
Article:
Lane, E.S., Azarmehr, N. orcid.org/0000-0002-6367-207X, Jevsikov, J. et al. (5 more authors) (2021) Multibeat echocardiographic phase detection using deep neural networks. Computers in Biology and Medicine, 133. 104373. ISSN 0010-4825

https://doi.org/10.1016/j.compbiomed.2021.104373

Multibeat Echocardiographic Phase Detection Using Deep Neural Networks

Elisabeth S Lane[1*], Neda Azarmehr[2], Jevgeni Jevsikov[1], James P Howard[2], Matthew J Shun-shin[2], Graham D Cole[2], Darrel P Francis[2], Massoud Zolgharni[1,2]

[1]*School of Computing and Engineering, University of West London, London, United Kingdom*

[2]*National Heart and Lung Institute, Imperial College, London, United Kingdom*

Abstract

Background: Accurate identification of end-diastolic and end-systolic frames in echocardiographic cine loops is important, yet challenging, for human experts. Manual frame selection is subject to uncertainty, affecting crucial clinical measurements, such as myocardial strain. Therefore, the ability to automatically detect frames of interest is highly desirable.

Methods: We have developed deep neural networks, trained and tested on multi-centre patient data, for the accurate identification of end-diastolic and end-systolic frames in apical four-chamber 2D multibeat cine loop recordings of arbitrary length. Seven experienced cardiologist experts independently labelled the frames of interest, thereby providing infallible annotations, allowing for observer variability measurements.

Results: When compared with the ground-truth, our model shows an average frame difference of -0.09±1.10 and 0.11±1.29 frames for end-diastolic and end-systolic frames, respectively. When applied to patient datasets from a different clinical site, to

* Corresponding author: School of Computing and Engineering, University of West London, St Mary's Rd, Ealing, London. United Kingdom. W5 5RF. E-mail address: Elisabeth.Lane@uwl.ac.uk.

which the model was blind during its development, average frame differences of -1.34±3.27 and -0.31±3.37 frames were obtained for both frames of interest. All detection errors fall within the range of inter-observer variability: [-0.87, -5.51]±[2.29, 4.26] and [-0.97, -3.46]±[3.67, 4.68] for ED and ES events, respectively.

Conclusions: The proposed automated model can identify multiple end-systolic and end-diastolic frames in echocardiographic videos of arbitrary length with performance indistinguishable from that of human experts, but with significantly shorter processing time.

Keywords: Echocardiography, Cardiac imaging, Deep learning, Phase detection

1. Introduction

Assessment of left ventricular (LV) function is of principal importance during an echocardiographic examination and is crucial for accurate patient evaluation. Echocardiography continues to be the most common technique in clinical practice for the quantification of LV function markers; such as ejection fraction (EF) and global longitudinal strain (GLS) [1]. Measurements usually relate to time points, such as end-diastole (ED) and end-systole (ES). Therefore, accurate detection of the end of the LV systole and diastole phases constitutes a critical step in any echocardiographic exam.

1.1. The need for fully automated systems

The importance of accurate identification of ED and ES frames has crucially been demonstrated by Mada et al. [2]. An error of just two to three frames in detecting ES

elicits an approximate 10% difference in segmental ES strain. Furthermore, the sensitivity of frame selection is greater in relation to the left bundle branch block. As highlighted by Amundsen [3], the consequence of misidentification of ED and ES frames can be extensive; impairing concordance between observers in both research and clinical practice. Therefore, automated methods for the resolution of accurate ED and ES phase detection could greatly contribute to improving the consistency of echocardiographic quantification.

The process of identifying ED and ES frames in video data is manually performed by trained clinicians via on-screen visual selection. ED frames can be determined using cues such as mitral valve closure, ECG R-wave and maximum LV volume. Whereas ES frames are commonly defined by mitral valve opening, minimal LV volume, aortic valve closure, or the end of the ECG T-wave. However, due to subtle frame-on-frame spatial differences, and complex temporal relationships virtually invisible to the human eye, manual detection presents a significant barrier to consistent diagnosis due to intra- and inter-observer variability lacking reproducibility and precision [4].

We previously identified the medial disagreement between accredited and experienced experts as 3 frames [5] when performing manual identification. Therefore, reliable and reproducible methods for ED and ES frame detection would allow for the development of fully automated techniques. Thus, meeting the objective of accurate quantification of LV function, in addition to automated calculation of EF and stroke volume, GLS and wall thickening.

1.2. Value of independence from ECG

Often, cardiac timing is determined through analysis of an accompanying ECG signal during an echocardiogram exam. Despite providing information enabling the computation of some clinically important parameters (such as temporal intervals from the R-wave peaks), ECG recordings require the connection of multiple cables which is time-consuming and, at times, inconvenient. In an era when highly portable scanners can be used to undertake focused studies lasting just a few minutes [6], the capacity of detecting cardiac timing events, independent from the ECG signal, has potential to be useful in implementing such automated technology on handheld devices.

In the absence of ECG signal, tissue Doppler data has been used to estimate cardiac cycle length [7] or detect ED frames [8]. Machine learning approaches have also been applied to automatically detect ED and ES frames from 2D echocardiography images (B-mode). This includes manifold learning [9], speckle tracking [10], correlation-based frame-to-frame deviation measures [11,12], nonlinear filtering and boundary detection techniques [13].

More recently, studies have focused on deep learning approaches, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Deep residual recurrent neural networks were applied to phase detection in apical-4-chamber (A4C) echocardiograms [14]. A major limitation of this study is the proposed model only accepts videos with a fixed length of frames, containing just one cardiac cycle. Presumably, this approach necessitates pre-processing of the input image sequence to isolate a single heartbeat.

The same authors later reported on combining CNN and RNN modules to detect frames of interest [15]. Although varying length inputs (22-59 frames) were used, again their results indicate the videos contained just one cardiac cycle. It is assumed this variation in length was probably due to different frame and heart rates.

In addition, 3D CNNs have been applied for the extraction of spatial-temporal features from A4C and apical-2-chamber (A2C) echocardiographic videos [16]. While the study states the model was trained on variable length sequences, the feasibility of the model was demonstrated only on a pair of detected ED/ES frames in each video with the QRS-complex in the accompanying ECG signal being used to detect an additional ED frame for the videos starting in systole phase; thereby providing ground-truth for a full cardiac cycle (ED-ES-ED).

A summary of all accuracies from previously reported studies [14-16], compared with those of our developed models, is provided in the results section.

1.3. Value of multibeat analysis

In clinical practice, longer recordings would allow for probing of physiological reactions after intervention, where detecting a subtle change in the mean value of a clinical maker, amongst much larger background beat-to-beat variability, is essential.

As stated, recent studies have failed to target the application of automated phase detection in arbitrarily long, uninterrupted echocardiogram recordings containing several full heartbeats. Clinically, it is necessary to monitor changes in crucial markers, such as EF or strain, from one examination to the next. Measurements taken from only one heartbeat may result in test-retest variability. Therefore, it would be impossible to

reliably conclude whether a patient's condition has deteriorated over time. Such variability and inaccuracy can be reduced by averaging measurements over several heartbeats, from the same acquisition. However, this is impractical when a proposed automated model is incapable of returning more than one single pair of ED/ES frame predictions.

We previously reported on the issue of beat-to-beat variability in echocardiography and potential bias due to using a single measurement from a single heartbeat [17-19]. When random variability between heartbeats is large, clinicians use "clinical judgement" to select which value to report; largely unaware of the devastating consequences for subsequent use. The ability to acquire and automatically analyse many heartbeats within reasonable time constraints would permit clinical protocols to be developed for multi-beat measurements, hence reducing undesirable variability between clinical assessments. In such measurements, the exact time of ED and ES events for each heartbeat is required.

1.4. Clinical deployability

Without exception, all previously reported studies related to echocardiographic phase detection have used 'single-centre' clinical datasets with one set of expert annotations for model developments and evaluations. Experience shows the performance of models trained using a single dataset may reduce considerably when transferred from one clinical site to another, and when applied to the images obtained from a different group of patients, collected using different equipment, and acquired/annotated following different protocols [20]. We have demonstrated this drawback in relation to the

echocardiographic LV segmentation where a model, designed and claimed to be superior by the authors using their own dataset, underperformed compared to the standard U-net architecture when applied to our patient data [21].

This limitation has proved prohibitive to the development of automated models becoming an acceptable mainstream methodology in daily clinical practice. Evaluating models on multi-centre clinical datasets/annotations naturally results in (i) greater patient numbers with a wider range of groups, representing the overall population, lower systematic bias in terms of (ii) protocols/guidelines and (iii) imaging equipment, and (iv) external validity. This would make the developed models less prone to overfitting, thus, resulting in their increased generalisability, in contrast to single-centre dataset studies.

1.5. Importance of external validation

Similarly, all previously reported studies on phase detection have used 'private' datasets. Therefore, accurate interpretation of the reported results, encompassing a wide range of accuracies, is not feasible since a direct comparison of the frame detection accuracy would require access to the same patient dataset. This highlights the importance of external validation of the automated algorithms and using publicly available benchmarks, before they can be incorporated into clinical use. To date, no study has used and reported accuracies on a publicly available echocardiography dataset.

1.6. Main contributions

A validated and clinically deployable solution is a central feature of our research approach. The main contributions of this research can be summarised as being the first study of its kind to:

- investigate the feasibility of using a deep learning framework to detect ED and ES frames in echocardiographic videos of arbitrary length, containing several heartbeats
- demonstrate the applicability of the developed framework by including several patient datasets from various clinical centres, where one dataset was used for model development and the others used for testing
- use annotations (ground-truth) from several cardiologist experts, allowing for the examination of inter- and intra-observer variability
- make our patient dataset and models publicly available, thereby providing a benchmark for future studies and allowing for external validation of our approach
- include performance reports on an independent external dataset, made available for the LV segmentation challenge, ensuring generalisability by transcontinental data inclusion

2. Methodology

2.1. Dataset, ethics and expert annotations

Descriptions of the datasets used in this study is as follows, with a brief summary provided in Table 1.

PACS-dataset

A large random sample of echocardiographic studies from different patients performed between 2010 and 2020 was extracted from Imperial College Healthcare NHS Trust's echocardiogram database. Ethical approval was obtained from the Health Regulatory Agency for the anonymised export of large quantities of imaging data. It was not necessary to approach patients individually for consent of data originally acquired for clinical purposes.

The images were acquired during examinations performed by experienced echocardiographers, according to the standard protocols for using ultrasound equipment from GE and Philips manufacturers. Only studies with full patient demographic data, and without intravenous contrast administration, were included. Automated anonymisation was performed to remove the patient-identifiable information. A detailed description, including patient characteristics, can be found in Howard et al. [22].

A CNN model, previously developed in our research group to detect different echocardiographic views [22], was then used to identify and separate the A4C views. A total of 1,000 videos from different patients of varying lengths, containing 1-3 heartbeats, were randomly selected.

Two accredited and experienced cardiology experts manually selected ED and ES frames, each blinded to the judgment of the other. We developed a custom-made program closely replicating the interface of clinical echocardiography hardware. Operators visually inspected the cine loops by controlled animation using a trackball, or arrow keys. The operators were asked to pick ED and ES frames in the A4C view, as they would in preparation for a Biplane Simpson's measurement in clinical practice. Selections were made in one or more sessions at their convenience and the time taken was recorded. The annotations were then used to define the reference ground-truth ED and ES frames for model developments (both training and testing).

Finally, the original DICOM-formatted image sequences were down sampled by cubic interpolation into a standardised size of 112×112 pixels.

MultiBeat-dataset

2D echocardiographic images were collected from 40 patients (18 males), with an age range of 27-80 years and a mean age of 59 years, who were referred for echocardiographic examination in the Echocardiography Department at St Mary's Hospital, London. There were no selection criteria, and all patients were in sinus rhythm. The study was approved by the local ethics committee and written- informed consent was obtained from all patients.

Standard transthoracic echocardiography was performed using a GE Vivid.i (GE Healthcare, London, United Kingdom) ultrasound machine equipped with a 1.5-3.6 MHz transducer (3S-RS). For each subject, an A4C view was obtained in left lateral decubitus position as per standard clinical guidelines [23]. The operators performing the exam were instructed not to change any machine setting (e.g. sector, gain, depth, etc.)

and the probe position during the acquisition period to obtain consistent data. The acquisition period was 20 seconds to make sure at least 10 cardiac cycles were present in all videos. The images were stored digitally for subsequent offline analysis. The ECG trace was present on all echocardiographic recordings.

Using the same platform described for the PACS-dataset and in a similar process, five other accredited and experienced cardiology experts manually selected ED and ES frames, again each blinded to the judgment of the others. All videos were then renamed and provided to one operator in a random order for second analysis, no previous result was shown. Thus, the operator was blinded from their own previous frame selections. To maintain independence, the operators annotating the MultiBeat-dataset were different from those who labelled the PACS-dataset.

Where an operator judged a beat to be of low quality, they declared it invalid and did not make a selection. Therefore, since the operators were blinded to each other and their own previous selections, there were heartbeats that were delineated on one or two viewings only by each operator. Only the heartbeats which had 6 delineations (540 in total) were used for testing the models. The location of the typical frames identified by the operators is plotted as red circular markers in Fig.3. DICOM-formatted image sequences were again down sampled by cubic interpolation into a standardised size of 112×112 pixels.

EchoNet-dataset

This publicly available dataset [24], originally shared for the task of LV segmentation, contains 10,030 A4C echocardiography videos from individuals who underwent imaging between 2016 and 2018 as part of routine clinical care at Stanford University

Hospital. Each video has been cropped and masked to remove text and information outside of the scanning sector.

The image sequences are provided with a dimension of 112×112 pixels. The videos are annotated by a registered sonographer. Although some videos may contain a couple of heartbeats, only one pair of ED/ES frames is labelled and were used as the reference ground-truth for testing the developed models (no training was performed using this dataset). A more detailed description of the EchoNet-dataset can be found in [26].

Table 1. A summary of the patient datasets used in this study.

Dataset Name	**PACS-dataset**	**MultiBeat-dataset**	**EchoNet-dataset**
Source	Private NHS Trust PACS Archives - Imperial College Healthcare	Private St Mary's Hospital Acquired for this study	Publicly available Stanford University Hospital echonet.github.io/dynamic
Ultrasound machine	Philips Healthcare (iE33 xMATRIX)	GE Healthcare (Vivid.i) and Philips Healthcare (iE33 xMATRIX)	Siemens Healthineers (Acuson SC2000) and Philips Healthcare (iE33, Epiq 5G, Epiq 7C)
Number of videos/patients	1,000	40	10,030
Length of videos	1-3 heartbeats	≥ 10 heartbeats	1 heartbeat
Ground-truth	2 annotations by 2 experts	6 annotations by 5 experts (twice by one expert)	1 annotation
Original size (pixels)	(300-768)×(400-1024)	422×636	112×112
Frame rate (fps)	23-102	52-80	50
Format	DICOM	DICOM	AVI
Use	Training/Testing	Testing	Testing

2.2. Ground-truth definition

The target output, or ground-truth, was generated using reference annotations provided by experts and subsequently used to train the deep learning models. As highlighted in section 1.2, and in order for the model to be entirely independent from the ECG, we did not rely on the ECG-derived information for model developments and testing because: (i) not all videos contained the ECG trace, and (ii) even if present, the ECG trace in many cases was too noisy or of insufficient quality for any meaningful analysis.

Treating the definition of ground-truth as a classification task, with three classes for frames (ED, ES, trivial), would result in an imbalanced problem since the 'trivial' class would be greatly over-represented. A recent study put forth the argument of a binary classification approach for cardiac phase detection [16]. However, by allocating the same label to all frames in the diastole phase (1) and systole phase (0), one risks ignoring high-level spatial and temporally related markers, including crucial physiological differences throughout the entire cardiac cycle.

Therefore, the problem was formulated as a regression task. To label individual cardiac frames, it was assumed the predictions for a cardiac sequence should decrease during the systole phase and increase during the diastole phase. Given two consecutive ground truth labels y_i and y_{i-1}, we expect $y_{i-1} \prec y_i$ in systole, and vice versa. Assigning the target values of 1 and 0 to ED and ES time-points, respectively, and using a linear interpolation function, the target outputs for all constituent farmers between the two events were defined as:

$$y_t = \begin{cases} -\dfrac{f_t - f_{ED}}{f_{ES} - f_{ED}}, & in\ systole\ phase \\ \dfrac{f_t - f_{ES}}{f_{ED} - f_{ES}}, & in\ diastole\ phase \end{cases}$$

Here, y_t is the ground-truth label for frame f_t at time-point *t,* and f_{ED} and f_{ES} are the frame numbers for ED and ES events, respectively. For a video containing multiple heartbeats, the ground-truth will therefore appear as a zigzag profile as illustrated in Fig.1. Due to varying video length, some contain a combination of singular or multiple events in the image sequence.

2.3. Neural network architecture

Considering the patient image sequences as visual time-series, we adopted Long-term Recurrent Convolutional Networks (CNN+LSTM) for analysing the echocardiographic videos. Such architectures are a class of models that is both spatially and temporally deep, specifically designed for sequence prediction problems (e.g., order of images) with spatial inputs (e.g. 2D structure or pixels in an image) [26].

Fig.1A. provides an overview of the network architecture. The model comprises (i) CNN unit for the encoding of spatial information for each frame of an echocardiographic video input, (ii) LSTM units for the decoding of complex temporal information, and (iii) a regression unit for the prediction of the frames of interest.

Spatial feature extraction: First, a CNN unit is used to extract a spatial feature vector from every cardiac frame in the image sequence. A series of state-of-the-

art architectures were employed for the CNN unit. These included ResNet50, InceptionV3, DenseNet, and InceptionResNetV2, details of which can be found in the relevant resources [27 - 30].

Temporal feature extraction: The CNN unit above is only capable of handling a single image, transforming it from input pixels into an internal matrix or vector representation. LSTM units are therefore used to process the image features extracted from the entire image sequence by the CNN, i.e. interpreting the features across time steps. Stacks of LSTM units (1-layer to 4-layers) were explored, where the output of each LSTM unit not in the final layer is treated as input to a unit in the next.

Regression unit: Finally, the output of the LSTM unit is regressed to predict the location of ED and ES frames. The model returns a prediction for each frame in the cardiac sequence (timestep).

2.4. Deep learning framework

For the model to be capable of processing a video input of arbitrary length, thus containing any number of heartbeats and events, a sliding window approach was adopted. As illustrated in Fig.1B., a sliding window with a fixed stride segments the cardiac image sequence into overlapping chunks of fixed length. Each segment is then fed into the neural network model, as described above, where a prediction vector p_k is returned. The final target output is computed as:

$$\hat{y}_t = \frac{1}{K}\sum_{k=1}^{K} p_{k,t}$$

Where $p_{k,t}$ is the prediction for frame t in the k^{th} segment, and K is the total number of predictions available for each frame, obtained from overlapping segments. A peak detection algorithm then searched for the local maxima and minima, representing the ED and ES frames, respectively.

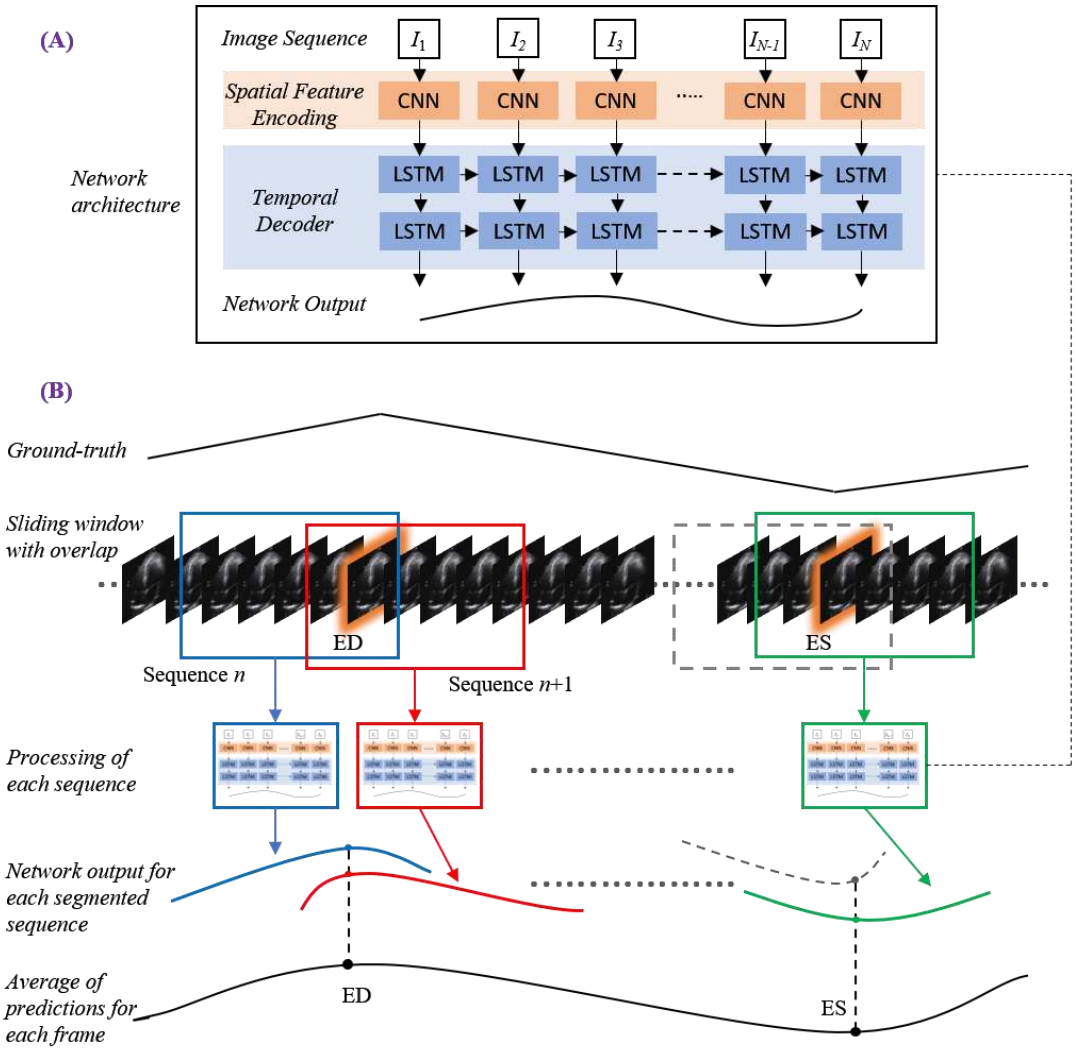

Fig. 1. Detailed schematic of the proposed deep learning framework: (A) the network architecture combining a CNN unit for spatial feature extraction with RNN (LSTM) blocks for temporal analysis; (B) the sliding window method processing fixed, overlapped, chunked sequences, generating multiple predictions for each frame with the mean calculated for each.

2.5. Implementation details

The models were implemented using the TensorFlow 2.0 deep learning framework [31] and trained using an NVIDIA GeForce ® GTX 1080 Ti GPU. Random, on the fly augmentation prevented overfitting, such as rotating between -10 and 10 degrees and spatial cropping between 0 and 10 pixels along each axis. The loss function was the mean squared error (MSE) with Adam optimiser [32] initialised with a learning rate of 10-5. Throughout the study, training was conducted over 70 epochs with a batch size of 2 for all models.

The PACS-dataset was used to train the models, with a data split of 60%, 20% and 20% for training, validation and testing, respectively. Early stopping was employed to avoid overfitting meaning training continued until the validation loss plateaued.

During testing, a sliding window of 30 frames in width with a stride of one was applied, allowing up to 30 predictions of differing temporal importance to be calculated for each timestep. Toward the end of each video, should a segment be fewer than 30 frames in length, it was zero-padded with the added frames removed after completion. Experimentation proved a stack of 2 LSTM layers was the optimum configuration across all models.

2.6. Evaluation metrics

As the primary endpoint for frame detection, evaluation of trained network predictions measures the difference between each labelled target y_t, either ED or ES, and the timestep prediction $\hat{y}_t$. Average Absolute Frame Difference (aaFD) notation is applied, where N is the number of events within the test dataset:

$$aaFD = \frac{1}{N}\sum_{t=1}^{N} |y_t - \hat{y}_t| \,,$$

The signed mean (μ) and standard deviation (σ) of the error (i.e. frame differences) were also calculated.

3. Results and discussion

3.1. PACS-dataset

The average time (mean±SD) taken by the operators to manually annotate ED/ES frames was 26±11 seconds, per event. The equivalent time for our automated models, executed on the GPU, was less than 1.5 seconds; significantly faster than the human-led process.

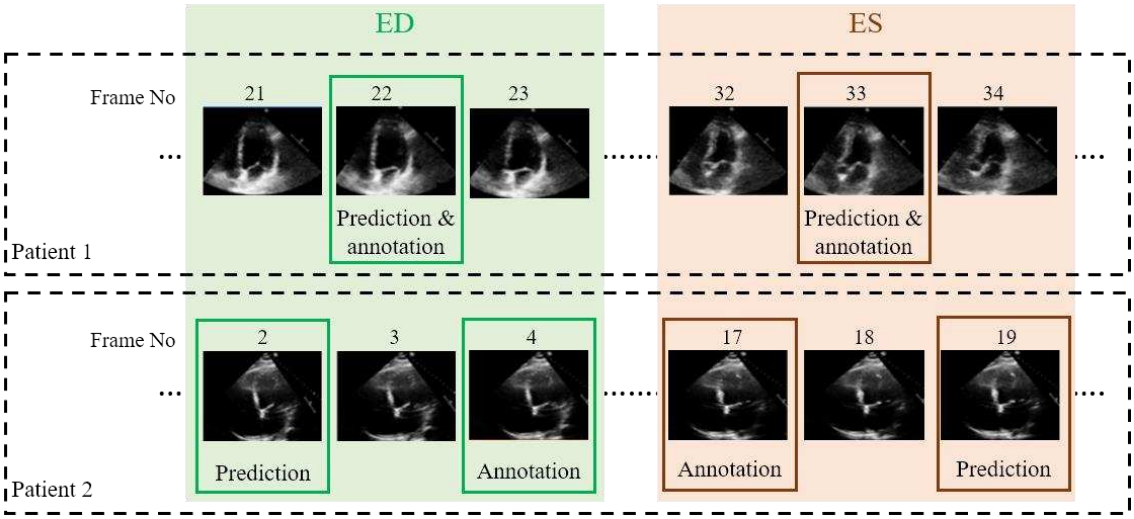

Fig.2. Examples of model's frame predictions (ResNet50 + 2x-LSTM) and Operator-1 annotations for two arbitrary patients from the PACS-dataset test set, when there is full agreement between the two (upper row) and conversely, when there is a mismatch (lower row).

Examples of two random patient videos for which the frame detection error is zero, as well as when there is a disagreement between the model's predictions and expert annotations. Table 2 details the error in ED and ES frame detection for all videos in the PACS-dataset. The results indicate the level of disagreement between Operator-1 annotations, considered as the ground-truth, compared with automated predictions and those made by Operator-2.

Table 2. Errors in ED and ES frame detection between Operator-1, the reference ground-truth, and predictions with Operator-2, for all testing videos in the PACS-dataset. Detection time is the average time it takes for the model (inference time) or the

operator (annotation time) to identify an ED/ES event. The best performing architecture, in terms of lowest detection error and shortest detection time, is highlighted.

Model/Operator	**ED**		**ES**		**Detection Time (s)**
	aaFD	$\mu \pm \sigma$	aaFD	$\mu \pm \sigma$	
ResNet50 + 2x-LSTM	**0.66**	**-0.09±1.10**	**0.81**	**0.11±1.29**	**0.776±0.33**
InceptionV3 + 2x LSTM	1.19	0.48±1.89	1.21	0.66±1.76	0.697±0.30
DenseNet + 2x LSTM	0.81	0.19±1.30	0.98	-0.01±1.53	1.379±0.59
InceptionResNetV2 + 2x LSTM	0.77	-0.02±1.38	0.83	0.23±1.29	1.07±0.46
Operator-2 (inter-observer)	1.55	-1.35±1.31	1.44	-0.90±1.80	26±11

The explored network architectures, which all employ the same type of RNN unit but use different state-of-the-art CNN modules, demonstrated comparable performance in terms of frame detection accuracy and inference time. In all models, the signed mean values are relatively small ($-0.09 \leq \mu \leq 0.66$) which indicate unbiased models; i.e. the models did not have a tendency to be consistently early or late, relative to the expert annotations. Conversely, Operator-2 was predominantly late in identifying both ED and ES events; $\mu = -1.35$ and -0.90 for ED and ES events, respectively.

'ResNet + 2x-LSTM' demonstrated a slim advantage, having the smallest discrepancy with Operator-1, and being the second fastest in terms of inference time of 0.78s for detecting each event. The aaFD was less than one frame in both events, with a mean difference of -0.09±1.10 and 0.11±1.29 frames for ED and ES events, respectively.

The discrepancy between Operator-1 and Operator-2 indicates a level of inter-observer variability; with an average absolute (and mean) frame difference of 1.55 (-1.35±1.31) and 1.44 (-0.90±1.80) frames for ED and ES events, respectively. Therefore, suggesting the discrepancy between automated models and Operator-1 is within the range of disagreement observed between two trained human operators.

Due to its lowest error, the 'ResNet + 2x-LSTM' architecture (hereinafter, referred to as the model) was selected for further analysis using the additional MultiBeat and EchoNet datasets. Table 3 provides a comparison between the performance of the model and previously reported deep learning results.

The model outperforms almost all existing approaches, indicating smaller discrepancies with the ground-truth from which it has learnt. However, caution is necessary, as different studies have used different private patient datasets, presumably with various levels of image quality and experience of human experts for annotations. Therefore, a direct comparison between the reported accuracies may not be as informative as desired. However, the proposed model's removal of all pre-processing steps and its capacity to identify multiple heartbeats in one long video is, however, an indisputable advantage.

It is also observed that ES frame detection error is consistently higher in all models than that for ED. Potentially owing to minute differences in consecutive frames indicating the mitral valve opening as the onset of the diastole phase is less apparent in the images; thus, resulting in a more challenging detection task for the model.

Table 3. Comparison of the proposed model with previously reported deep learning architectures regarding aaFD in ED and ES event detection.

Model	aaFD ED	aaFD ES
ResNet50 + 2x-LSTM	0.66	0.81
ResNet + 2x-LSTM [14]	3.7	4.1
3D CNN + LSTM [16]	1.6	1.7
DenseNet + 2x-Bi-GRU [15]	0.20	1.43

3.2 Multibeat-dataset

An ECG signal was recorded simultaneously alongside image acquisition for the MultiBeat-dataset and appears as a transverse trace on the echo image sequence. The ECG was extracted using a combination of constraints where the trace was assumed to be (i) continuous, (ii) have a consistent colour profile, and (iii) distinct from the background. The extracted signal for a random patient is used in Fig.3. to plot the identified frames by the human operators (6 annotations) and the automated model.

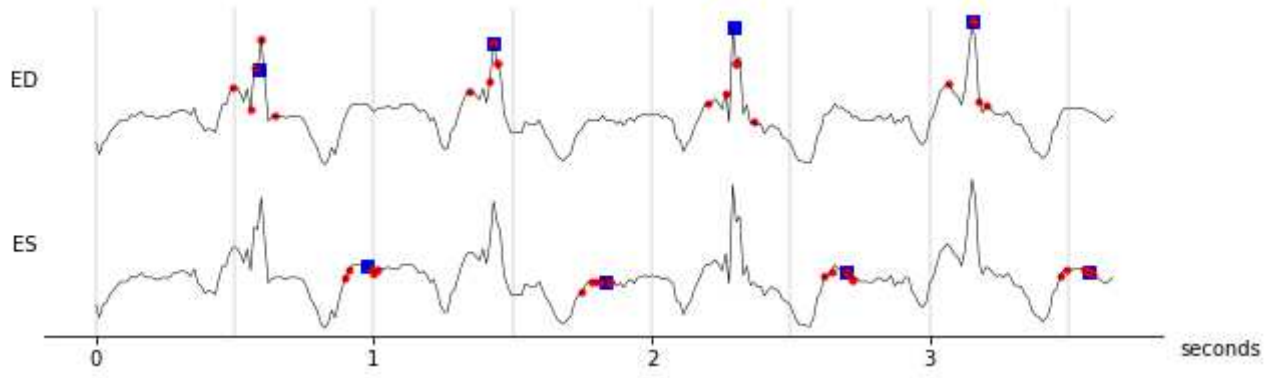

Fig. 3. Extracted ECG trace spanning 4 heartbeats for a random patient, delineated showing the 6 annotations from 5 operators (red circles) and automatically identified (blue squares) ED and ES frames.

Table 4 details detection errors between Operator-1 and detections made by the model and other operators. The model disagrees with Operator-1, as do Operators 2-5. Indeed, Operator-1 disagreed with themselves on their second annotation attempt (denoted as Operator-1b). The smallest error was the discrepancy between the two annotations on

separate occasions by the same operator (i.e. intra-observer variability), with a mean difference -0.22±2.76 and 0.25±3.75 for ED and ES events, respectively.

The range of mean difference between two different operators (i.e. inter-observer variability) was [-0.87, -5.51]±[2.29, 4.26] and [-0.97, -3.46]±[3.67, 4.68] for ED and ES events, respectively. The model discrepancy falls within the range of inter-observer variability. Clearly demonstrating the reliability of the model in frame detection, compared with the experienced human experts.

Significantly, both intra- and inter-observer variability measures suggest the experts' disagreement is greater when identifying ES frames. This is consistent with the model's performance, for which higher errors are observed when detecting ES frames.

Table 4. Errors in ED and ES frame detection between Operator-1a (considered as ground-truth) and predictions made by the other operators and the model for all testing videos in the MultiBeat-dataset. Operator-1b denotes the second set of annotations by the first human operator, indicating intra-observer variability.

Model/Operator	**ED**		**ES**	
	aaFD	$\mu \pm \sigma$	aaFD	$\mu \pm \sigma$
Operator-1a vs Operator-1b	1.96	-0.22 ± 2.76	1.90	0.25 ± 3.75
Operator-1a vs Operator-2	2.65	-1.22 ± 4.26	3.67	-2.25 ± 4.68
Operator-1a vs Operator-3	5.82	-5.51 ± 3.77	4.80	-4.46 ± 3.77
Operator-1a vs Operator-4	1.72	-0.87 ± 2.29	2.01	-0.97 ± 3.48
Operator-1a vs Operator-5	3.27	-2.96 ± 2.57	4.11	-3.64 ± 3.67
Operator-1a vs model	2.62	-1.34 ± 3.27	1.86	-0.31 ± 3.37

To ensure fair comparison between model performance and operators, Fig. 4. plots detection errors. Each human operator is compared with other 5, their consensus (mean) is considered as the reference annotation (red boxplots). The model is also compared with the consensus of the same 5 human annotations (blue boxplots). All 12 panels suggest performance of the model is similar, if not better, to that of an individual operator when using the other operators as a reference standard.

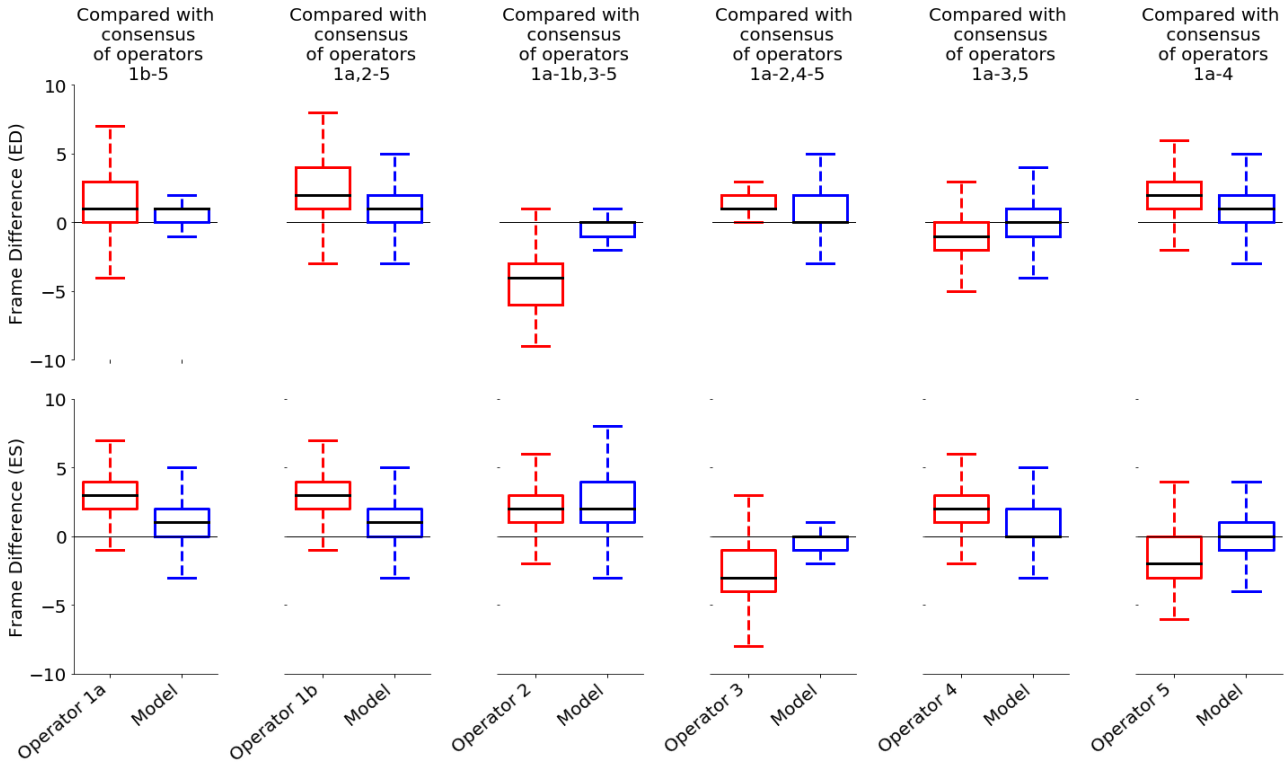

Fig.4. Errors in ED and ES frame identification by each operator, expressed relative to the consensus (mean) of all other 5 human annotations (red boxplots). In each case, alongside these errors, are those identified by the model expressed relative to the consensus of the same 5 annotations (blue boxplots). In the box-and-whisker plots, the

thick line represents the median, the box represents the quartiles, and the whiskers represent the 2.5% and 97.5% percentiles.

Because different human experts make different judgments, it is not possible for any automated model to agree with all expert annotations all the time. However, it is desirable for automated models to have fewer discrepancies when compared with the performance of human judgment. Given the model was never exposed to this dataset (image sequences, and any of the corresponding annotations), its predictions in ED and ES frame detection can be treated as one of the independent assessors.

Hence, for each heartbeat, there were 7 assessments of the desired frame; 6 human and one automated. Therefore, for each assessor, 6 frame differences were calculated when compared to other human or automated assessors. The pool of these differences across all heartbeats and image sequences indicates the overall performance for each assessor and is shown as boxplots in Fig.5.

Operator-4 demonstrates the smallest range of discrepancies in identification of ED frames (standard deviation of 3.47), but was consistently late, with a bias of -1.50 frames when compared to the consensus of other assessments.

The model had a relatively acceptable discrepancy from the consensus of the human operators, with a mean difference of 0.39±3.97 and 1.54±3.80 frames in ED and ES events, respectively. Indicating the model can be used to detect the frames of interest and that it is as reliable as the experienced human experts.

The range of human operator judgments for each heartbeat (i.e. difference between the earliest and latest manually identified frames) may be assumed as the uncertainty of the

reference method and, therefore, the highest accuracy obtainable. The mean frame intervals among all heartbeats was 8.10±3.84 and 7.01±4.28 frames for ED and ES events, respectively.

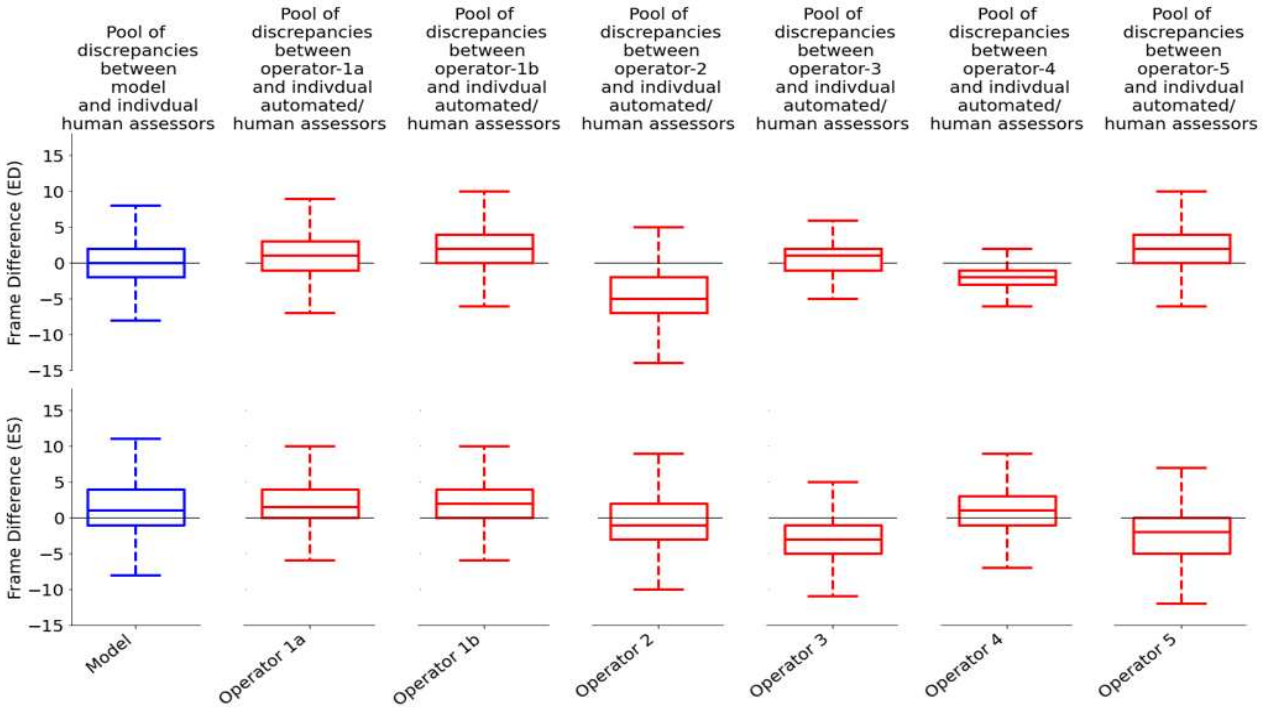

Fig.5. Errors in ED and ES frame identification by each of the assessors across all heartbeats and all patients. For each heartbeat there were 7 assessments (6 human and one automated). Errors are expressed as the pooled data from frame differences between each individual assessor and the 6 others. In the box-and-whisker plots, the thick line represents the median, the box represents the quartiles, and the whiskers represent the 2.5% and 97.5% percentiles.

3.3. EchoNet-dataset

In section 3.1, the proposed model was compared against previously reported approaches. However, each study used a different private dataset, making a direct comparison extremely difficult. Here, we applied our model to the publicly available EchoNet-dataset, allowing for future studies to be benchmarked against ours. Like the MultiBeat-dataset, no further training was carried out, and the dataset was used in its entirety for testing.

From the total number of videos (10,000), 810 were excluded owing to one of the ED or ES events occurring in the penultimate or final frame in the video, hence being unsuitable. EchoNet was made available for a challenge focused on segmentation of the left ventricular. Therefore, it was acceptable to have ED or ES events occurring in first or last frames. The retained 9,190 videos were fed into the model, when no resampling of the images was required as the dataset is provided with a resolution of 112×112 pixels; identical to the input size of our model.

An aaFD of 2.30 and 3.49 frames was obtained for ES and ES events, respectively and the mean frame difference was 0.16±3.56 and 2.64±3.59 for ED and ES; well within the range of inter-observer variability observed in section 3.2.

4. Conclusion

This study sought to investigate the feasibility of fully automated identification of ED and ES frames derived from 2D echocardiographic images and independent from an accompanying ECG signal using deep neural networks. The performance of the proposed method was examined by comparisons to gold standard reference data,

obtained from multiple cardiologist experts. It has been demonstrated that the performance of the proposed model is like that of human experts, with its detection error falling within the range of inter-observer variability and can therefore be used to reliably identify multiple ED and ES frames from videos of arbitrary length.

The performance of the automated model, measured as the processing time, is superior to that of human operators, where an improvement of >20 times was observed.

The proposed framework was tested on A4C views; however, it is the authors' belief that the utilised deep learning approaches could be applied to other echocardiographic views (no view-specific assumptions were made during the model developments). This will be the subject of future studies. As in previous studies, ours investigates 2D echocardiography as the clinically relevant modality. Currently, 3D echocardiography suffers from a considerable reduction in frame rate and image quality, hindering its adoption into routine practice [33]. When such issues are resolved, automatic frame detection in 3D images could be explored. Meanwhile, 2D echocardiography remains unrivalled, particularly when high frame rates are required.

Interpreting the results of our proposed model alongside other published architectures from the literature was not feasible. A direct comparison of detection accuracy would require access to the same patient dataset. At present, no echocardiography dataset, and corresponding annotations specifically prepared for cardiac phase detection, is publicly available. Additionally, representative multi-centre patient data, essential for ensuring any developed model would scale up well to other sites and environments, is currently scarce.

To address such broadly acknowledged shortcomings in the application of deep learning to echocardiography, we are developing Unity, a UK collaborative of cardiologists, physiologists and computer scientists under the aegis of the British Society of Echocardiography.

An image analysis interface has been developed in the form of a web-based, interactive, real-time platform to capture carefully curated expert annotations from numerous echo specialists with patient data provided from over twelve sites across the UK. Thus, ensuring coverage of multiple vendors, systems and environments. The Unity patient datasets, including the ones used in this study, are available for download at intsav.github.io. Additionally, all developed models designed using this annotation biobank and Unity datasets, are made available under open-source agreements at github.com/intsav, open to anyone to inspect, modify and improve upon them. This access to both datasets and models would allow an external validation of our findings.

Acknowledgment: This work was supported in part by the British Heart Foundation (Grant no. PG/19/78/34733). E Lane is supported by the Vice Chancellor's Scholarship at the University of West London. This research and open-access release of the has been conducted under: The Imperial AI Echocardiography Dataset [IRAS: 279328, REC:20/SC/0386].

References

[1] Knackstedt, C., Bekkers, S., Schummers, G., Schreckenberg, M., Muraru, D., Badano, L., Franke, A., Bavishi, C., Omar, A. and Sengupta, P., 2015. Fully Automated Versus Standard Tracking of Left Ventricular Ejection Fraction and Longitudinal Strain. *Journal of the American College of Cardiology*, 66(13), pp.1456-1466.

[2] Mada, R., Lysyansky, P., Daraban, A., Duchenne, J. and Voigt, J., 2015. How to Define End-Diastole and End-Systole?. *JACC: Cardiovascular Imaging*, 8(2), pp.148-157.

[3] Amundsen, B., 2015. It Is All About Timing!. *JACC: Cardiovascular Imaging*, 8(2), pp.158-160.

[4] Darvishi, S., Behnam, H., Pouladian, M. and Samiei, N., 2012. Measuring Left Ventricular Volumes in Two-Dimensional Echocardiography Image Sequence Using Level-set Method for Automatic Detection of End-Diastole and End-systole Frames. *Research in Cardiovascular Medicine*, 1(2), pp.39-45.

[5] Zolgharni, M., Negoita, M., Dhutia, N., Mielewczik, M., Manoharan, K., Sohaib, S., Finegold, J., Sacchi, S., Cole, G. and Francis, D., 2017. Automatic detection of end-diastolic and end-systolic frames in 2D echocardiography. *Echocardiography*, 34(7), pp.956-967.

[6] Testuz, A., Muller, H., Keller, P., Meyer, P., Stampfli, T., Sekoranja, L., Vuille, C. and Burri, H., 2012. Diagnostic accuracy of pocket-size handheld echocardiographs used by cardiologists in the acute care setting. *European Heart Journal - Cardiovascular Imaging*, 14(1), pp.38-42.

[7] Zolgharni, M., Francis, D., Dhutia, N., Cole, G., Bahmanyar, M., Jones, S., Sohaib, S., Tai, S., Willson, K. and Finegold, J., 2014. Automated Aortic Doppler Flow Tracing for Reproducible Research and Clinical Measurements. *IEEE Transactions on Medical Imaging*, 33(5), pp.1071-1082.

[8] Jahren, T., Steen, E., Aase, S. and Solberg, A., 2020. Estimation of End-Diastole in Cardiac Spectral Doppler Using Deep Learning. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, 67(12), pp.2605-2614.

[9] Shalbaf, A., Behnam, H., Gifani, P. and Alizadeh-Sani, Z., 2011. Automatic detection of end systole and end diastole within a sequence of 2-D echocardiographic images using modified Isomap algorithm. 2011 1st Middle East Conference on Biomedical Engineering.

[10] Aase, S., Snare, S., Dalen, H., Stoylen, A., Orderud, F. and Torp, H., 2010. Echocardiography without electrocardiogram. European Journal of Echocardiography, 12(1), pp.3-10.

[11] Tridandapani, S., Fowlkes, J. and Rubin, J., 2005. Echocardiography-Based Selection of Quiescent Heart Phases. Journal of Ultrasound in Medicine, 24(11), pp.1519-1526.

[12] Wick, C., Mcclellan, J., Ravichandran, L. and Tridandapani, S., 2013. Detection of Cardiac Quiescence From B-Mode Echocardiography Using a Correlation-Based Frame-to-Frame Deviation Measure. IEEE Journal of Translational Engineering in Health and Medicine, 1, pp.1900211-1900211.

[13] Ravichandran, L., Wick, C., McClellan, J., Liu, T. and Tridandapani, S., 2014. Detection of Quiescent Cardiac Phases in Echocardiography Data Using

Nonlinear Filtering and Boundary Detection Techniques. Journal of Digital Imaging, 27(5), pp.625-632.

[14] Dezaki, F., Dhungel, N., Abdi, A., Luong, C., Tsang, T., Jue, J., Gin, K., Hawley, D., Rohling, R. and Abolmaesumi, P., 2017. Deep Residual Recurrent Neural Networks for Characterisation of Cardiac Cycle Phase from Echocardiograms. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp.100-108.

[15] Taheri Dezaki, F., Liao, Z., Luong, C., Girgis, H., Dhungel, N., Abdi, A., Behnami, D., Gin, K., Rohling, R., Abolmaesumi, P. and Tsang, T., 2019. Cardiac Phase Detection in Echocardiograms With Densely Gated Recurrent Neural Networks and Global Extrema Loss. IEEE Transactions on Medical Imaging, 38(8), pp.1821-1832.

[16] Fiorito, A., Ostvik, A., Smistad, E., Leclerc, S., Bernard, O. and Lovstakken, L., 2018. Detection of Cardiac Events in Echocardiography Using 3D Convolutional Recurrent Neural Networks. 2018 IEEE International Ultrasonics Symposium (IUS).

[17] Pabari, P., Willson, K., Stegemann, B., van Geldorp, I., Kyriacou, A., Moraldo, M., Mayet, J., Hughes, A. and Francis, D., 2010. When is an optimization not an optimization? Evaluation of clinical implications of information content (signal-to-noise ratio) in optimization of cardiac resynchronization therapy, and how to measure and maximize it. Heart Failure Reviews, 16(3), pp.277-290.

[18] Moraldo, M., Cecaro, F., Shun-Shin, M., Pabari, P., Davies, J., Xu, X., Hughes, A., Manisty, C. and Francis, D., 2013. Evidence-based recommendations for

PISA measurements in mitral regurgitation: systematic review, clinical and in-vitro study. International Journal of Cardiology, 168(2), pp.1220-1228.

[19] Shun-Shin, M. and Francis, D., 2012. Why Are Some Studies of Cardiovascular Markers Unreliable? The Role of Measurement Variability and What an Aspiring Clinician Scientist Can Do Before It Is Too Late. Progress in Cardiovascular Diseases, 55(1), pp.14-24.

[20] Zech, J., Badgeley, M., Liu, M., Costa, A., Titano, J. and Oermann, E., 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine, 15(11), p.e1002683.

[21] Azarmehr, N., Xujiong Ye, Faraz Janan, J. P. Howard, D. Francis and M. Zolgharni., 2019. Automated Segmentation of Left Ventricle in 2D echocardiography using deep learning. In: MIDL. https://arxiv.org/abs/2003.07628

[22] Howard, J., Tan, J., Shun-Shin, M., Mahdi, D., Nowbar, A., Arnold, A., Ahmad, Y., McCartney, P., Zolgharni, M., Linton, N., Sutaria, N., Rana, B., Mayet, J., Rueckert, D., Cole, G. and Francis, D., 2020. Improving ultrasound video classification: an evaluation of novel deep learning methods in echocardiography. Journal of Medical Artificial Intelligence, 3, pp.4-4.

[23] Lang, R., Bierig, M., Devereux, R., Flachskampf, F., Foster, E., Pellikka, P., Picard, M., Roman, M., Seward, J. and Shanewise, J., 2006. Recommendations for chamber quantification☆. *European Journal of Echocardiography*, 7(2), pp.79-108.

[24] Echonet.github.io. 2021. *Echonet Dynamic*. [online] Available at: <https://echonet.github.io/dynamic/> [Accessed 19 January 2021].

[25] Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C., Heidenreich, P., Harrington, R., Liang, D., Ashley, E. and Zou, J., 2020. Video-based AI for beat-to-beat assessment of cardiac function. *Nature*, 580(7802), pp.252-256.

[26] Donahue, J., Hendricks, L., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K. and Darrell, T., 2017. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4), pp.677-691.

[27] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[28] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z., 2016. Rethinking the Inception Architecture for Computer Vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[29] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K., 2017. Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[30] Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A., 2017. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 31(1).

[31] TensorFlow. 2021. Tensorflow Core | Machine Learning For Beginners And Experts. [online] Available at: <https://www.tensorflow.org/overview> [Accessed 20 January 2021].

[32] Kingma, D. and Ba, J., 2015. Adam: A Method For Stochastic Optimization. Published as a conference paper at ICLR 2015, pp.1-15.

[33] Cheng, K., Monaghan, M., Kenny, A., Rana, B., Steeds, R., Mackay, C. and van der Westhuizen, D., 2018. 3D echocardiography: benefits and steps to wider implementation. British Journal of Cardiology.