyd E, Laguna P, de Freitas N, Neubauer S, Watkins H, Rodriguez B. Distinct ECG phenotypes identified in hypertrophic cardiomyopathy using machine learning associate with arrhythmic risk markers. *Front Physiol* 2018;**9**:213.
49. Arevalo HJ, Vadakkumpadan F, Guallar E, Jebb A, Malamas P, Wu KC, Trayanova NA. Arrhythmia risk stratification of patients after myocardial infarction using personalized heart models. *Nat Commun* 2016;**7**: doi: 10.1038/ncomms11437.
50. Dawes TJW, de Marvao A, Shi W, Fletcher T, Watson GMJ, Wharton J, Rhodes CJ, Howard LSGE, Gibbs JSR, Rueckert D, Cook SA, Wilkins MR, O'Regan DP. Machine learning of three-dimensional right ventricular motion enables outcome prediction in pulmonary hypertension: a cardiac MR imaging study. *Radiology* 2017;**283**:381–390.
51. Warriner DR, Jackson T, Zacur E, Sammut E, Sheridan P, Hose DR, Lawford P, Razavi R, Niederer SA, Rinaldi CA, Lamata P. An asymmetric wall-thickening pattern predicts response to cardiac resynchronization therapy. *JACC Cardiovasc Imaging* 2018;**11**:1545–1546.
52. Mirams GR, Pathmanathan P, Gray RA, Challenor P, Clayton RH. Uncertainty and variability in computational and mathematical models of cardiac physiology. *J Physiol* 2016;**594**:6833–6847.
53. Pathmanathan P, Gray RA. Ensuring reliability of safety-critical clinical applications of computational cardiac models. *Front Physiol* 2013;**4**:358.
54. Winslow RL, Trayanova N, Geman D, Miller MI. Computational medicine: translating models to clinical care. *Sci Transl Med* 2012;**4**:158rv11–158rv11.
55. Kurokawa J, Kodama M, Clancy CE, Furukawa T. Sex hormonal regulation of cardiac ion channels in drug-induced QT syndromes. *Pharmacol Ther* 2016;**168**: 23–28.
56. Niemeijer MN, van den Berg ME, Deckers JW, Aarnoudse ALHJ, Hofman A, Franco OH, Uitterlinden AG, Rijnbeek PR, Eijgelsheim M, Stricker BH. ABCB1 gene variants, digoxin and risk of sudden cardiac death in a general population. *Heart* 2015;**101**:1973–1979.
57. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. *Korean J Radiol* 2019;**20**:405.
58. Chang KC, Dutta S, Mirams GR, Beattie KA, Sheng J, Tran PN, Wu M, Wu WW, Colatsky T, Strauss DG, Li Z. Uncertainty quantification reveals the importance of data variability and experimental design considerations for in silico proarrhythmia risk assessment. *Front Physiol* 2017;**8**:917.
59. Dey P, Ross JS, Ritchie JD, Desai NR, Bhavnani SP, Krumholz HM. Data sharing and cardiology. *J Am Coll Cardiol* 2017;**70**:3018–3025.
60. Schiltz M. Science without publication paywalls: cOAlition S for the realisation of full and immediate open access. *PLoS Med* 2018;**15**:e1002663.
61. Britton OJ, Bueno-Orovio A, Van Ammel K, Lu HR, Towart R, Gallacher DJ, Rodriguez B. Experimentally calibrated population of models predicts and explains intersubject variability in cardiac cellular electrophysiology. *Proc Natl Acad Sci USA* 2013;**110**:E2098–E2105.
62. Sánchez C, Bueno-Orovio A, Wettwer E, Loose S, Simon J, Ravens U, Pueyo E, Rodriguez B. Inter-subject variability in human atrial action potential in sinus rhythm versus chronic atrial fibrillation. *PLoS One* 2014;**9**:e105897.
63. Passini E, Britton OJ, Lu HR, Rohrbacher J, Hermans AN, Gallacher DJ, Greig RJH, Bueno-Orovio A, Rodriguez. Human in silico drug trials demonstrate higher accuracy than animal models in predicting clinical pro-arrhythmic cardiotoxicity. *Front Physiol* 2017;**8**: doi: 10.3389/fphys.2017.00668.
64. Sánchez C, Bueno-Orovio A, Pueyo E, Rodríguez B. Atrial fibrillation dynamics and ionic block effects in six heterogeneous human 3D virtual atria with distinct repolarization dynamics. *Front Bioeng Biotechnol* 2017;**5**:29.
65. Liang L, Liu M, Martin C, Elefteriades JA, Sun W. A machine learning approach to investigate the relationship between shape features and numerically predicted risk of ascending aortic aneurysm. *Biomech Model Mechanobiol* 2017;**16**: 1519–1533.
66. Sims CR, Delima LR, Calimaran A, Hester R, Pruett WA. Validating the physiologic model HumMod as a substitute for clinical trials involving acute normovolemic hemodilution. *Anesth Analg* 2018;**126**:93–101.
67. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available: table 1. *Am J Epidemiol* 2016;**183**:758–764.
68. Colatsky T, Fermini B, Gintant G, Pierson JB, Sager P, Sekino Y, Strauss DG, Stockbridge N. The Comprehensive in Vitro Proarrhythmia Assay (CiPA) initiative—update on progress. *J Pharmacol Toxicol Methods* 2016;**81**:15–20.
69. Cavero I, Holzgrefe H. CiPA: ongoing testing, future qualification procedures, and pending issues. *J Pharmacol Toxicol Methods* 2015;**76**:27–37.
70. Pappalardo F, Russo G, Tshinanu FM, Viceconti M. In silico clinical trials: concepts and early adoptions. *Brief Bioinform* 2019:**20**:1699–1708.
71. Viceconti M, Henney A, Morley-Fletcher E. In silico clinical trials: how computer simulation will transform the biomedical industry. *Int J Clin Trials* 2016; **3**:37.
72. Haddad T, Himes A, Thompson L, Irony T, Nair R; MDIC Computer Modeling and Simulation Working Group Participants. Incorporation of stochastic engineering models as prior information in Bayesian medical device trials. *J Biopharm Stat* 2017;**27**:1089–1103.

73. Faris O, Shuren J. An FDA viewpoint on unique considerations for medical-device clinical trials. *N Engl J Med* 2017;**376**:1350–1357.
74. Kovatchev BP, Breton M, Man CD, Cobelli C. In silico preclinical trials: a proof of concept in closed-loop control of type 1 diabetes. *J Diabetes Sci Technol* 2009;**3**: 44–55.
75. Zemzemi N, Bernabeu MO, Saiz J, Cooper J, Pathmanathan P, Mirams GR, Pitt-Francis J, Rodriguez B. Computational assessment of drug-induced effects on the electrocardiogram: from ion channel to body surface potentials. *Br J Pharmacol* 2013;**168**:718–733.
76. Rocatello G, El Faquir N, De Santis G, Iannaccone F, Bosmans J, De Backer O, Sondergaard L, Segers P, De Beule M, de Jaegere P, Mortier P. Patient-specific computer simulation to elucidate the role of contact pressure in the development of new conduction abnormalities after catheter-based implantation of a self-expanding aortic valve. *Circ Cardiovasc Interv* 2018;**11**:e005344.
77. Andreu D, Ortiz-Pérez JT, Fernández-Armenta J, Guiu E, Acosta J, Prat-González S, De Caralt TM, Perea RJ, Garrido C, Mont L, Brugada J, Berruezo A. 3D delayed-enhanced magnetic resonance sequences improve conducting channel delineation prior to ventricular tachycardia ablation. *Europace* 2015;**17**:938–945.
78. Cedilnik N, Duchateau J, Dubois R, Sacher F, Jaïs P, Cochet H,Sermesant M. Fast personalized electrophysiological models from computed tomography images for ventricular tachycardia ablation planning. *Europace* 2018;**20**:iii94–iii101.
79. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, Forshee R, Walderhaug M, Botsis T. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *J Biomed Inform* 2017;**73**:14–29.
80. Politou E, Alepis E, Patsakis C. Forgetting personal data and revoking consent under the GDPR: challenges and proposed solutions. *J Cybersecurity* 2018;**4**: doi: 10.1093/cybsec/tyy001.
81. Darcy AM, Louie AK, Roberts LW. Machine learning and the profession of medicine. *JAMA* 2016;**315**:551.
82. Esfandiari N, Babavalian MR, Moghadam A-ME, Tabar VK. Expert systems with applications knowledge discovery in medicine: current issue and future trend. *Expert Syst Appl* 2014;**41**:4434–4463.
83. Nordling L. A fairer way forward for AI in health care. *Nature* 2019;**573**: S103–S105.
84. Pathmanathan P, Cordeiro JM, Gray RA. Comprehensive uncertainty quantification and sensitivity analysis for cardiac action potential models. *Front Physiol* 2019; **10**:721.
85. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep* 2016;**6**:26094.
86. Land S, Gurev V, Arens S, Augustin CM, Baron L, Blake R, Bradley C, Castro S, Crozier A, Favino M, Fastl TE, Fritz T, Gao H, Gizzi A, Griffith BE, Hurtado DE, Krause R, Luo X, Nash MP, Pezzuto S, Plank G, Rossi S, Ruprecht D, Seemann G, Smith NP, Sundnes J, Rice JJ, Trayanova N, Wang D, Jenny Wang Z, Niederer SA. Verification of cardiac mechanics software: benchmark problems and solutions for testing active and passive material behaviour. *Proc R Soc A Math Phys Eng Sci* 2015;**471**:20150641.
87. Niederer SA, Kerfoot E, Benson AP, Bernabeu MO, Bernus O, Bradley C, Cherry EM, Clayton R, Fenton FH, Garny A, Heidenreich E, Land S, Maleckar M, Pathmanathan P, Plank G, Rodríguez JF, Roy I, Sachse FB, Seemann G, Skavhaug O, Smith NP. Verification of cardiac tissue electrophysiology simulators using an N-version benchmark. *Philos Trans R Soc A Math Phys Eng Sci* 2011;**369**: 4331–4351.
88. Eden C, Johnson KW, Gottesman O, Bottinger EP, Abul-Husn NS. Medical student preparedness for an era of personalized medicine: findings from one US medical school. *Per Med* 2016;**13**:129–141.

Appendix B

Confidence metrics for Paper II and Paper IV

Confidence metrics are an important part of the implementation of deep learning algorithms in clinical workflows. These metrics allow networks to ignore many inputs from an unseen distribution that would otherwise be incorrectly classified (as shown in Paper I). A confidence metric was presented in Paper I based on the last fully connected layer in the classification network. Here confidence metrics are also shown for Paper II and Paper IV. These metrics were not included in the original articles because it was not developed yet for Paper II and was outside the scope of the work for Paper IV.

B.1 Confidence metrics for Paper II: left ventricle dimension measurement

The confidence metric for Paper II consisted of two components. The objective of Paper II was the prediction of endpoints for performing caliper measurements in 2D cardiac ultrasound. The most likely mistake made by the network is to predict multiple potential endpoint locations in the image. This will be represented in the heatmap as multiple areas of activation ().

Using the center of mass method for coordinate extraction (see Paper II), the predicted caliper location will be in between the two regions of activation. using an argmax instead will pick one of the activation regions, but the position of the argmax is generally more random than the center of mass and is therefore less desirable in general. Squaring the heatmap and re-normalizing helps eliminate some smaller regions of activation but is still susceptible to dual activations.Occurrences of dual activations can be easily determined by measuring the activation level beneath a small window surrounding the predicted caliper location and comparing it to an expected threshold.

The second component of the confidence metric consists of measuring the angle between the predicted calipers. A high angle between calipers will indicate to an observer that the measurement is incorrect. Although the loss function (see Paper II includes a term to minimize this, it still may occur in some cases. Therefore, the total angle difference between caliper pairs (calculated by cosine similarity) can be summed and the image flagged if it exceeds a threshold.

These heatmap (Equation (B.1)) and angle (Equation (B.2)) confidence metrics can be described mathematically as:

$$\left(\sum_{i=x-w}^{x+w} \sum_{j=y-w}^{y+w} H_n(i,j) \right) < T_H \quad \forall H_n \in \{H_0, H_1, ...H_N\} \tag{B.1}$$

$$\frac{1}{2} \left(\sum_{\vec{c_0} \in C} \sum_{\vec{c_1} \in C} 1 - \frac{\vec{c_0} \cdot \vec{c_1}}{\|\vec{c_0}\| \ \|\vec{c_1}\|} \right) < T_A \tag{B.2}$$

Where x, y is the predicted caliper location, H_n is the heatmap for endpoint n, N is the total number of heatmaps ($2\times$ the number of calipers), w is the window size, $\vec{c_0}$ & $\vec{c_1}$ are individual calipers in the set of calipers C, and T_H & T_A are the thresholds determined for the heatmap and angle metrics respectively. These thresholds can be modified depending on the desired trade-off in detectability and accuracy. Note that (B.1) calculated for each heatmap while (B.2) is calculated for a sum of all calipers.

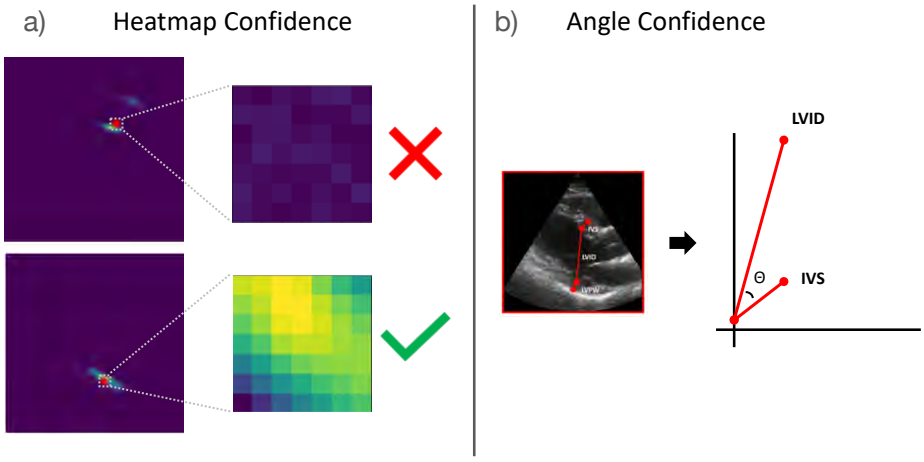


Figure B.1: Confidence metrics for dimension measurements: a) heatmap confidence will be triggered when multiple high-likelihood regions cause the final point to be pulled off either likely region, b) confidence based on the angle between calipers will detect cases where the endpoints were detected well, but the relationship between calipers is incorrect.

B.2 Confidence metric for Paper IV: left ventricle segmentation

The confidence metric for Paper IV was derived from the unsupervised evaluation metrics originally proposed by Zhu et al. [136] and applied to left ventricle segmentation in [137]. The authors propose evaluating a segmentation in

terms of it's *convexity* and *simplicity* defined below for a segmentation S in Equation (B.3) and Equation (B.4) respectively.

$$convexity(S) = \frac{Area(S)}{Area(ConvexHull(S))} \tag{B.3}$$

$$simplicity(S) = \frac{\sqrt{4\pi * Area(S)}}{Perimeter(S)} \tag{B.4}$$

These metrics are high for smooth, simple shapes and both equal 1 for a perfect circle. While the left ventricle is more elliptical than circular in shape, these two metrics generally which mimic the desired appearance of a segmentation. We thus define a confidence metric for left ventricle segmentation in terms of the harmonic mean of convexity and simplicity:

$$\frac{Simplicity(S) * Convexity(S)}{Simplicity(S) + Convexity(S)} < T_S \tag{B.5}$$

This metric evaluates the realism of a segmentation. Like the metrics in B.1, the threshold T_S can be modified depending on the desired trade-off between detectability and accuracy.