Real-time Standard View Classication in Transthoracic Echocardiography using Convolutional Neural Networks

Andreas Østvik, Erik Smistad, Svein Arne Aase, Bjørn Olav Haugen, and Lasse Løvstakken

Citation:
Andreas Østvik, Erik Smistad, Svein Arne Aase, Bjørn Olav Haugen, and Lasse Løvstakken (2018) Real-time Standard View Classication in Transthoracic Echocardiography using Convolutional Neural Networks. Ultrasound in Medicine & Biology, 2018, DOI: https://doi.org/10.1016/j.ultrasmedbio.2018.07.024

Real-time Standard View Classification in Transthoracic Echocardiography using Convolutional Neural Networks

Andreas Østvik[a], Erik Smistad[a,b], Svein Arne Aase[c], Bjørn Olav Haugen[a], and Lasse Lovstakken[a,*]

[a]*Norwegian University of Science and Technology, Faculty of Medicine and Health Sciences, Department of Circulation and Medical Imaging, Trondheim, Norway*
[b]*SINTEF Medical Technology, Trondheim, Norway*
[c]*GE Vingmed Ultrasound AS, Horten, Norway*

Abstract

Transthoracic echocardiography examinations are usually performed according to a protocol comprising different probe postures providing standard views of the heart. These are used as a basis when assessing cardiac function, and it is essential that the morphophysiological representations are correct. Clinical analysis is often initialized with the current view, and automatic classification can thus be useful in improving today's workflow. In this article, convolutional neural networks (CNNs) are used to create classification models predicting up to seven different cardiac views. Data sets of 2-D ultrasound acquired from studies totaling more than 500 patients and 7000 videos were included. State-of-the-art accuracies of (98.3±0.6)% and (98.9±0.6)% on single frames and sequences, respectively, and real-time performance with (4.4±0.3) ms per frame was achieved. Further, it was found that CNNs have the potential for use in automatic multiplanar reformatting and orientation guidance. Using 3-D data to train models applicable for 2-D classification, we achieved a median deviation of (4±3)° from the optimal orientations.

Keywords: Transthoracic echocardiography, Standard view classification, Convolutional neural network, Deep learning

Introduction

Transthoracic echocardiography (TTE) is widely used for assessment of cardiac function. The examinations are usually performed according to protocols involving different probe postures providing several views of the heart (Lang et al., 2015). Image quality varies substantially between patients and is operator dependent, which increases inter-observer variability and decreases the feasibility of detailed quantitative measurements in the clinic. Cardiac view classification (CVC), that is, determining the image plane through the heart, is the essential first interpretation step in any TTE examination. Clinical implementation of automatic solutions is currently limited, but we believe it could affect several elements of everyday practice.

Finding valid cardiac views has traditionally been difficult for apprentices. The European Association of Echocardiography recommends a minimum of 350 examinations to acquire basic competence for standard TTE (Popescu et al., 2009). Together with the requirement for expert resources, didactic tools using real-time CVC can potentially reduce this number by providing standardization through active quality assurance and probe alignment guidance. Further, a new group of users are adopting echocardiography through the introduction of hand-held devices, making ultrasound (US) more available in general. An implementation with low hardware requirements can be used on such devices and thus provide support in point-of-care situations where cardiologists normally are absent (Morris, 2015).

Tools used when diagnosing cardiac diseases are of-

*Corresponding Author: Andreas Østvik, NTNU, Faculty of Medicine and Health Sciences, Department of Circulation and Medical Imaging, 7491 Trondheim, Norway. E-mail: andreas.ostvik@ntnu.no.

Conflict of Interest: A.Ø, E.S and B.O.H holds positions at CIUS, a center of research-based innovation funded by the Research Council of Norway (Grant 237887) and industry from 2016 to 2022.

ten initialized with specification of current view, and in most cases this must be done manually by the operator. Automatic classification can improve the workflow and adaptivity of quantitative tools and allow continuous scanning and on-site analysis of several quantitative parameters without pushing a single button on the ultrasound scanner. In addition, such a solution could enhance user experience in 3-D US acquisitions by improving automatic extraction of relevant 2-D image planes from volumes (Lu et al., 2008).

Finally, CVC can also complement patient database archives by automatically labeling recordings and thus enable better search functionality, data mining and categorization utilities. In turn, this could, for example, improve follow-up by automatically extracting corresponding views at different stages of patient care.

Related work and state of the art

Penatti et al. (2015) reviewed cardiac view classification for TTE up to 2013. Most studies consider a selection of three or four of the most common cardiac views: apical two chamber (A2C), apical four-chamber (A4C) and apical long-axis (ALAX), as well as the parasternal long-axis (PLAX) and parasternal short-axis (PSAX). Some consider additional views, such as apical five chamber, subcostal four-chamber (SC4C) and vena cava inferior (SCVC), together with a class for unknown data. Examples of relevant views are shown in Fig. 1.

Prior studies claim to achieve overall accuracies as high as 98% on image sequences, such as reported by Wu et al. (2013). In general, inclusion of more views have reduced accuracy considerably. To the best of our knowledge, Park et al. (2007) reported the largest data set, containing 1080 and 223 image sequences for training and validation, respectively.

Most previous studies have used a support vector machine classifier on features extracted with various methods. Recently, deep convolutional neural networks (CNNs) have had great success in many image classification tasks (LeCun et al., 2015). As opposed to traditional machine learning approaches with hand-crafted features, these methods learn both the feature extraction and classification directly from the training data. After Krizhevsky et al. (2012) won the annual ImageNet challenge (ILSVRC) in 2012 using a CNN (Russakovsky et al., 2015), it has become the predominant approach for solving computer vision and recognition tasks.

CNNs have attracted significant attention from the US image analysis community, where hand-crafting generic features can be difficult. Chen et al. (2015) was among the first to report use of CNNs for US view classification, more specifically for locating the fetal abdominal standard plane. Currently, a body of related work in the domain of fetal US image classification exists that methodologically also uses CNNs (Baumgartner et al., 2017; Huang et al., 2017; Bridge et al., 2017). In addition, much research for TTE currently involves the use of CNNs. Perrin et al. (2017) and Narula et al. (2016) have used it for evaluation of cardiac function. Abdi et al. (2017) used it to automatically assess the quality of up to five views using a regression-based recurrent approach. Recently, Gao et al. (2017) used CNNs for classifying eight different cardiac views using a method fusing hand-crafted and learned features. Their database consisted of 432 image sequences, and they achieved an average accuracy of 92.1% validating on 152 image sequences.

Main contributions

In the work described here, our aim was to develop fully automated and robust methods for real-time CVC using CNNs and facilitate their use in a clinical setting. We also investigated the potential for applying these methods to automatic extraction of 2-D views from 3-D volumes and orientation guidance for finding optimal views in 2-D US. Compared with previous studies, the contributions of this paper are as follows:

- Annotation and training on significantly more patient data than previously included, with extensive patient-based cross-validation and testing ensuring unbiased results
- Consideration of up to seven of the most common cardiac views: A2C, A4C, ALAX, PSAX, PLAX, SC4C and SCVC, in addition to a class for unknown data
- Analysis of two common network topologies and a proposed network design based on recent work in the field with the aim of being both accurate and effective

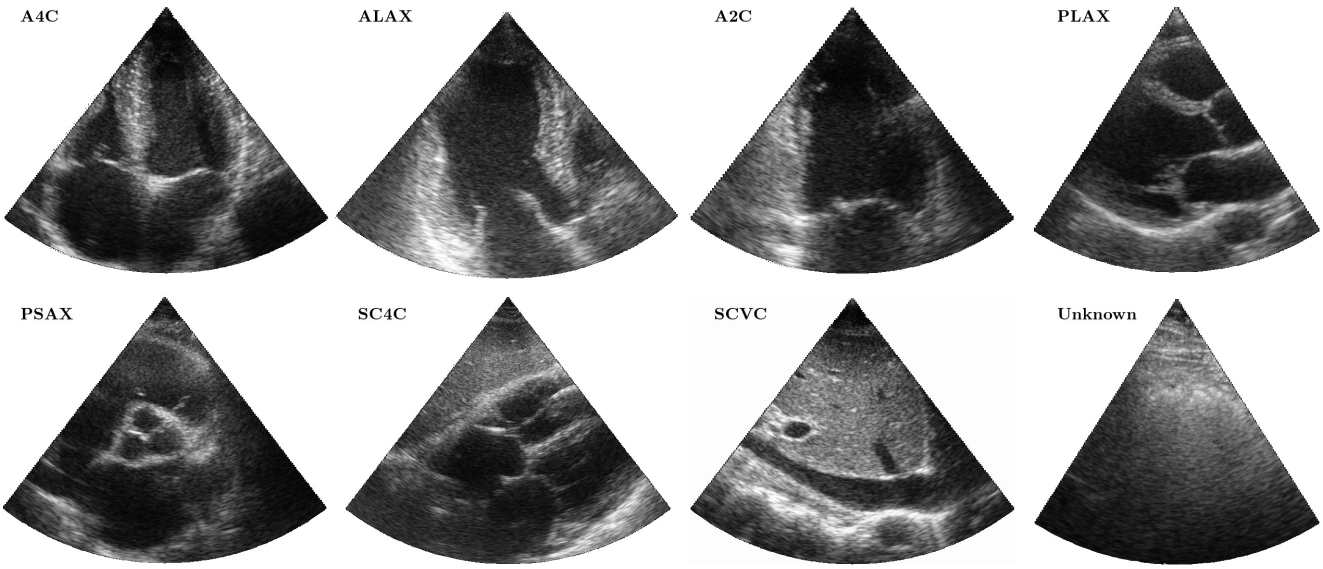

Figure 1: Seven cardiac views in transthoracic echocardiography obtained in arbitrary stages of the heart cycle. Examples of the apical four chamber (A4C), long-axis (ALAX), two chamber (A2C), parasternal long axis (PLAX), short-axis (PSAX), subcostal four-chamber (SC4C) and vena cava inferior view (SCVC) is illustrated, in addition to a nonassignable sample labeled unknown.

- Experiments on orientation guidance for finding optimal apical views and a comparison between models trained with either 2-D or 3-D data
- Analysis of computational requirements and performance

Convolutional neural networks

Three CNNs were investigated for cardiac view classification. Compared to the problems in which typical image classification networks are designed, we consider CVC easier. The consensus on increasing network depth to achieve better results does not necessarily hold for such tasks, and we believe that competitive performance can be achieved with less complex networks. We therefore address this issue by combining observations from relevant work and propose a network that aims to balance the accuracy and effectiveness for this use.

For extensive details of the investigated networks, the reader is referred to relevant articles (Krizhevsky et al., 2012; Szegedy et al., 2016). Herein, we introduce them briefly and emphasize their differences and our changes to the original topology. Furthermore, we accentuate the background of our design choices.

AlexNet architecture

The winner of ILSVRC 2012 is a CNN referred to as AlexNet. It is a simple feed-forward network with five blocks of convolutional layers followed by rectified linear activation units (ReLU) and maximum pooling. Local response normalization is used after the first two convolution layers. The final part of the network is composed of two fully connected layers with ReLU and dropout regularization, whereas the final classifier is a fully connected layer followed by softmax activation.

Compared with the original topology, the local response normalization layers were removed for this study, and batch normalization(Ioffe and Szegedy, 2015) was used instead for additional regularization, as suggested by Canziani et al. (2016).

Inception architecture

Some of the most influential proposals after AlexNet came from the authors of "Network In Network" (NIN) (Lin et al., 2013), who suggested using bottlenecks (e.g., convolutions with kernel size 1×1) to combine features between layers. The key insights from their article inspired Szegedy et al. (2015) to create the Inception architecture (introduced as GoogleNet). The principal difference from other networks is the building blocks, referred to as Inception modules. Each block consists of parallel routes of convolutions with varying kernel size, in addition to a pathway with pooling. Fig. 2 is a schematic of a typical module with bottlenecks and batch normalization. Several editions of Inception have been pre-

sented since its introduction, but the fundamental philosophy of parallel routes in depth is the same.

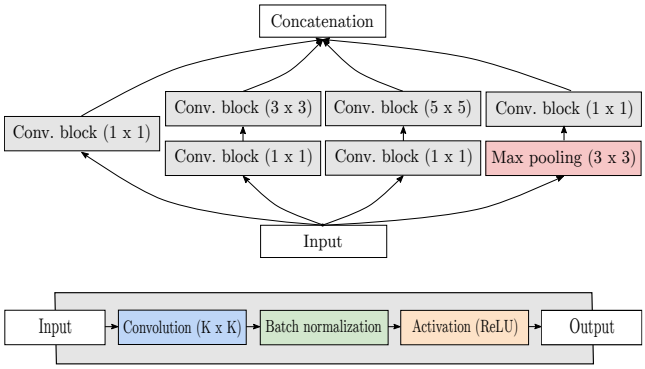

Figure 2: The inception module is a combination of parallel convolution blocks with different kernel sizes and a pooling branch concatenated into a single output. Each convolution block (Conv.) consists of convolutions followed by batch normalization and non-linear activation.

In this study, the third edition of the Inception architecture (Szegedy et al., 2016) was employed. The major architectural difference compared with the original topology is spatial factorization of large spatial filters. On this basis, three different modules were designed and used throughout the network. In the lower parts, where the feature maps are relatively large, the module is similar to that in Fig. 2, except that the (5×5) convolutions are factorized into two layers of (3×3) convolutions. The other modules use asymmetric convolutions, for example, a (3×1) followed by a (1×3) convolution. In addition to factorization, batch normalization is used after convolution layers. Here we use smaller input images than intended for this architecture, and to allow better information flow and avoid convolution filters larger than the feature maps, we removed the second max-pooling layer.

Cardiac view classification architecture

The network we propose resembles that in the discussed work and employs a combination of introduced concepts. Fig. 3 is an overview of the architecture. The fundamental building blocks consist of convolutions, batch normalization (Ioffe and Szegedy, 2015) and non-linear activation units. Batch normalization was added to speed up training by allowing higher learning rates and avoiding use of network resources to compensate for outlying filter weights during backpropagation. Parametric rectified linear units (PReLU) were chosen as the activation unit in all blocks (He et al., 2015). Compared with the frequently used ReLU, which is zero for negative values, PReLU allows non-zero gradients for inactive units. The negative part is a linear function with a learned slope.

Initially, input is propagated through two component blocks with (3×3) convolution kernels, followed by max pooling. The first and second convolution layer have 16 and 32 filters, respectively. We use pooling of size (2×2) and equal strides to downsample without overlap. After the second pooling layer, data are processed through an Inception module with three parallel routes. Each route consists of a bottleneck, two of which were followed by blocks with larger convolution kernels, (3×3) and (5×5), respectively. This is equivalent to the module in Fig. 2 without the pooling route. The bottlenecks in the Inception module reduce the number of filters by 25%, 25% and 50% in the order of small to large convolution kernels, respectively. Furthermore, the number of filters is increased by 25% in the following convolution block.

Inspired by the connection scheme in DenseNet (Huang et al., 2016), the input of the Inception module is concatenated with the output and processed into a transition module with bottleneck and max pooling. This step is repeated three times, and as emphasized by Redmon and Farhadi (2016) in the base classifier of the YOLO object detection system, we doubled the number of filters before every new pooling layer. As opposed to their implementation, we control this behavior in the bottleneck of the transition block. The dense connectivity pattern further alleviates the vanishing gradient problem, and perhaps more importantly, it can enhance feature propagation and reusability.

After the third transition, the data are processed through two Inception blocks with a constant number of filters and no pooling. The route with (5×5) convolution kernels was omitted in these modules, and dropout regularization was used between them. The final classification block consisted of a compressing convolution layer with 11 kernels and number of filters equal to the class count. This was activated with another PReLU, before features were spatially averaged and fed into a softmax activation as in NIN. The spatial pooling replaces the more typical fully connected layers. This reduces the parameter count and, it is also claimed, makes the network less vulnerable to overfitting (Lin et al., 2013).

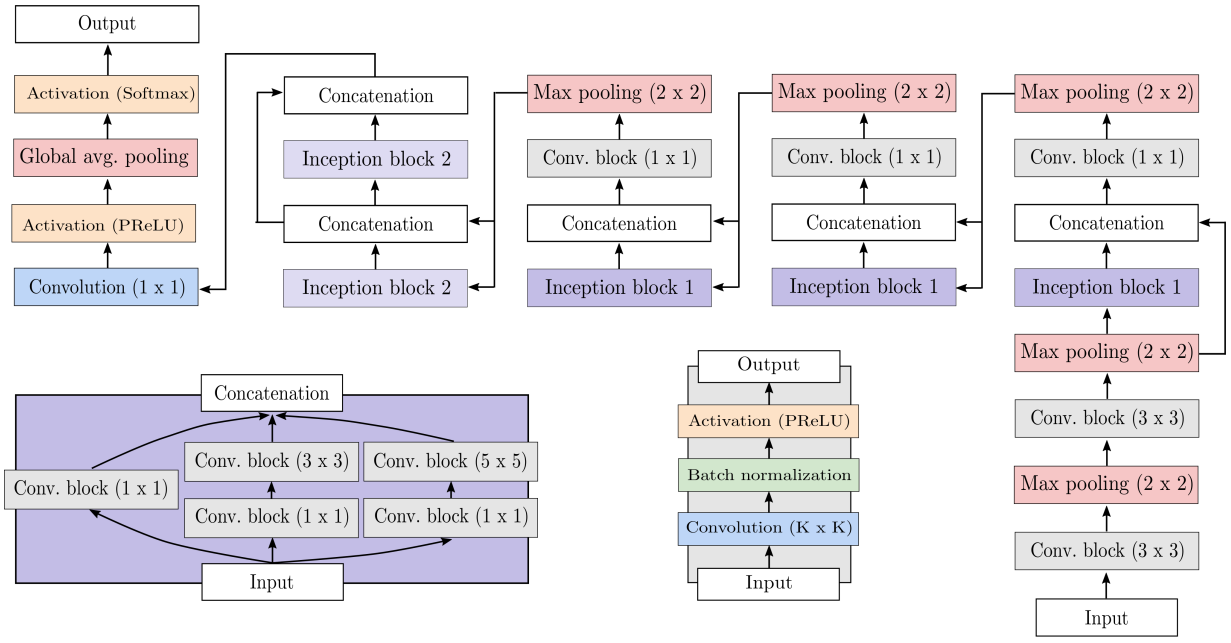

Figure 3: Schematic of the proposed network architecture used for cardiac view classification. Convolution blocks (gray boxes) are composed of convolutions, batch normalization and PReLU activations. Two versions of the Inception module are employed: the illustrated one being used in the lower part of the network (dark purple) and a simplified one without the (5×5) route in higher parts of the network (bright purple). The final classifier block consist of another compressing convolution layer with kernel size (1×1) and filter amount equal to the number of views. The output is activated with a PReLU layer. Finally, global average pooling followed by softmax activation yields a prediction vector as output.

Experimental setup

Experiments were divided into two parts. First, training and evaluation on annotated 2-D data were conducted using three different CNNs: AlexNet, Inception and the proposed CVC architecture. Afterward, 2-D data extracted from 3-D volumes were included and used to train new models using the CVC architecture. Three-dimensional data were then evaluated, together with a comparison between the models trained with the same architecture on 2-D data.

Database and annotation

Three different data sets of anonymous US data were included in this study. All data originated from patient studies approved by the Regional Committee for Medical Research Ethics and conducted according to the Helsinki Declaration. Written informed consent was obtained from all patients. The sample data are considered representative of a regular cardiological clinic and give a distribution of both healthy and ill participants in the relevant age groups.

2D US image sequences

he first data set consists of 4582 US videos with varying numbers of frames from 205 patients. Acquisition was performed by three senior cardiologists according to a standard protocol for echocardiography using a GE Vivid E9 US scanner (GE Vingmed Ultrasound, Horten, Norway) with a GE M5S phased-array transducer. Fifty-six of the patients were diagnosed with systolic or diastolic cardiac dysfunction. The population age ranged from 20 to 91 years with an average age of 64 years. The second data set was randomly drawn from the Nord-Trøndelag Health Study (HUNT) population study (Dalen et al., 2009) and consisted of 2559 US videos from 265 subjects. Acquisition was performed by one senior cardiologist according to the same protocol using a GE Vivid 7 scanner with a GE M5S phased-array transducer. All subjects were free from known cardiac dysfunction, and the population had an average age of 49 years.

The videos were annotated manually and categorized into seven different classes: A4C, ALAX, A2C, PLAX, PSAX, SC4C and SCVC. Subcostal acquisitions were not included in the HUNT study. Fig. 4. summarizes the data indicating the class balance. Non-assignable images were labeled unknown, but the number was not considered sufficient for training relative to the other classes. Thus, samples from a laboratory experiment with the goal of acquiring

arbitrary US images without clinical relevance were added. The total was 41,450 images from 460 videos.

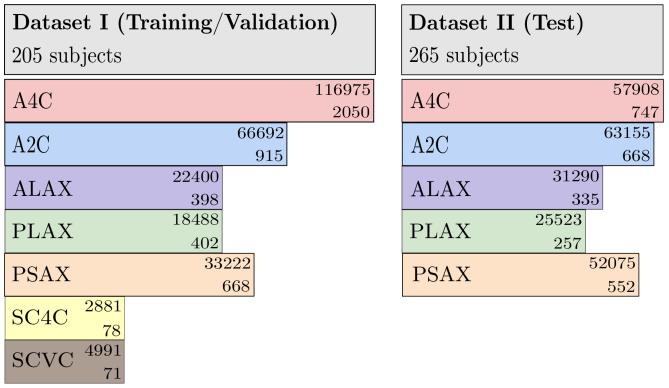

Figure 4: Overview of the two 2-D data sets. The upper value is the number of frames in the given class, and the lower value is the number of videos.

Considerable variations in image quality were discovered, and in a parallel annotation task, the images were labeled as poor, acceptable or good. The relative distribution labeled by an expert cardiologist pre-analysis was (32%, 41%, 27%) from poor to good respectively.

3D US volume sequences

The 3-D data set consists of 60 anonymous US volumetric exams with varying numbers of volumes from the same number of patients. Acquisition was performed by two senior cardiologists by placing the probe in the apical position using a GE Vivid E9 US scanner with a 3V 4D sector array transducer probe.

Data were generated by extracting 2-D images, or planes, from the 3-D volume around a fixed depth axis placed in the frustum center. This mimicks the scenario of rotating a 2-D probe in an apical position, generating all possible views oriented with respect to the depth axis. Here we extracted one frame per degree, yielding a total of 360 images per volume. Eyeballing and simple caliper measures were used to choose three different frames (angles) from each volume as optimal apical views (A4C, A2C, ALAX). These angles were further used as the peak of an asymmetric Gaussian weighting when labeling the data. The tail of the Gaussian label l is determined by the distance between adjacent peaks and is given by

$$l^{\text{view}}_{\leftarrow,\rightarrow} = \exp\left\{-\left(\frac{\Delta\theta_{\leftarrow,\rightarrow}}{\sqrt{2}\sigma_{\leftarrow,\rightarrow}}\right)^2\right\}. \quad (1)$$

Here, $\Delta\theta_{\leftarrow,\rightarrow}$ is the angular distance from the peak of a specific view in a given direction. The standard deviation is the fractional distance to the nearest adjacent peak in either direction, that is, $\sigma_{\leftarrow,\rightarrow} = |\theta^{\text{view}} - \theta^{\text{adj.view}}_{\leftarrow,\rightarrow}|/3$. An example annotation with reference to the 17-segment left ventricle model (Lang et al., 2015) is provided in Fig. 5. This annotation scheme was chosen to allow a connection between adjacent peaks. Unlike a binary classification, this enables a more robust transition region between optimal views and may be more suitable for orientation guiding and quality assurance while scanning. It could also be used to extract the desired 2-D planes automatically from 3-D volumes.

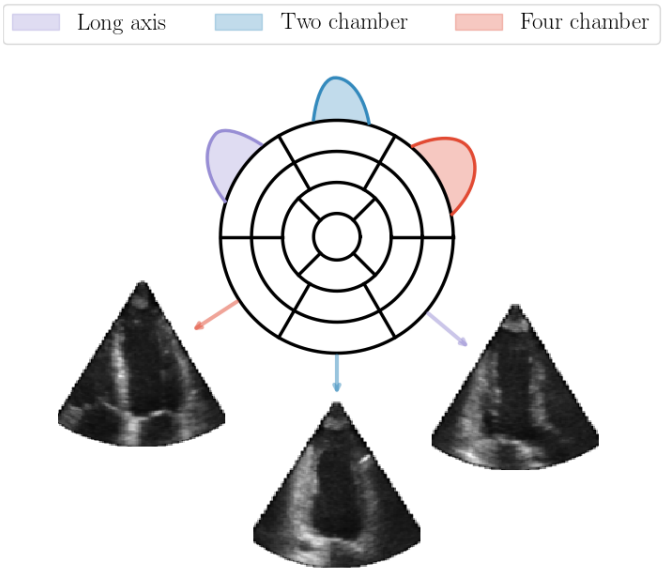

Figure 5: Sketch of an example annotation with reference to the left ventricle segment model (17 divisions). The curves correspond to the confidence label of a specific cardiac view, where higher values suggests optimal orientation.

Preprocessing

The data were scan converted from beamspace data stored in DICOM format. The 3-D data were stitched when necessary. No image enhancement filters were applied. For training, the images were intensity normalized and downsampled to a size of (128 × 128) pixels. No data augmentation was applied.

Learning details

Training was performed over a maximum of 100 epochs using mini-batch gradient descent with a batch size of 64. In machine learning, one epoch is defined as a complete pass of training data, whereas

the batch size is the number of examples shown for each weight update. We used the categorical cross-entropy and mean absolute error (MAE) loss functions (Goodfellow et al., 2016)for training on 2-D and 3-D data, respectively. An adaptive moment estimation method for stochastic optimization named Adam (Kingma and Ba, 2014) was used with a maximum learning rate of 10^{-4}. Uniform Glorut initialization (Glorot and Bengio, 2010) was used on the convolution layers before training. The model was evaluated on unknown data between epochs, where the best model was saved underway. The data were fully shuffled after every epoch. To avoid unnecessary training time and overfitting, early stopping routines based on validation accuracy were used with a patience of 20 epochs.

As seen in Fig. 4, the training data are clearly unbalanced, with a ratio of 1:29 between the least and most represented class. To combat possible bias toward high representations, the training data were downsampled before every new epoch by randomly drawing frames from each US acquisition based on its ratio compared with the least represented class. This allows training on equal amounts of data from each class and every epoch; by performing this on a per-sequence basis, representations from each acquisition are also included. Note that we still use the term epoch, although it breaks the definition of passing the entire dataset.

To setup the learning environment, the framework Keras was utilized with Tensorflow (Abadi et al., 2015) as backend. Experiments was carried out on a workstation installed with an Ubuntu 16.04 operating system. The hardware consisted of an Intel Core i7-6820HK CPU with a clock speed of 4.10 GHz, 32 GB RAM and a NVIDIA GeForce GTX 1070 GPU with 8GB of memory.

Methods and metrics for evaluation

A 10-fold patient-based cross-validation technique was performed, separating the first data set into training and validation partitions. For each run, this corresponds to omitting 20 or 21 patients from the 2-D data. The same was done for the 3-D data, in which each fold consists of six patients. Such patient-based model validation will give a better impression of the expected results on new patient data. To the best of our knowledge, this is the first publication on the topic in which patient-based cross-validation is extensively used. To further evaluate the model we included an independent data set for testing purposes only.

In addition to accuracy, validation metrics such as precision and recall were used because of the imbalanced class frequency in the 2-D data. They are defined as $TP/(TP+FP)$ and $TP/(TP+FN)$, respectively, where TP is the true positives, FP the false positive and FN the false negative. The model accuracy is defined as the ratio of true predictions to all predictions.

Further, for validation on 3-D data, the MAE is calculated over the angle interval for all volumes of every subject as

$$\mathrm{MAE} = \frac{\sum_{\theta=0}^{\theta_{\max}} |l_\theta^{\mathrm{true}} - l_\theta^{\mathrm{pred}}|}{\theta_{\max}}, \tag{2}$$

where the angle $\theta \in [0, \theta_{\max}] = [0, 2\pi)$, and l_θ^{true}, l_θ^{pred} is the true and predicted labels for a given angle. In addition, we performed a qualitative inspection comparing the predictions to ground truth by visualizing them together.

To determine the classification time per incoming image in a deployed setting, an experiment in which images are loaded individually in a loop and classified with the trained models was conducted. This mimics a clinical scenario in which frames are acquired and processed one by one. A total of 30,000 images were processed for each experiment, and for every model we investigated the change in inference time using the GPU. As a hardware invariant measure for runtime, the number of floating point operations was added. This was calculated using the profiler tool released through the Tensorflow framework.

Together with the network definition, the storage requirements are determined mainly by the number of parameters needed to initialize the network. This number is calculated using the Keras framework.

Results

Analysis on 2D data

Experimental results from patient-based cross-validation using three different network topologies trained on 2-D data are given in Table. 1. The trained models were tested on an independent and unknown test set, yielding the results outlined in Table. 2. The sequence validation was performed using a majority vote approach. The CVC model yielded competitive results despite having significantly fewer learned parameters. Compared with the other models, the

model variance is lower for the CVC network. Low average inference time per image was achieved for all networks using the frameworks Tensorflow and FAST (Smistad et al., 2015). This was without any emphasis on inference optimization. Utilizing the GPU, the CVC network classifies approximately 230 frames per second, while AlexNet manages twice that amount. This was without any emphasis on inference optimization. Using the GPU, the CVC network classifies approximately 230 frames per second, whereas AlexNet manages twice that number. This is well within the limits of real-time view classification in this context.

To the best of our knowledge, the results surpass current state of the art on 2-D B-mode data and indicate that neural networks are well suited for ultrasound view classification tasks. Accessible benchmark data would be needed for a proper comparison with related work, but it is believed that the diversity and size of the data set used in this study at worst yield an equal baseline.

Analysis on 3D data

The averaging of MAE over all volumes of every subject is illustrated in Fig. 6. This is calculated from the classification of 360 angles/images per volume. The worst and best cases are indicated, together with the medians of all subjects. The predictions and the ground truth from these cases are illustrated in Fig 7. The median MAE of all subjects was (3.8±2.4)%, and the median deviation from true to predicted peak was $(4 \pm 3)°$. The median MAE using the CVC model trained on 2D data only was $(11.3 \pm 9.7)\%$.

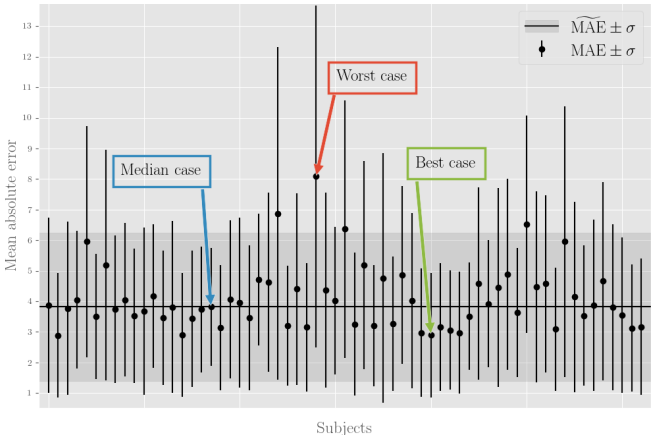

Figure 6: Mean absolute error (MAE) values of all subject volumes. The median MAE of all subjects is given by the horizontal line.

Discussion

Technical considerations

Patient-based cross-validation indicates that the CVC network is best in terms of relevance and accuracy metrics. The standard deviation is almost halved for the cross-validation models, and it has the smallest parameter space. Testing on an independent and unknown data set also suggests better generalization. Excluding the subcostal and unknown views, the overall accuracy from cross-validation is $(98.1 \pm 0.7)\%$, making the test results within the calculated variation. The slight and consistent underestimation can have multiple origins; for instance, it could be a small degree of overfitting toward the training/validation data set (e.g., scanner, probe and operators and their preferences). The trained models would probably benefit from a broader representation domain.

Compared with AlexNet, the other networks have smaller receptive fields and less coarse downsampling and, at least for the first layers, preserve more pixel information from the input image. On the other hand, less expressiveness is captured in the learned features. Though the Inception module can retain this to some extent by having a route with a semi-large kernel, it may seem that adding features benefits generalization more in this scenario. Though it is hard to pinpoint the specific reason why CVC models surpass the results of the other networks on this task, we believe that the combination of Inception modules, dense connections, activation, bottlenecks and number of features (more than AlexNet, fewer than Inception) strengthens the generalization.

The subcostal window proves to be the hardest to classify; arguably, lack of training data is the probable cause. Even if this is the driving factor of these algorithms, still views with distinct characteristics tend to simplify the classification. For example, in Fig. 1, we see that the parasternal views seem to have more interclass variance than the apical views and have a higher success rate on unknown data despite learning from less.

Image quality is dependent on the acquisition environment and setup: the parameters used on the scanner, expertise of the physician and status of patient morphophysiology. On an abstract level, this information is embedded into the sequences from a specific patient, and by omitting the use of patient-based validation, the model would gain a fictitious

Table 1: Experimental results from cross-validation on dataset I using three different network topologies. Validations are per single frame and image sequence (in parenthesis for precision and recall). Bold metric indicate best score. Runtime measurements, number of floating point operations and trainable parameters are also given.

(a) AlexNet with BN

	Precision (%)	Recall (%)
A4C	97.7 (98.4)	96.0 (97.6)
ALAX	92.1 (95.0)	95.9 (97.7)
A2C	94.7 (96.5)	96.2 (97.0)
PLAX	96.4 (97.6)	98.1 (99.0)
PSAX	95.7 (97.8)	95.2 (96.7)
SC4C	88.4 (93.6)	96.8 (97.3)
SCVC	94.3 (98.5)	92.2 (94.4)
Unknown	99.2 (95.8)	98.7 (**100.0**)
Overall accuracy(%):		
Frame		96.4 ± 1.2
Sequence		97.5 ± 1.3
Runtime [ms]:		
GPU		$\mathbf{2.0 \pm 0.2}$
CPU		$\mathbf{8.1 \pm 0.2}$
Operations [GFLOPS]:		**0.25**
Parameters:		~20.6M

(b) Inception ver. 3 (Modified)

	Precision (%)	Recall (%)
A4C	97.9 (98.8)	97.8 (99.0)
ALAX	96.8 (99.0)	95.6 (96.2)
A2C	96.5 (**97.5**)	96.7 (98.2)
PLAX	97.0 (97.8)	98.4 (99.5)
PSAX	96.5 (98.6)	97.0 (97.6)
SC4C	92.6 (**96.1**)	96.3 (94.9)
SCVC	97.7 (100.0)	92.9 (**95.8**)
Unknown	99.1 (99.1)	98.8 (**100.0**)
Overall accuracy(%):		
Frame		97.4 ± 1.1
Sequence		98.5 ± 0.8
Runtime [ms]:		
GPU		10.7 ± 0.6
CPU		20.4 ± 0.5
Operations [GFLOPS]:		1.45
Parameters:		~21.8M

(c) Proposed CVC Network

	Precision (%)	Recall (%)
A4C	**98.5** (**99.0**)	**98.5** (**99.3**)
ALAX	**98.1** (**99.2**)	**96.2** (**98.0**)
A2C	**96.9** (**97.5**)	**97.8** (**98.3**)
PLAX	**98.5** (**99.5**)	**99.1** (**100.0**)
PSAX	**98.7** (**100.0**)	**97.9** (**98.2**)
SC4C	**92.7** (94.0)	**99.1** (**100.0**)
SCVC	**99.4** (**100.0**)	**95.3** (94.4)
Unknown	**99.6** (**99.8**)	**99.6** (**100.0**)
Overall accuracy(%):		
Frame		$\mathbf{98.3 \pm 0.6}$
Sequence		$\mathbf{98.9 \pm 0.6}$
Runtime [ms]:		
GPU		4.4 ± 0.3
CPU		15.9 ± 0.4
Operations [GFLOPS]:		0.80
Parameters:		~**10.6M**

Table 2: Experimental results on test dataset II using three different network topologies. Validations are per single frame and image sequence (in parenthesis for precision and recall). Bold metric indicate best score.

(a) AlexNet with BN

	Precision (%)	Recall (%)
A4C	93.7 (96.0)	99.3 (99.7)
ALAX	97.3 (99.0)	90.7 (93.1)
A2C	95.4 (96.4)	93.1 (95.2)
PLAX	93.1 (96.6)	98.3 (99.4)
PSAX	98.9 (99.7)	95.9 (98.3)
Overall accuracy(%):		
Frame		95.5 ± 0.7
Sequence		97.3 ± 0.6

(b) Inception ver. 3 (Modified)

	Precision (%)	Recall (%)
A4C	94.7 (96.7)	99.5 (99.8)
ALAX	97.7 (99.1)	90.0 (92.2)
A2C	95.1 (96.1)	94.3 (96.0)
PLAX	94.6 (96.6)	98.9 (99.6)
PSAX	99.1 (99.7)	96.9 (98.3)
Overall accuracy(%):		
Frame		96.1 ± 1.6
Sequence		97.5 ± 1.4

(c) Proposed CVC Network

	Precision (%)	Recall (%)
A4C	**96.2** (**97.8**)	**99.6** (**99.8**)
ALAX	**98.6** (**99.5**)	**93.1** (**95.3**)
A2C	**96.6** (**97.6**)	**96.0** (**97.4**)
PLAX	**97.5** (**99.3**)	**98.7** (**99.7**)
PSAX	**99.4** (**99.9**)	**98.3** (**99.5**)
Overall accuracy(%):		
Frame		$\mathbf{97.4 \pm 0.6}$
Sequence		$\mathbf{98.5 \pm 0.5}$

advantage in predicting allegedly unknown data. Examples of poor images from the data set are provided in Fig. 8, where the model has predicted the views as indicated under every image. The variation in quality from Fig. 1 is apparent, and we discover that the model has more conflicts with ground truth when images are poor. Of 54 misclassified sequences in the cross-validation, 42 were classified in this category, whereas the remainder were acceptable.

Another interesting observation is that the images in Fig. 8 were acquired from two different patients and amount to approximately 15% of the total sequential error. By observation, sequences from the patient shown in the upper part of the figure have an abnormal artifact in the left ventricle. The other patients generally have noisy and virtually invisible structures. Both types of issues can cause classification problems and could potentially be present in all image sequences from a specific patient. By distributing sequences from the same patient in both the training and validation data sets, the model could effectively adapt to the irregularity. Patient-based cross-validation and independent tests should thus be emphasized when assessing results from generated

(a) Worst - MAE = $(8.1 \pm 5.6)\%$ (b) Median - MAE = $(3.8 \pm 1.9)\%$ (c) Best - MAE = $(2.9 \pm 2.0)\%$

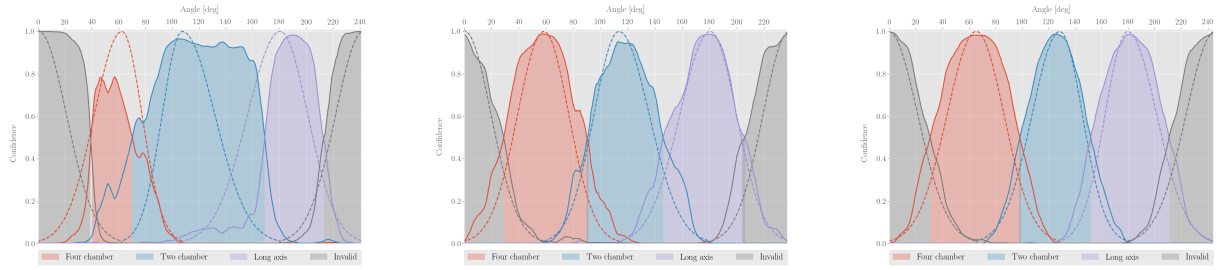

Figure 7: Evaluation on 2D images extracted from 3D volumes with orientation angle with respect to the depth axis. The models used are trained with data from the 3D volumes. The dotted curves correspond to the assigned labels, while the filled curves is the model predictions.

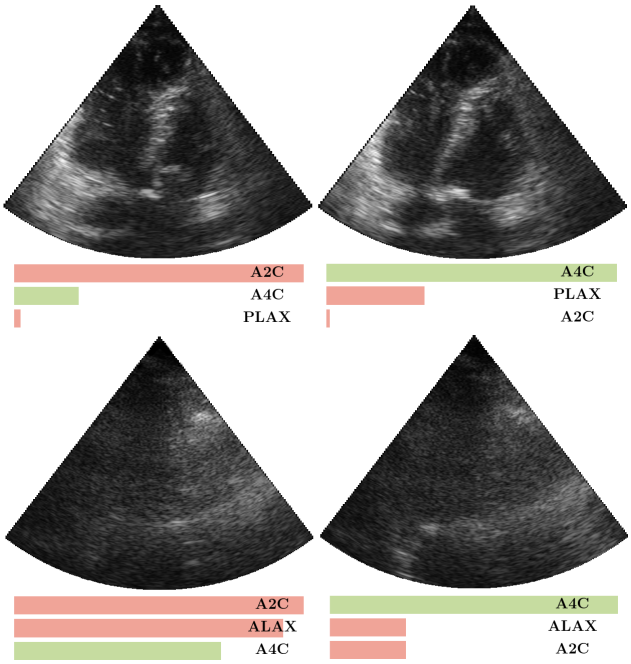

Figure 8: Example of poor cardiac ultrasound images from two different apical four chamber sequences classified by the proposed view detection model. Green label indicates the ground truth label, and the size corresponds to the fractional prediction of the model. The left side shows frames where the model has conflicts with ground truth.

models.

Compared with other work, our results seem promising. Methods and potential applications have some overlap with the research conducted by Abdi et al. (2017) on quality assessment of cine loops. However, their multistream regression network required 20 consecutive image frames to assess one label; to discriminate between views, every frame had to be passed through a shared layer architecture and into five different view-specific layers. This could be feasible for distinguishing views because Abdi et al. state that it is in real time, but their focus is on quality assessment of a given view; classification is not investigated.

In three dimensions, annotation of optimal views was difficult becaue variations in image features were insignificant for small angle intervals. This held especially for the four- and two-chamber views, whereas the long-axis view was easier because the diameter of the left ventricular outflow tract could be used as a reference in most cases. With this in mind, we still achieve a low median deviation of the predicted to true peaks in all patients, and by inspecting Fig. 7, we argue that the long-axis view appears more robust. In general, models trained with 3-D data achieve a low MAE. The performance of models trained with 2-D data, as expected, experiences more fluctuation, and it can be difficult to detect the optimal view. The reason might be variations in image quality and views slightly off orientation. The latter are not distinguished in the 2-D data set, as we assumed that every examination contains the best possible view for every patient. Results could therefore be expected to have a saturating behavior around the optimal view.

Clinical perspective

As stated in the Introduction, automatic CVC has several clinical applications, such as improving workflow, enabling more automation and guiding inexperienced users. The results on the second independent data set in this study indicate that the accuracy of the proposed CVC methods based on CNNs is real even for data acquired with other scanners and by different operators. This accuracy, together with the measured low runtime and the real-time video, suggests that this method is ready for further testing in a

clinical setting. Development and end-to-end fashion allow low threshold deployment and applicability in many settings without any tuning or in-depth knowledge of the methods. No parameters are required; only an input image is needed. Results also indicate that including training data from heart volumes can improve guiding utilities and quality assurance while scanning. Despite this, models trained with 2-D data will probably be better suited for database utilities, such as data mining, search and categorization. It is easier to add more views, and the accuracy is very high.

Training opportunities for new health care personnel are limited, and expert knowledge is often captivated by workload or centralization. We believe these increasingly relevant problems could be addressed by tools such as automatic CVC. However, separate clinical studies on training effects, standardization and workflow must be induced to support this statement.

Conclusion

In the study described here, different neural networks were investigated for cardiac view classification. State-of-the-art results for standard 2-D echocardiography were achieved. The proposed network had a small number of trainable parameters and achieved real-time inference with high accuracy. Although the demonstration looks robust when training on 2-D data, our initial experiments into apical view guidance based on 3-D data indicated room for further work. Using slices of 3-D volumes for training improved the results significantly, and we believe that further development toward real-time quality assurance and guidance from US images is plausible when including such data.

Conflict of interest disclosure

A.Ø, E.S. and B.O.H. hold positions at the Centre for Innovative Ultrasound Solutions (CIUS), a center of research-based innovation funded by the Research Council of Norway (Grant 237887), and in industry from 2016 to 2022.

Acknowledgments

We gratefully acknowledge Håvard Dalen, Stein Samstad and Ole Christian Mjlstad at the Department of Circulation and Medical Imaging, NTNU and Clinic of Cardiology, St. Olavs Hospital, for acquisition of data used in this project. This work has received funding from the Centre of Innovative Ultrasound Solution, a Norwegian Resarch Council appointed centre for research-based innovation (SFI), Project grant 237887.

References

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mané D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viégas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015.

Abdi AH, Luong C, Tsang T, Jue J, Gin K, Yeung D, Hawley D, Rohling R, Abolmaesumi P. Quality assessment of echocardiographic cine using recurrent neural networks: Feasibility on five standard view planes. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2017. pp. 302–310.

Baumgartner CF, Kamnitsas K, Matthew J, Fletcher TP, Smith S, Koch LM, Kainz B, Rueckert D. Sononet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE transactions on medical imaging, 2017;36:2204–2215.

Bridge CP, Ioannou C, Noble JA. Automated annotation and quantitative description of ultrasound videos of the fetal heart. Medical image analysis, 2017;36:147–161.

Canziani A, Paszke A, Culurciello E. An analysis of deep neural network models for practical applications. arXiv preprint arXiv:1605.07678, 2016.

Chen H, Ni D, Qin J, Li S, Yang X, Wang T, Heng PA. Standard Plane Localization in Fetal Ultrasound via Domain Transferred Deep Neural Networks. IEEE Journal of Biomedical and Health Informatics, 2015;19:1627–1636.

Dalen H, Thorstensen A, Aase SA, Ingul CB, Torp H, Vatten LJ, Stoylen A. Segmental and global longitudinal strain and strain rate based on echocardiography of 1266 healthy individuals: the hunt study in norway. European Journal of Echocardiography, 2009;11:176–183.

Gao X, Li W, Loomes M, Wang L. A fused deep learning architecture for viewpoint classification of echocardiography. Information Fusion, 2017;36:103–113.

Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 2010. pp. 249–256.

Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press, 2016.

He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision, 2015. pp. 1026–1034.

Huang G, Liu Z, Weinberger KQ, van der Maaten L. Densely connected convolutional networks. arXiv preprint arXiv:1608.06993, 2016.

Huang W, Bridge CP, Noble JA, Zisserman A. Temporal heartnet: towards human-level automatic analysis of fetal cardiac screening video. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2017. pp. 341–349.

Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, 2015. pp. 448–456.

Kingma D, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. Advances in neural information processing systems, 2012:1097–1105.

Lang RM, Badano LP, Mor-Avi V, Afilalo J, Armstrong A, Ernande L, Flachskampf FA, Foster E, Goldstein SA, Kuznetsova T, 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. European Heart Journal-Cardiovascular Imaging, 2015;16:233–271.

LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015;521:436–444.

Lin M, Chen Q, Yan S. Network in network. arXiv preprint arXiv:1312.4400, 2013.

Lu X, Georgescu B, Zheng Y, Otsuki J, Comaniciu D. Autompr: Automatic detection of standard planes in 3d echocardiography. In: Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on. IEEE, 2008. pp. 1279–1282.

Morris AE. Point-of-care ultrasound: Seeing the future. Current Problems in Diagnostic Radiology, 2015;44:3 – 7.

Narula S, Shameer K, Omar AMS, Dudley JT, Sengupta PP. Machine-learning algorithms to automate morphological and functional assessments in 2d echocardiography. Journal of the American College of Cardiology, 2016;68:2287–2295.

Park JH, Zhou SK, Simopoulos C, Otsuki J, Comaniciu D. Automatic cardiac view classification of echocardiogram. In: Proceedings of the IEEE International Conference on Computer Vision, 2007.

Penatti OAB, Werneck RdO, de Almeida WR, Stein BV, Pazinato DV, Mendes Júnior PR, Torres RdS, Rocha A. Mid-level image representations for real-time heart view plane classification of echocardiograms. Computers in Biology and Medicine, 2015;66:66–81.

Perrin DP, Bueno A, Rodriguez A, Marx GR, Pedro J. Application of convolutional artificial neural networks to echocardiograms for differentiating congenital heart diseases in a pediatric population. In: Medical Imaging 2017: Computer-Aided Diagnosis. Vol. 10134. International Society for Optics and Photonics, 2017. p. 1013431.

Popescu BA, Andrade MJ, Badano LP, Fox KF, Flachskampf FA, Lancellotti P, Varga A, Sicari R, Evangelista A, Nihoyannopoulos P, et al. European association of echocardiography recommendations for training, competence, and quality improvement in echocardiography. European Journal of Echocardiography, 2009;10:893–905.

Redmon J, Farhadi A. Yolo9000: Better, faster, stronger. arXiv preprint arXiv:1612.08242, 2016.

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 2015;115:211–252.

Smistad E, Bozorgi M, Lindseth F. FAST: framework for heterogeneous medical image computing and visualization. International Journal of Computer Assisted Radiology and Surgery, 2015;10:1811–1822.

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. pp. 1–9.

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. pp. 2818–2826.

Wu H, Bowers DM, Huynh TT, Souvenir R. Echocardiogram view classification using low-level features. In: IEEE 10th International Symposium on Biomedical Imaging, 2013. pp. 752–755.