



UWL REPOSITORY

repository.uwl.ac.uk

Neural architecture search of echocardiography view classifiers

Azarmehr, Neda, Ye, Xujiang, Howard, James P., Lane, Elisabeth S., Labs, Robert, Shun-Shin, Matthew J., Cole, Graham D., Bidaut, Luc, Francis, Darrel P. and Zolgharni, Massoud ORCID: <https://orcid.org/0000-0003-0904-2904> (2021) Neural architecture search of echocardiography view classifiers. *Journal of Medical Imaging*, 8 (03). ISSN 2329-4302

<http://dx.doi.org/10.1117/1.JMI.8.3.034002>

This is the Accepted Version of the final output.

UWL repository link: <https://repository.uwl.ac.uk/id/eprint/8041/>

Alternative formats: If you require this document in an alternative format, please contact: open.research@uwl.ac.uk

Copyright:

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy: If you believe that this document breaches copyright, please contact us at open.research@uwl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Neural Architecture Search of Echocardiography View Classifiers

Neda Azarmehr^{a,*}, Xujiang Ye^a, James P Howard^b, Elisabeth Sarah Lane^c, Robert Labs^c, Matthew J Shun-shin^b, Graham D Cole^b, Luc Bidaut^a, Darrel P Francis^b, Massoud Zolgharni^{b,c}

^aUniversity of Lincoln, School of Computer Science, UK

^bImperial College London, National Heart and Lung Institute, UK

^cUniversity of West London, School of Computing and Engineering, UK

Abstract.

Purpose: Echocardiography is the most commonly used modality for assessing the heart in clinical practice. In an echocardiographic exam, an ultrasound probe samples the heart from different orientations and positions, thereby creating different viewpoints for assessing the cardiac function. The determination of the probe viewpoint forms an essential step in automatic echocardiographic image analysis.

Approach: In this study, convolutional neural networks are used for the automated identification of 14 different anatomical echocardiographic views (larger than any previous study) in a dataset of 8,732 videos acquired from 374 patients. Differentiable architecture search approach was utilised to design small neural network architectures for rapid inference while maintaining high accuracy. The impact of the image quality and resolution, size of the training dataset, and number of echocardiographic view classes on the efficacy of the models were also investigated.

Results: In contrast to the deeper classification architectures, the proposed models had significantly lower number of trainable parameters (up to 99.9% reduction), achieved comparable classification performance (accuracy 88.4-96.0%, precision 87.8-95.2%, recall 87.1-95.1%) and real-time performance with inference time per image of 3.6-12.6ms.

Conclusion: Compared with the standard classification neural network architectures, the proposed models are faster and achieve comparable classification performance. They also require less training data. Such models can be used for real-time detection of the standard views.

Keywords: Deep Learning, Echocardiography, Neural Architecture Search, View Classification, AutoML.

*Neda Azarmehr, n.azarmehr@gmail.com

1 Introduction

Echocardiography or cardiac ultrasound imaging is the modality of choice for the diagnosis of cardiac pathology. Echocardiographic (echo) measurements provide quantitative diagnostic markers of cardiac function. Portability, speed, and affordability are the advantages of echo.

Echo examinations are typically focused upon protocols containing diverse probe positions and orientations providing several views of the heart anatomy. Standard echo views require imaging the heart from multiple windows. Each window is specified by the transducer position and includes

8 parasternal, apical, subcostal and suprasternal. The orientation of the echo imaging plane produces
9 views such as long axis, short axis, four-chamber, and five-chamber.¹

10 Interpretation of echo images begins with view detection. This is a time-consuming and man-
11 ual process that requires specialised training and is prone to inter- and intra-observer variability.
12 Echo images are very similar and can be particularly challenging for an operator to successfully
13 categorise.

14 Therefore, accurate automatic classification of heart views has several potential clinical appli-
15 cations such as improving workflow, guiding inexperienced users, reducing inter-user discrepancy,
16 and improving accuracy for high throughput of echo data and subsequent diagnosis.

17 In most current clinical practice, images from different modalities are managed and stored in
18 Picture Archiving and Communication Systems (PACS). Recently, add-on echo software packages,
19 such as EchoPAC (GE Healthcare) and QLAB (Philips), attempt to automate the analysis and
20 diagnosis process. However, they still necessitate human involvement in detecting relevant views.
21 As previously stated, echocardiography image frames are not easily discernible by the operator,
22 plus there is often background noise. Therefore, automatic view classification could be widely
23 beneficial for pre-labelling large datasets of unclassified images.^{2,3}

24 Application of machine learning algorithms in computer vision has improved the accuracy and
25 time-efficiency of automated image analysis, particularly automated interpretation of medical im-
26 ages.⁴⁻⁷ However, traditional machine learning methods are constructed using complex processes
27 and tend to have a restricted scope and effectiveness.^{8,9} Recent advances in the design and appli-
28 cation of deep neural networks have resulted in increased possibilities when automating medical
29 image-based diagnosis.^{10,11}

30 *1.1 Approaches to neural network design*

31 Convolutional neural networks (CNNs) are extremely effective at learning patterns and features
32 from digital images and have demonstrated success in many image classification tasks.^{12,13} How-
33 ever, this success has been accompanied by a growing demand for architecture engineering of
34 increasingly more complex deep neural networks through a time-consuming and arduous man-
35 ual process. Moreover, the developed architectures are usually dependent on the particular image
36 dataset used in the design process, and adapting the architectures to new datasets remains a very
37 difficult task that relies on extensive trial and error process and expert knowledge.

38 Recently, increased attention has been paid to emerging algorithmic solutions, such as Neural
39 Architecture Search (NAS), to automate the manual process of architecture design, and these have
40 accomplished highly competitive performance in image classification tasks.¹⁴⁻¹⁷ NAS can actually
41 be considered as a subfield of automated machine learning (AutoML).¹⁸

42 Pivotal to the NAS architecture is the creation of a large collection of potential network ar-
43 chitectures. These options are subsequently explored to determine an ideal output with a specific
44 combination of training data and constraints, such as network size. Initial NAS approaches, such as
45 reinforcement learning^{19,20} and evolution,²¹ search for complete network topology, thus involving
46 extremely large search spaces comprised of arbitrary connections and operations between neural
47 network nodes. Such complexity results in using massive amounts of energy and requiring thou-
48 sands of GPU hours or million-dollar cloud compute bills²² to design neural network architectures.

49 Successful NAS approaches, such as Efficient Neural Architecture Search (ENAS) from Google
50 Brain¹⁵ and more recently Differentiable Architecture Search (DARTS),¹⁶ have been shown to re-
51 duce the search costs by orders of magnitude, requiring $\sim 100x$ fewer GPU hours. These methods

52 leverage an important observation that popular CNN architectures often contain repeating blocks
53 or are stacked sequentially. Their effectiveness is thus owing to the key idea of focusing on find-
54 ing a small optimal computational cell (as the building block of the final architecture), rather than
55 searching for a complete network. The size of the search space is therefore significantly reduced
56 since the computational cells contain considerably fewer layers than the whole network architec-
57 ture, which would make such approaches potentially viable for solving real-world challenges.

58 The DARTS method has been shown to outperform ENAS in terms of the GPU hours required
59 for the search process.¹⁶ While most NAS studies report experimental results using standard image
60 datasets such as CIFAR and ImageNet, the effectiveness of DARTS on scientific datasets, including
61 medical images, has also been demonstrated. In this study, the DARTS method for designing
62 customised architectures has been adopted.

63 *1.2 Related work on echocardiography view classification*

64 Most previous studies on automatic classification of echocardiographic views have used hand-
65 crafted features and traditional machine learning techniques, achieving varying degrees of success
66 in classifying a limited number of common echocardiographic views.²²⁻³⁰ Following the recent
67 success of deep convolutional neural networks in computer vision, and particularly for image clas-
68 sification tasks, there has been a handful of reports on the application of deep learning for cardiac
69 ultrasound view detection. Herein, we have focused on such studies.

70 Gao et al.³⁰ proposed a fused CNN architecture by integrating a deep learning network along
71 the spatial direction, and a hand-engineered feature network along the temporal dimension. The
72 final classification result for the two-strand-network was obtained through a linear combination of
73 the classification scores obtained from each network. They used a dataset of 432 image sequences

74 acquired from 93 patients. For each strand of CNN network implemented using Matlab, it took
75 2 days to process all images. Their model achieved an average accuracy rate of 92.1% when
76 classifying 8 different echocardiographic views.

77 In another study,³¹ view identification formed part of an automated pipeline designed for the
78 interpretation of echocardiograms. The standard VGG architecture was employed as the CNN
79 model, and 6 different echocardiographic views were included in the study. The class label for
80 each video was assigned by taking the majority decision of predicted view labels on the 10 frames
81 extracted from the video. The overall classification accuracy, calculated from the reported confu-
82 sion matrix, was 97.7%, and no results for single image classification was reported. In a follow-up
83 study,³ they included 23 views (9 of which were 3 apical planes, each one divided into 'no oc-
84 clusions', 'occluded LA', and 'occluded LV' categories) from 277 echocardiograms. The reported
85 overall accuracy of the VGG model dropped to 84% at an individual image level, with the greatest
86 challenge being distinctions among the various apical views. By averaging across multiple images
87 from each video, higher accuracies could be achieved.

88 Madani et al.³² proposed a CNN model to classify 12 standard B-mode echocardiographic
89 views (15 views, including Doppler modalities) using a dataset of 267 transthoracic studies (90%
90 used for training-validation, and 10% for testing). An inference latency of 21ms per image was
91 achieved for images with a size of 60×80 pixels. They also reported an average overall accuracy
92 of 91.7% for classifying single frames, compared to an average of 79.4% for expert echocardiog-
93 raphers classifying a subset of the same test images. However, this may not be a fair comparison as
94 the expert humans were given the same downsampled images that were fed into the CNN model,
95 but the human experts are not trained and have no experience of working with such low-resolution
96 images. Later on, they reported an improved classification accuracy of 93.64% by first applying

97 a segmentation stage, where the field of view was extracted from the images using U-net model³³
98 and the isolated image segment was then fed into the classifier.³⁴

99 In a more recent study,⁶ a CNN model was proposed with the aim to balance accuracy and
100 effectiveness. The design was inspired by the Inception³⁵ and DenseNet³⁶ architectures. The per-
101 formance of the model was examined using a dataset of 2559 image sequences from 265 patients,
102 and an overall accuracy of 98.3% was observed for classifying 7 echocardiographic views. The
103 reported inference time was 4.4 ms and 15.9 ms when running the model on the GPU and CPU,
104 respectively, for images with a size of 128×128 pixels.

105 Vaseli et al.³⁷ reported on designing a lightweight model with the knowledge of three state-of-
106 the-art networks (VGG16, DenseNet, and ResNet) for classifying 12 echocardiographic views. A
107 maximum accuracy of 88.1% was observed using their lightweight models, with a minimum infer-
108 ence time of $52 \mu\text{s}$ for images with a size of 80×80 pixels. However, the reported accuracies are
109 provided for classifying cine loops, and are computed as the average of the predictions for all con-
110 stituent frames in each cine loop. It is unclear how many frames constituted a cine loop. For a cine
111 loop containing 120 frames (time-window of 2s acquired at 60 frames/s), therefore, an inference
112 time of $\geq 6.2\text{ms}$ would be required to achieve the reported accuracy. A more rigorous examina-
113 tion of their models also seems necessary and, as apparent from the provided confusion matrices,
114 a great majority of the reported misclassifications, seen as a failure of the models, occurred for
115 parasternal short-axis views.

116 *1.3 Main contributions*

117 Given our two competing objectives of minimising the neural network size and maximising its
118 prediction accuracy, this study aims to adopt the recent NAS solution of DARTS for designing

119 efficient neural networks. To the best of our knowledge, no other study has applied DARTS to the
120 complex problem of echocardiographic views classification.

121 In our study, we also aimed at including subclasses of a given echocardiographic view. In
122 general, the more numerous the view classes, the more difficult the task of distinguishing the
123 views for the CNN model. This is because if a group of images is considered as a single view in
124 one study and as multiple views in another, those multiple views are likely to be relatively similar
125 in appearance. Perhaps this is one of the primary reasons for the wide range of accuracies (84-97%)
126 reported in the literature.

127 We have previously reported on preparation and annotation of a large patient dataset, covering
128 a range of pathologies and including 14 different echocardiographic views, which we used for
129 evaluating the performance of existing standard CNN architectures.³⁸ In this study, we will use
130 this dataset to design customised network architectures for the task of echo view classification.

131 The input image resolution could potentially impact the classification performance. In case
132 of aggressively downsampled images, the relevant features may in fact be lost, thus lowering the
133 classification accuracy. On the other hand, unnecessarily large images would result in more com-
134 putations. Nevertheless, all previous reports considered one particular (but dissimilar in different
135 studies) image resolution, the selection of which was always unexplained. Herein, we have thus
136 looked at the impact of different input image resolutions.

137 The accuracy of deep learning classifiers is largely dependent on the size of high-quality initial
138 training datasets. Collecting an adequate training dataset is often the primary obstacle of many
139 computer vision classification tasks. This could be particularly challenging in medical imaging
140 where the size of training datasets are scarce, e.g. because the images can only be annotated by
141 skilled experts. Hence, it would be advantageous to require less training data. Therefore, we

142 examined the influence of the size of training data on the model's performance for each of the
143 investigated networks in this study.

144 No matter how ingenious the deep learning model, image quality places a ceiling on the reli-
145 ability of any automated image analysis. Echocardiograms inherently suffer from relatively poor
146 image quality. Therefore, we also looked at the impact of image quality on the classification per-
147 formance.

148 In light of the above, the main contributions of this study can be summarised as follows:

- 149 • Inclusion of 14 different anatomical echocardiographic views (outlined in Figure 1); larger
150 than any previous study. We also examined the cases when only 7 or 5 different views were
151 included to investigate the impact of the number of views on the detection accuracy.
- 152 • Analysis of three well-known network topologies and of a proposed neural network, ob-
153 tained from applying NAS techniques to design network topologies with far fewer trainable
154 parameters and comparable/better accuracy for echo view classification.
- 155 • Analysis of computational and accuracy performance of the developed models using our
156 large-scale test dataset.
- 157 • Analysis of the impact of the input image resolution; 4 different image sizes were investi-
158 gated.
- 159 • Analysis of the influence of the size of training data on the model's performance for all
160 investigated networks.
- 161 • Analysis of the correlation between the image quality and accuracy of the model for view
162 detection.



Fig 1 The 14 cardiac views in transthoracic echocardiography: apical two-chamber (A2CH), apical three-chamber (A3CH), apical four-chamber left ventricle focused (A4CH-LV), apical four-chamber right ventricle focused (A4CH-RV), apical five-chamber (A5CH), parasternal long-axis (PLAX-Full), parasternal long-axis tricuspid valve focused (PLAX-TV), parasternal long-axis valves focused (PLAX-Valves), parasternal short-axis aortic valve focused (PSAX-AV), parasternal short-axis left ventricle focused (PSAX-LV), subcostal (Subcostal), subcostal view of the inferior vena cava (Subcostal-IVC), suprasternal (Suprasternal), and apical left atrium mitral valve focused (LA/MV).

163 2 Dataset

164 In this section, a brief account of the patient dataset used in this study is provided. A detailed
 165 description, including patient characteristics, can be found in Howard et al.³⁸

166 A random sample of 374 echocardiographic examinations of different patients and performed
 167 between 2010 and 2020 was extracted from Imperial College Healthcare NHS Trust’s echocardiogram
 168 database. The acquisition of the images was performed by experienced echocardiographers
 169 and according to standard protocols, using ultrasound equipment from GE and Philips manufac-
 170 turers.

171 Ethical approval was obtained from the Health Regulatory Agency (Integrated Research Ap-

172 plication System identifier 243023). Only studies with full patient demographic data and without
 173 intravenous contrast administration were included. Automated anonymization was performed to
 174 remove all patient-identifiable information.

175 The videos were annotated manually by an expert cardiologist (JPH), categorising each video
 176 into one of 14 classes which are outlined in Figure 1. Videos thought to show no identifiable
 177 echocardiographic features, or which depicted more than one view, were excluded. Altogether,
 178 this resulted in 9,098 echocardiographic videos. Of these, 8,732 (96.0%) videos could be classified
 179 as one of the 14 views by the human expert. The remaining 366 videos were not classifiable as a
 180 single view, either because the view changed during the video loop, or because the images were
 181 completely unrecognisable. The cardiologist’s annotations of the videos were used as the ground
 182 truth for all constituent frames of that video.

183 **DICOM-formatted videos of varying image sizes (480×640, 600×800, and 768×1024 pixels)**
 184 were then split into constituent frames, and three frames were randomly selected from each video

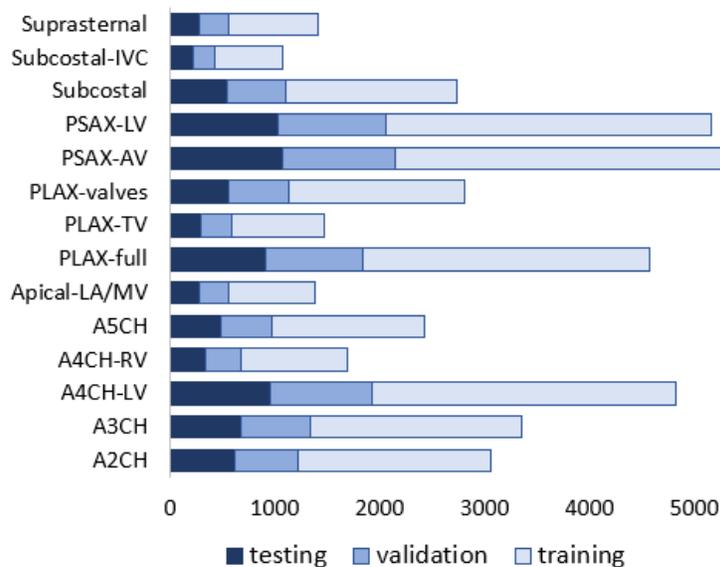


Fig 2 Distribution of data in the training, validation and test dataset; values show the number of frames in a given class.

185 to represent arbitrary stages of the heart cycle, resulting in 41,321 images. The dataset was then
186 randomly split into training (24791 images), validation (8265 images), and testing (8265 images)
187 sub-datasets in a 60:20:20 ratio. Each sub-datasets contained frames from separate echo studies to
188 maintain sample independence.

189 The relative distribution of echo view classes labelled by the expert cardiologist is displayed in
190 Figure 2 and indicates an imbalanced dataset, with a ratio of 3% (Subcostal-IVC view as the least
191 represented class) to 13% (PSAX-AV view as the dominant view).

192 **3 Method**

193 Details of the well-known classification network architectures investigated in this study (i.e., VGG16,
194 ResNet18, and DenseNet201) can be found in relevant resources.^{36,39,40} Here, a detailed descrip-
195 tion of the designed CNN models will be provided.

196 *3.1 DARTS method*

197 Proposed by Liu et al. in 2019,¹⁶ DARTS formulates the architecture search task in a differentiable
198 manner. Unlike conventional approaches of applying evolution^{21,41} or reinforcement learning^{14,42}
199 over a discrete and non-differentiable search space, DARTS is based on the continuous relaxation
200 of the architecture representation, allowing an efficient search of the architecture using gradient
201 descent.

202 DARTS method consists of two stages: architecture search and architecture evaluation. Given
203 the input images, it first embarks on an architecture search to explore for a computation cell (a
204 small unit of convolutional layers) as the building block of the neural network architecture. After
205 the architecture search phase is complete and the optimal cell is obtained based on its validation

206 performance, the final architecture could be formed from one cell or a sequential stack of cells.
 207 The weights of the optimal cell learnt during the search stage are then discarded, and are initialised
 208 randomly for retraining the generated neural network model from scratch.

209 A cell, depicted in Figure 3, is an ordered sequence of several nodes in which one or multi-
 210 ple edges meet. Each node $C^{(i)}$ represents a feature map in convolutional networks. Each edge
 211 (i,j) is associated with some operation $O^{(i,j)}$, transforming the node $C^{(i)}$ to $C^{(j)}$. This could be a
 212 combination of several operations, such as convolution, max-pooling, and ReLU.

213 Each intermediate node $C^{(j)}$ is computed based on all of its predecessors as:

$$C^{(j)} = \sum_{i < j} O^{(i,j)} (C^{(i)}) \quad (1)$$

214 Instead of applying a single operation (e.g., 5×5 convolution), and evaluating all possible oper-
 215 ations independently (each trained from scratch), DARTS places all candidate operations on each
 216 edge (e.g., 5×5 convolution, 3×3 convolution, and max-pooling represented in Figure 3 by red,
 217 blue, and green lines, respectively). This allows sharing and training their weights in a single pro-
 218 cess. The task of learning the optimal cell is effectively finding the optimal placement of operations
 219 at the edges.

220 The actual operation at each edge is then a linear combination of all candidate operations $O(i,j)$,
 221 weighted by the softmax output of the architecture parameters $\alpha^{(i,j)}$:

$$\bar{O}^{(i,j)}(C) = \sum_{o \in \partial} \frac{\exp(\alpha_o^{(i,j)})}{\sum_{o' \in \partial} \exp(\alpha_{o'}^{(i,j)})} O(C) \quad (2)$$

222 Optimization of the continuous architecture parameters α is carried out using gradient descent

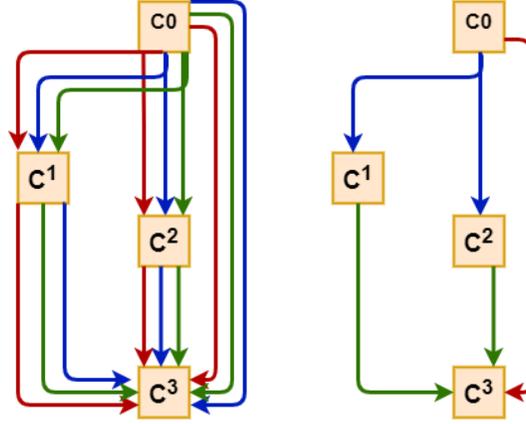


Fig 3 Schematic of a DARTS cell. Left: a computational cell with four nodes C^0 - C^3 . Edges connecting the nodes represent some candidate operations (e.g., 5×5 convolution, 3×3 convolution, and max-pooling represented in Figure 3 by red, blue, and green lines, respectively). Right: the best-performing cell learnt from retaining the optimal operations. Figure inspired by Elsken et al.⁴³

223 on the validation loss. The mixed operation $\bar{O}^{(i,j)}$ is then replaced by the operation $O^{(i,j)}$ correspond-
 224 ing to the highest weight:

$$O^{(i,j)} = \underset{o \in \partial}{\operatorname{argmax}} \alpha_0^{(i,j)} \quad (3)$$

225 An example final cell architecture is displayed in the right panel, in Figure 3. The task of archi-
 226 tecture search is learning a set of continuous variables in vector $\alpha^{(i,j)}$.

227 The training loss \mathcal{L}_{train} and validation loss \mathcal{L}_{val} are determined by the architecture parameters
 228 α and the weights ω in the network. The learning of α is performed in conjunction with learning
 229 of ω within all the candidate operations (e.g., weights of the convolution filters).

230 DARTS seeks to find the architecture α^* that minimises $\mathcal{L}_{val}(\omega^*, \alpha^*)$, where the weights ω^*
 231 associated with the architecture minimise the training loss $\omega^* = \operatorname{argmin}_{\omega} \mathcal{L}_{train}(\omega, \alpha^*)$. This indi-

232 cates a bi-level optimization problem as:

$$\min_{\alpha} \mathcal{L}_{val}(\omega^*(\alpha), \alpha) \quad (4)$$

233

$$\text{such.that } \omega^*(\alpha) = \operatorname{argmin}_{\omega} \mathcal{L}_{train}(\omega, \alpha) \quad (5)$$

234 It is computationally expensive to solve the optimization problem precisely; i.e., computing the

235 true loss by training ω for each architecture. Utilising a one-step approximation, the training of α

236 and ω is performed by alternating the gradient steps in the weights and the architecture parameters.

237 The weights are optimized by descending in the direction $\nabla_{\omega} \mathcal{L}_{train}(\omega, \alpha)$, while α is optimized

238 by descending in the direction $\nabla_{\alpha} \mathcal{L}_{val}(\omega - \xi \nabla_{\omega} \mathcal{L}_{train}(\omega, \alpha), \alpha)$, where ξ is equal to the learning

239 rate for the weights optimiser.

240 Two types of cells are defined and optimized in DARTS:

241 • Normal Cell which maintains the output spatial dimension the same as input

242 • Reduction Cell which reduces the output spatial dimension while increasing the number of

243 filters/channels

244 The final architecture is then formed by stacking these cells.

245 3.2 DARTS parameters for architecture search

246 For the stage of architecture search, 80% of the dataset was held out for equally-sized training and

247 validation subsets, and 20% for testing. Images were normalised and downsampled to 4 different

248 sizes of 32×32 , 64×64 , 96×96 , and 128×128 pixels, with corresponding batch sizes of 64, 14, 8,
249 and 4, respectively.

250 The following candidate operations were included in the architecture search stage: 3×3 and
251 5×5 separable convolutions, 3×3 and 5×5 dilated separable convolutions, 3×3 max-pooling, 3×3
252 average-pooling, skip-connection, and zero. For the convolutional operations, a ReLU-Conv-BN
253 order was used. If applicable, the operations were of stride one. The convolved feature maps were
254 padded to preserve their spatial size.

255 A network of 8 cells was then used to conduct the search for a maximum of 30 epochs. The
256 initial number of channels was 16 to make sure the network could fit into a single GPU. Stochastic
257 Gradient Decent (SGD) with a momentum of 0.9, initial learning rate of 0.1, and weight decay of
258 3×10^{-4} was used to optimise the weights. To obtain enough learning signal, DARTS utilises zero
259 initialization for architecture variables indicating the same amount of attention over all possible
260 operations as it is taking the softmax after each operation.

261 Adam optimiser⁴⁴ with an initial learning rate of 0.1, momentum of (0.5, 0.999), and weight
262 decay of 10^{-3} were used as the optimiser for α .

263 *3.3 Models training parameters*

264 Training occurred subsequently, using annotations provided by the expert cardiologist. It was
265 carried out independently for each of the 4 different image sizes of 32×32 , 64×64 , 96×96 , and
266 128×128 pixels. Identical training, validation, and testing datasets were used in all network mod-
267 els. The validation dataset was used for early stopping to avoid redundant training and overfitting.
268 Each model was trained until the validation loss plateaued. The test dataset was used for the per-
269 formance assessment of the final trained models. The DARTS models were kept blind to the test

270 dataset during the stage of architecture search.

271 Adam optimiser with a learning rate of 10^{-4} and a maximum number of 800 epochs was used
272 for training the models. The cross-entropy loss was used as the networks objective function. For
273 training the DARTS model, a learning rate of 0.1 deemed to be a better compromise between speed
274 of learning and precision of result and was therefore used. A batch size of 64 or the maximum
275 which could be fitted on the GPU (if <64) was employed.

276 It is evident from Figure 2 that the dataset is fairly imbalanced with unequal distribution of
277 different echo views. To prevent potential biases towards more dominant classes, we used online
278 batch selection where the equal number of samples from each view were randomly drawn (by
279 over-sampling of underrepresented classes). This led to training on a balanced dataset representing
280 all classes in every epoch. An epoch was still defined as the number of iterations required for the
281 network to meet all images in the training dataset.

282 *3.4 Evaluation metrics*

283 Several metrics were employed to evaluate the performance of the investigated models in this study.
284 Overall accuracy was calculated as the number of correctly classified images as a fraction of the
285 total number of images. Macro average precision and recall (average overall views of per-view
286 measures) were also computed. F1 score was calculated as the harmonic mean of the precision
287 and recall. *Since this study is a multi-class problem, F1 score was the weighted average, where the
288 weight of each class was the number of samples from that class.*

289 PyTorch⁴⁵ was used to implement the models. For the computationally intensive stage of archi-
290 tecture search, a GPU server equipped with 4 NVIDIA TITAN RTX GPUs with 64 GB of memory
291 was rented. For the subsequent training of the searched networks and also the standard models, the

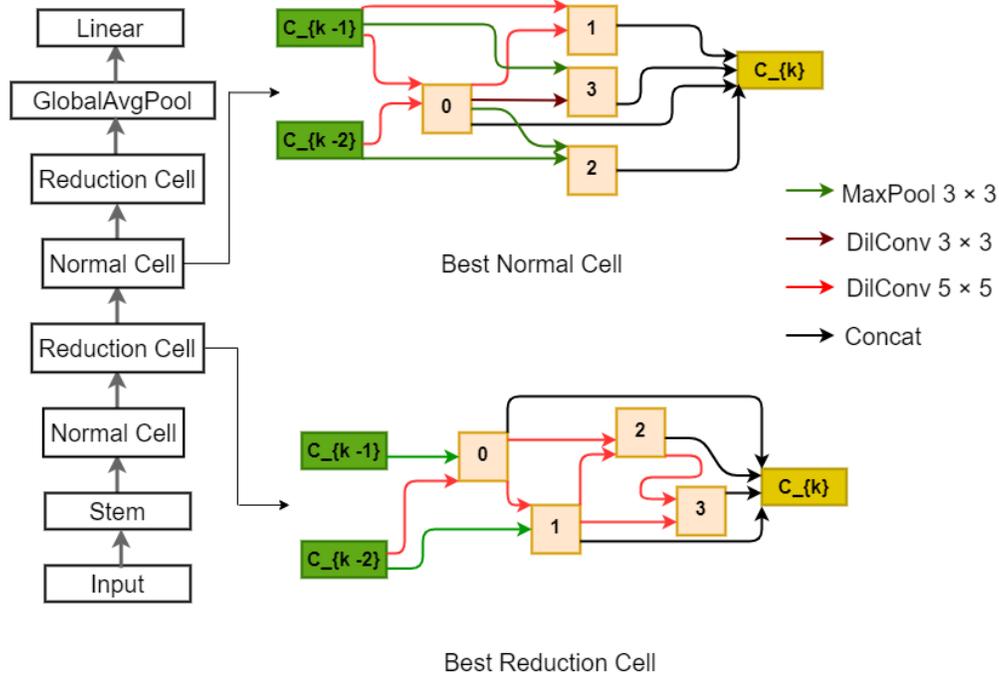


Fig 4 Optimal normal and reduction cells for the input image size of 128×128 pixels, as suggested by the DARTS method, where 3×3 and 5×5 dilated separable convolutions, 3×3 max-pooling, and skip-connection operations have been retained from the candidate operations initially included. Each cell has 2 inputs which are the cell outputs in the previous two layers. The output of the cell is defined as the depth-wise concatenation of all nodes in the cell. A schematic view of the "2-cell-DARTS", formed from a sequential stack of 2 cells, is also displayed on the left. Stem layer incorporates a convolution layer and a batch normalisation layer.

292 utilised GPU was an Nvidia QUADRO M5000 with 8 GB of memory, representing a more widely
 293 accessible hardware for real-time applications. Inference time (latency time for classifying each
 294 image) was also estimated with the trained models running on the GPU. To this end, a total of 100
 295 images were processed in a loop, and the average time was recorded. All training/prediction com-
 296 putations were carried using identical hardware and software resources, allowing for a fair com-
 297 parison of computational time-efficiency between all network models investigated in this study.

298 The number of trainable parameters in the model, as well as the training time per epoch was
 299 also recorded for all CNN networks.

Table 1 Experimental results on the test dataset for input sizes of (32×32) , (64×64) , (96×96) and (128×128) and different network topologies. Accuracy is ratio of correctly classified images to the total number of images; precision and recall are the macro average measures (average overall views of per-view measures); F1 score is the harmonic mean of precision and recall. The values in bold indicate the best performance for each measure.* For these experiments, a maximum batch size of <64 could be fitted on the GPU.

Network	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Parameters (thousands)	Inference Time (ms)	Time/epoch (s)
(32×32)							
1-cell-DARTS	88.4	87.8	87.1	87.4	58	3.6	41
2-cell-DARTS	93.0	92.5	92.3	92.3	411	7.0	46
ResNet18	90.6	89.9	89.7	89.8	11,177	11.8	184
Vgg16	90.7	89.9	89.5	89.6	134,316	8.3	210
DenseNet201	88.3	87.9	87.0	87.4	20,013	119	1303
(64×64)							
1-cell-DARTS	90.0	89.4	88.7	89.0	92	6.5	81
2-cell-DARTS	95.0	94.7	94.2	94.4	567	12.6	121
ResNet18	92.1	91.5	91.7	91.5		12.0	185
Vgg16	92.4	91.5	92.2	91.8		8.5	240
DenseNet201	93.1	92.5	92.8	92.6		127.3	1322
(96×96)							
1-cell-DARTS	93.2	92.8	92.3	92.5	101	7.2	141
2-cell-DARTS	95.4	95.1	94.9	94.9	669	14.2	264
ResNet18	93.1	92.4	92.2	92.3		12.1	186
Vgg16	93.6	92.9	93.0	92.9		8.6	276
DenseNet201	93.8	93.0	93.3	93.1		129.0	1336
(128×128)							
1-cell-DARTS	92.5	92.3	91.4	91.8	89	5.9	180
2-cell-DARTS	96.0	95.2	95.1	95.1	545	11.8	380*
ResNet18	92.9	92.6	92.2	92.4		12.2	196
Vgg16	93.2	92.1	92.7	92.3		9.0	429*
DenseNet201	93.8	93.1	93.2	93.1		129.4	1605*

300 **4 Results and Discussion**

301 *4.1 Architecture search*

302 The search took $\sim 6, 23, 42,$ and 92 hours for image sizes of $32 \times 32, 64 \times 64, 96 \times 96,$ and 128×128
303 pixels, respectively, on the computing infrastructure described earlier (section 3.4). Figure 4
304 displays the best convolutional normal and reduction cells obtained for the input image size of
305 128×128 pixels. The retained operations were 3×3 and 5×5 dilated convolutions, 3×3 max-
306 pooling, and skip-connection. Each cell is assumed to have 2 inputs which are the outputs from the
307 previous and penultimate cells. The output of the cell is defined as the depth-wise concatenation
308 of all nodes in the cell.

309 Two network architectures were assembled from the optimal cell; "1-cell-DARTS" comprised
310 of one cell only, and "2-cell-DARTS" formed from a sequential stack of 2 cells. Addition of
311 more cells to the network architecture did not significantly improve the prediction accuracy, as
312 reported in the next section, but increased the number of trainable parameters in the model and
313 thus the inference time for view classification. Therefore, the models with more than 2 cells, i.e.
314 architectures with redundancy, were judged as being comparatively inefficient and thus discarded.
315 Figure 4 (left side) also displays the full architecture for the "2-cell-DARTS" model for the input
316 image size of 128×128 pixels.

317 *4.2 View classification*

318 Results for 5 different network topologies and different image sizes are provided in Table 1. De-
319 spite having significantly fewer trainable parameters, the two DARTS models showed competitive
320 results when compared with the standard classification architectures (i.e., VGG16, ResNet18, and
321 DenseNet201). The 2-cell-DARTS model, with only $\sim 0.5m$ trainable parameters, achieves the

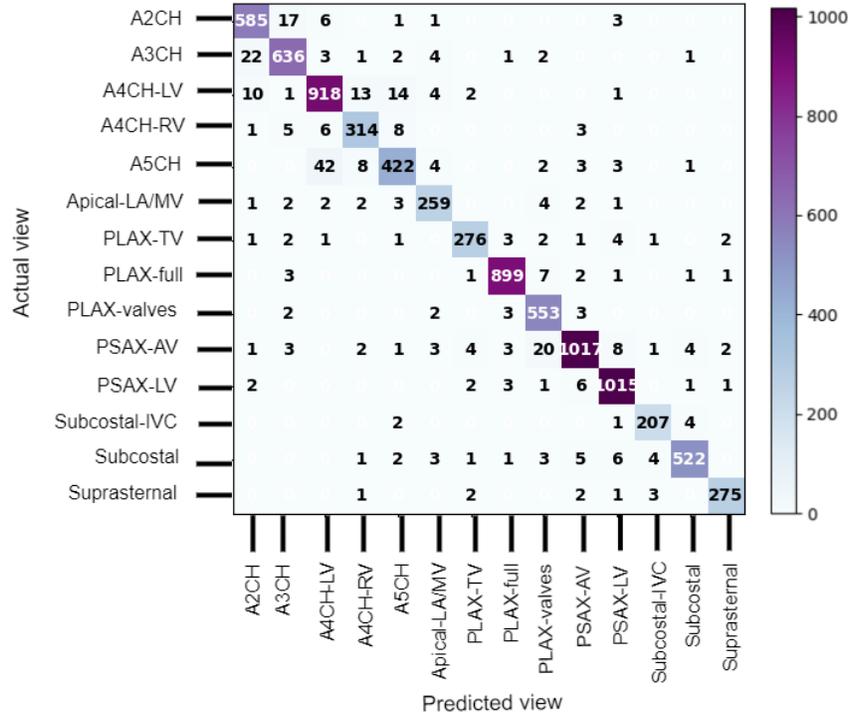


Fig 5 Confusion matrix for the 2-cell-DARTS model and input image resolution of 128×128 pixels.

322 best accuracy (93-96%), precision (92.5-95.2%), and recall (92.3-95.1%) among all networks and
 323 across all input image resolutions. Deeper standard neural networks, if employed for echo view
 324 detection, would therefore be significantly redundant, with up to 99% redundancy in trainable
 325 parameters.

326 On the other hand, while maintaining a comparable accuracy to standard network topologies,
 327 the 1-cell-DARTS model has $\leq 0.09m$ trainable parameters and the lowest inference time amongst
 328 all models and across different image resolutions (range 3.6-7.2ms). This would allow processing
 329 about 140-280 frames per second, thus making real-time echo view classification feasible.

330 Compared with manual decision making, this is a significant speedup. Although the identifi-
 331 cation of the echo view by human operators is almost instantaneous (at least for easy cases), the
 332 average time for the overall process of displaying/identifying/recording the echo view takes several
 333 seconds.

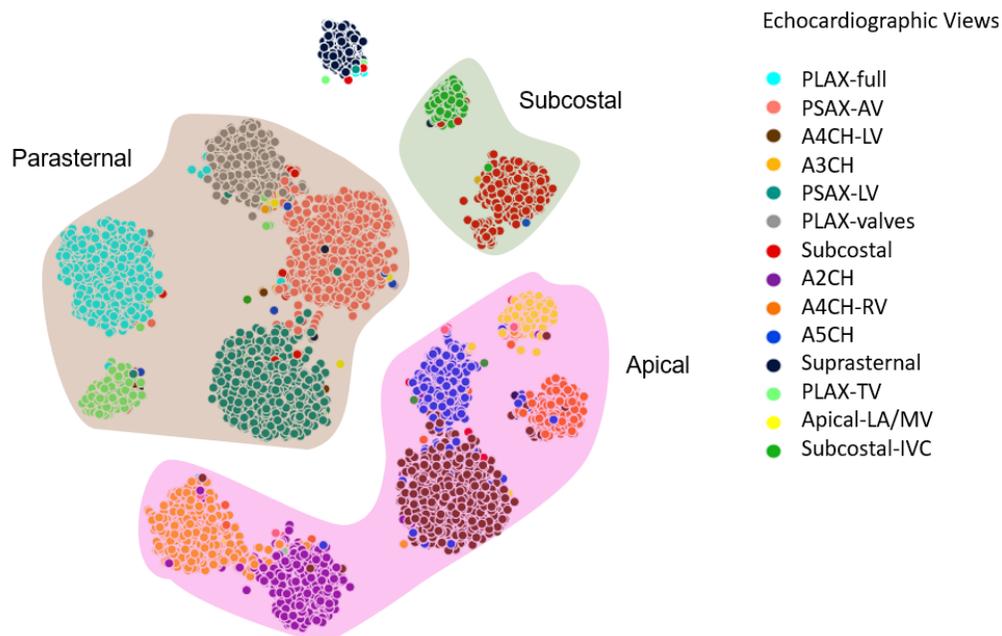


Fig 6 t-Distributed Stochastic Neighbor Embedding (t-SNE) visualisation of 14 echo views from the 2-cell-DARTS model (128×128 image size). Each point represents an echo image from the test dataset, and different colored points represent different echo view classes.

334 Having fewer trainable parameters, both DARTS models also exhibit faster convergence and
 335 shorter training time per epoch than standard deeper network architectures: 157 ± 116 s vs. 622 ± 576 s,
 336 respectively, for the training dataset we used.

337 The confusion matrix for the 2-cell-DARTS model and image resolution of 128×128 pixels
 338 is provided in Figure 5. The errors appear predominantly clustered between a certain pair of
 339 views which represent anatomically adjacent imaging planes. The A5CH view proves to be the
 340 hardest one to detect (accuracy of about 80%), as the network is confused between this view and
 341 other apical windows. This is in line with previous observations that the greatest challenge lies in
 342 distinguishing between the various apical views.³¹

343 Interestingly, the two views the model found most difficult to correctly differentiate (A4CH-
 344 LV versus A5CH, and A2CH versus A3CH) were also the two views on which the two experts

Ground truth	A3CH	A5CH	PSAX-LV
Prediction	PLAX-valves	A4CH-LV	PSAX-AV

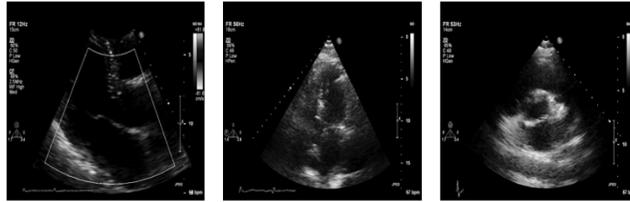


Fig 7 Three different misclassified examples predicted by the 2-cell-DARTS model for the image resolution of 128×128 pixels.

345 disagreed most often.³⁸ The A4CH view is in an anatomical continuity with the A5CH view. The
 346 difference is whether the scanning plane has been tilted to bring the aortic valve into view, which
 347 would make it A5CH. When the valve is only partially in view, or only in view during part of the
 348 cardiac cycle, the decision becomes a judgement call and there is room for disagreement. Similarly,
 349 the A3CH view differs from the A2CH view only in a rotation of the probe anticlockwise, again to
 350 bring the aortic valve into view

351 It is also interesting to note that the misclassification is not fully asymmetrical. For instance,
 352 while 42 cases of A5CH images are confused with A4CH-LV, there are only 14 occasions of
 353 A4CH-LV images mistaken for A5CH.

354 On the other hand, echo views with distinct characteristics are easier for the model to distin-
 355 guish. For instance, PLAX-full and Suprasternal seem to have higher rates of correct identification,
 356 and the network is confused only on one occasion between these two views.

357 This is also evident on the t-Distributed Stochastic Neighbor Embedding (t-SNE) plot in Figure
 358 6, which displays a planar representation of the internal high-dimensional organization of the 14
 359 trained echo view classes within the network’s final hidden layer (i.e. input data of the fully
 360 connected layer). Each point in the t-SNE plot represents an echo image from the test dataset.

361 Noticeably, not only has the network grouped similar images together (a cluster for each view,
362 displayed with different color), but it has also grouped similar views together (highlighted with a
363 unique background color). For instance, it has placed A5CH (blue) next to A4CH (dark brown),
364 and indeed there is some "interdigitation" of such cases, e.g. for those whose classification between
365 A4CH and A5CH might be debatable. Similarly, at the top right, the network has discovered that
366 the features of the Subcostal-IVC images (green) are similar to the Subcostal images (red). This
367 shows that the network can point to relationships and organizational patterns efficiently.

368 Figure 7 shows examples of misclassified cases, when the prediction of the 2-cell-DARTS
369 model disagreed with the expert annotation. The error can be explained by the inherent difficulty
370 of deciding, even for cardiologist experts, between views that are similar in appearance to human
371 eyes and are in spatial continuity (case of A4CH / A5CH mix-up), images of poor quality (case of
372 A4CH / PSAX mix-up), or views in which a same view-defining structure may be present (case of
373 PSAX-LV / PSAX/AV mix-up).

374 *4.3 Impact of image resolution, quality, and dataset size*

375 The models seem to exhibit a plateau of accuracy between the two larger image resolutions of
376 96×96 and 128×128 pixels (Fig 8). On the other hand, for the smaller image size of 32×32
377 pixels, the classification performance seems to suffer across all network models, with a 2.3-5.1%
378 reduction in accuracy relative to the resolution of 96×96 pixels.

379 Shown in Figure 9's upper panel, is the class-wise view detection accuracy for various input
380 image resolutions. Notably, not all echo views are affected similarly by using lower image reso-
381 lutions. The drop in overall performance is therefore predominantly caused by a marked decrease
382 in detection accuracy of only certain views. For instance, A4CH-RV suffers a sharp reduction of

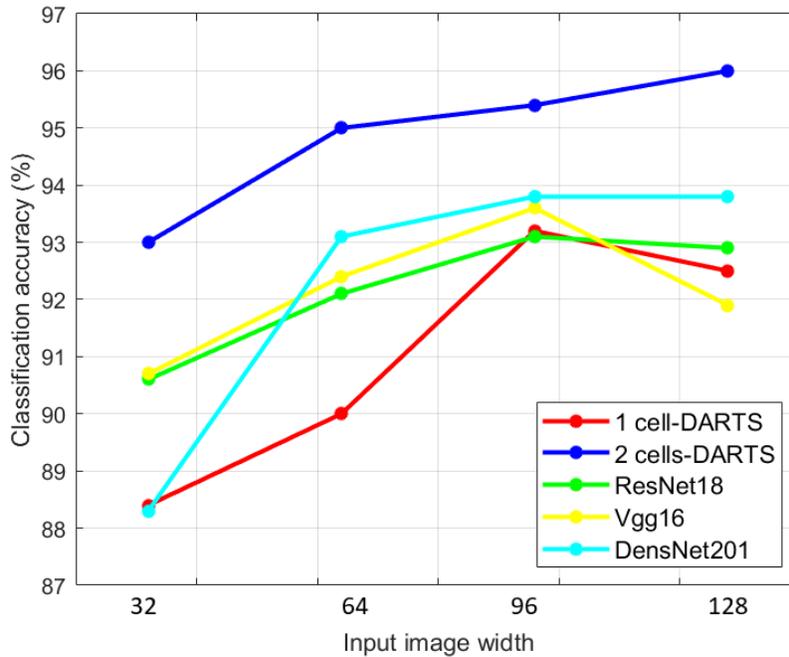


Fig 8 Comparison of accuracy for different classification models and different image resolutions; image width of 32 correspond to the image resolution of 32×32 pixels.

383 $>10\%$ in prediction accuracy when dealing with images of 32×32 pixels.

384 Figure 9's lower panel shows the relative confusion matrix, illustrating the improvement asso-
 385 ciated with using image resolution of 96×96 versus 32×32 pixels. Being already a difficult view
 386 to detect even in higher resolution images, A5CH will have 47 more cases of misclassified images
 387 when using images of 32×32 pixels. Overall, apical views seem to suffer the most from lower
 388 resolution images, being mainly misclassified as other apical views. For instance, the two classes
 389 associated with the A4CH will primarily be mistaken for one another. This is likely because, with
 390 a decreased resolution, the details of their distinct features would be less discernible by the net-
 391 work. Conversely, parasternal views seem to be less affected, and still detectable in downsampled
 392 images. This could be owing to the fact that the relevant features, on which the model relies for
 393 identifying this view, are still present and visible to the model.

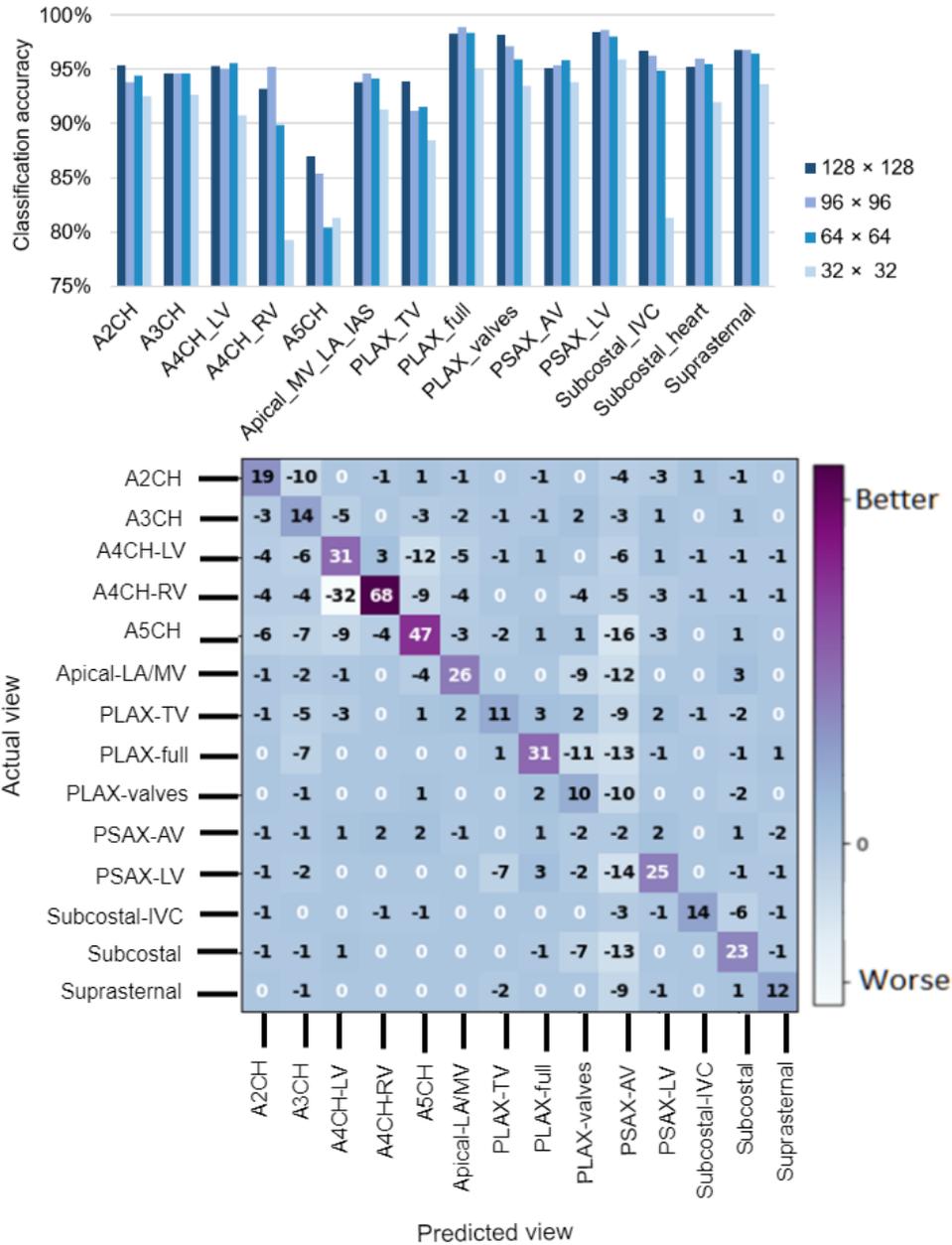


Fig 9 Accuracy of the 2-cell-DARTS model for various input image resolutions. Upper: class-wise prediction accuracy. Lower: relative confusion matrix showing improvement associated with using image resolution of 96×96 versus 32×32 pixels.

394 Overall, and for almost all echo views, the image size of 96×96 pixels appeared to be a good
 395 compromise between classification accuracy and computational costs.

396 To examine the influence of the size of the training dataset on the model’s performance, we

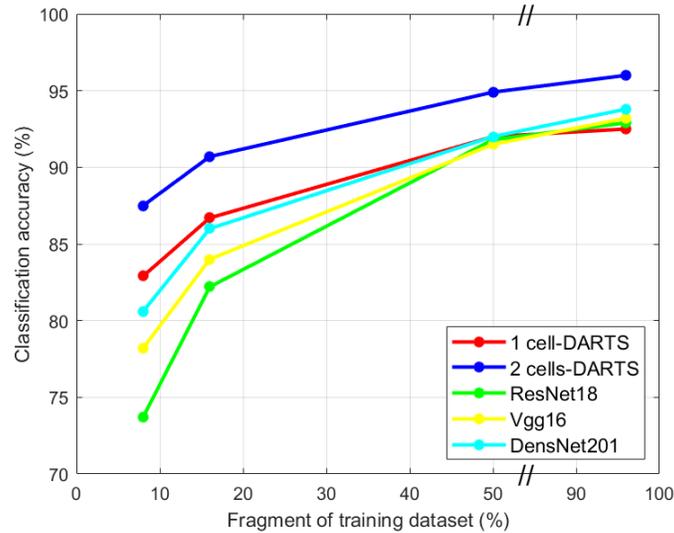


Fig 10 Comparison of accuracy of different classification models for image size of 128×128 versus different fragments of training dataset used when training the models. For each sub-dataset, all models were retrained from scratch.

397 conducted an additional experiment where we split the training data into sub-datasets with strict
 398 inclusion relationship (i.e., having the current sub-dataset a strict subset of the next sub-dataset),
 399 and ensured all the sub-datasets were consistent (i.e., having the same ratio for each echo view as in
 400 the original training dataset). We then retrained all targeted neural networks on these sub-datasets
 401 from scratch, and investigated how their accuracy varied with respect to the size of the dataset
 402 used for training the model. The size of the validation and testing datasets, however, remained
 403 unchanged.

404 Figure 10 shows a drop in the classification accuracy across all models when smaller sizes of
 405 training data are used for training the networks. However, various models are impacted differently.
 406 Suffering from redundancy, deeper neural networks require more training data to achieve similar
 407 performances. DenseNet, with the largest number of trainable parameters, appears to be the one
 408 which suffers the most, with a 20% reduction in its classification accuracy, when only 8% of the
 409 training dataset is used.

410 However, the DARTS-based models appear to be relatively less profoundly affected by the size
411 of the training dataset, where both models demonstrate no more than 8% drop in their prediction
412 accuracy when deprived of the full training dataset. When using fewer than 12,400 images (i.e.,
413 50% of the training dataset), both DARTS-based models exhibit superior performance over the
414 deeper networks.

415 Additionally, we hypothesised that the more numerous the echo view classes, the more difficult
416 the task of distinguishing the views for deep learning models, e.g. because of more chances of
417 misclassifications among classes. This is potentially the underlying reason for the inconsistent
418 accuracies (84-97%) reported in the literature when classifying between 6 to 12 different view
419 classes. To investigate this premise, we considered cases when only 5 or 7 different echo views
420 were present in the dataset. [To this end, rather than reducing the number of classes by merging
421 several views to create new classes which may not be clinically very helpful, we were selective in
422 choosing some of the existing classes.](#) For each study, we aimed at including views representing
423 anatomically adjacent or similar imaging planes such as apical windows (thus challenging for the
424 models to distinguish), as well as other echo windows. The list of echo views included in each
425 study is provided in Table 2.

426 The results show an increase in the overall prediction accuracy for the two DARTS-based
427 models, when given the task of detecting fewer echo view classes and despite having relatively
428 smaller training datasets to learn from. The 1-cell-DARTS model shows 8% improvement in its
429 performance when the number of echo views is reduced from 14 to 5. The 2-cell-DARTS model
430 reaches a maximum accuracy of 99.3%, i.e. higher than any previously reported accuracies for
431 echo view classification. This highlights the fact that for a direct comparison of the classification
432 accuracy between the models reported in literature, the number of different echo windows included

Table 2 The dependence of overall accuracy on the number of echo views; experimental results on the test dataset with 5, 7, and 14 classes for different network topologies, and image resolution of 64×64 pixels. The 7-class study included A2CH, A3CH, A4CH-LV, A5CH, PLAX-full, PSAX-LV, Subcostal-IVC, and a total of 24464 images. The 5-class study included A4CH-LV, PLAX-full, PSAX-AV, Subcostal, Suprasternal, and a total of 18896 images. Accuracy is ratio of correctly classified images to the total number of images; precision and recall are the macro average measures (average overall views of per-view measures); F1 score is the harmonic mean of precision and recall.

Network	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Parameters (thousands)	Inference Time (ms)	Time/epoch (s)
1-cell-DARTS							
14-classes	90.0	89.4	88.7	89.0	92	6.5	81
7-classes	96.4	96.1	96.1	96.1	110	7.8	58
5-classes	98.1	98.3	97.9	98.1	85	6.6	38
2-cell-DARTS							
14-classes	95.0	94.7	94.2	94.4	567	12.6	121
7-classes	97.0	96.9	96.7	96.8	709	15.6	85
5-classes	99.3	99.3	99.1	99.2	556	12.9	55

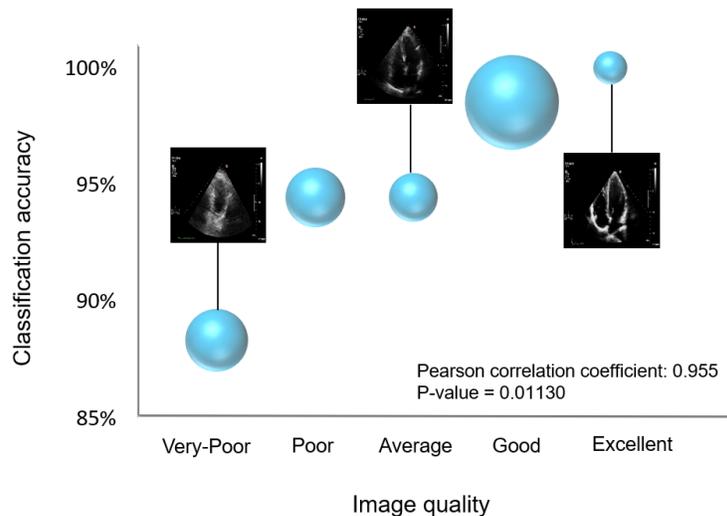


Fig 11 Correlation between the classification accuracy and the image quality (judged by the expert cardiologist) of A4CH-LV view in the test dataset. Area of the bubbles represent the relative frequency of the images in that quality score category. Results correspond to the the 2-cell-DARTS model and image resolution of 128×128 pixels. Here, p-value is the probability that the null hypothesis is true; i.e., the probability that the correlation between image quality and classification accuracy in the sample data occurred by chance.

433 in the study must be taken into account.

434 Finally, in order to study the impact of image quality on the classification performance, we

435 asked a second expert cardiologist to provide an assessment of image quality in the A4CH-LV
436 views, and assign a quality label to each image where the quality was classified into 5 grades:
437 very poor, poor, average, good, and excellent. Figure 11 displays the relationship between the
438 classification accuracy of the 2-cell-DARTS model and the image quality in the test dataset. The
439 area of the bubbles represents the relative frequency of the images in that quality score category,
440 with the "good" category as the dominant grade. This is likely because the image acquisition had
441 been performed mainly by experienced echocardiographers.

442 The correlation between the classification accuracy and the image quality is evident (p -value of
443 0.01). Images labelled as having "excellent" quality, indicated the highest classification accuracy
444 of $\sim 100\%$. It is apparent that the discrepancy between the model's prediction and the expert
445 annotation is higher in poor quality images. This could potentially be due to the fact that poorly
446 visible chambers with a low degree of endocardial border delineation could result in some views
447 being mistaken for other apical windows.

448 *4.4 Study limitations and future work*

449 This study sheds light on several possible directions for future work. Herein, we have focused on
450 the rapid and accurate classification of individual frames from an echo cine loop. Such a task will
451 be crucial for a real-time view detection system in clinical scenarios where images need to be pro-
452 cessed while they are acquired from the patient and/or where the system is to be used for operator
453 guidance. However, for offline studies and when the entire cine loop is available, classification of
454 the echo videos could also be of practical use. Some studies have attempted video classification
455 using the majority vote on some or all frames from a given video.^{6,34} However, this approach does
456 not use the temporal information available in the cine loop, such as the movement of structures

457 during the cardiac cycle. Therefore, a future study could look into using all available information
458 for view detection.

459 Our study investigated 2D echocardiography as the clinically relevant modality. Currently,
460 3D echocardiography suffers from a considerable reduction in frame rate and image quality, and
461 this has limited its adoption into routine practice over the past decade.^{46–48} When such issues are
462 resolved, automatic processing of the 3D modality could also be explored. In the meantime, 2D
463 echocardiography remains unrivalled, particularly when high frame rates are needed.

464 We investigated the impact of image quality on the classification accuracy for apical four-
465 chamber views only. A more comprehensive examination of the image quality and its influence on
466 the detection of different echo views would be informative.

467 The dataset used in this study was comprised of images acquired using ultrasound equipment
468 from GE and Philips manufacturers. Although the proposed models do not make any *a priori*
469 assumptions on data obtained from specific vendors and therefore should be vendor-neutral, echo
470 studies using more diverse ultrasound equipment should still be explored.

471 Similar to all previous studies, our dataset originated from one medical centre, i.e. Imperial
472 College Healthcare NHS Trust’s echocardiogram database. Representative multi-centre patient
473 data will be essential for ensuring that the developed models will scale up well to other sites and
474 environments.

475 Interpreting the results of the proposed models alongside other proposed architectures in the
476 literature (with a wide range of reported accuracies) was not feasible. This is due to the fact that a
477 direct comparison of the classification accuracy would require access to the same patient dataset.
478 At present, no echocardiography dataset and corresponding annotations for view detection are
479 publicly available.

480 In order to address such broadly acknowledged shortcomings in the application of deep learn-
481 ing to echocardiography, we are now developing Unity (data.unityimaging.net), a UK collabora-
482 tive of cardiologists, physiologists and computer scientists, under the aegis of the British Society
483 of Echocardiography. An image analysis interface has been developed in the form of a web-based,
484 interactive, real-time platform to capture carefully-curated expert annotations from numerous echo
485 specialists, with patient data provided by over a dozen sites across the UK, thus ensuring cover-
486 age of multiple vendors, systems and environments. All developed models designed using this
487 annotation biobank (e.g., automated cardiac phase detection,⁴⁹ left ventricular segmentation,⁵⁰ and
488 view classification in current study), will be made available under open-source agreements on
489 [intsav.github.io](https://github.com/intsav).

490 **5 Conclusion**

491 In this study, efficient CNN architectures are proposed for automated identification of the 2D
492 echocardiographic views. The DARTS method was used in designing optimized architectures
493 for rapid inference while maintaining high accuracy. A dataset of 14 different echocardiographic
494 views was used for training and testing the proposed models. Compared with the standard classi-
495 fication CNN architectures, the proposed models are faster and achieve comparable classification
496 performance. Such models can thus be used for real-time detection of the standard echo views.

497 The impact of image quality and size of the training dataset on the efficacy of the models was
498 also investigated. Deeper neural network models, with a large number of redundant trainable pa-
499 rameters, require more training data to achieve similar performances. A direct correlation between
500 the image quality of classification accuracy was observed.

501 The number of different echo views to be detected has a direct impact on the performance of

502 the deep learning models, and must be taken into account for a fair comparison of classification
503 models.

504 Aggressively downsampled images will result in losing relevant features, thus lowering the
505 prediction accuracy. On the other hand, while much larger images may be favoured for some
506 fine grained applications (e.g., segmentation), their use for echo view classification would offer
507 only slight improvements in performance (if any) at the expense of more processing and memory
508 requirements.

509 *Disclosures*

510 No conflicts of interest are declared by the authors.

511 *Acknowledgments*

512 This work was supported in part by the British Heart Foundation (Grant no. PG/19/78/34733).
513 N. Azarmehr is supported by the School of Computer Science, PhD scholarship at University of
514 Lincoln, UK. We would like to express our gratitude to Piotr Bialecki for his valuable suggestions.
515 We also thank Apostolos Vrettos for providing the expert annotations used to assess the impact of
516 image quality.

517 *References*

518 1 R. M. Lang, L. P. Badano, V. Mor-Avi, *et al.*, “Recommendations for cardiac chamber quan-
519 tification by echocardiography in adults: an update from the american society of echocardi-
520 ography and the european association of cardiovascular imaging,” *European Heart Journal-
521 Cardiovascular Imaging* **16**(3), 233–271 (2015).

- 522 2 H. Khamis, G. Zurakhov, V. Azar, *et al.*, “Automatic apical view classification of echocardiogram-
523 grams using a discriminative learning dictionary,” *Medical Image Analysis* **36**, 15–21 (2017).
- 524 3 J. Zhang, S. Gajjala, P. Agrawal, *et al.*, “Fully automated echocardiogram interpretation
525 in clinical practice: feasibility and diagnostic accuracy,” *Circulation* **138**(16), 1623–1635
526 (2018).
- 527 4 J. H. Park, S. K. Zhou, C. Simopoulos, *et al.*, “Automatic cardiac view classification of
528 echocardiogram,” in *2007 IEEE 11th International Conference on Computer Vision*, 1–8,
529 IEEE (2007).
- 530 5 K. Siegersma, T. Leiner, D. Chew, *et al.*, “Artificial intelligence in cardiovascular imaging:
531 state of the art and implications for the imaging cardiologist.,” *Netherlands Heart Journal:
532 Monthly Journal of the Netherlands Society of Cardiology and the Netherlands Heart Foun-
533 dation* **27**(9), 403–413 (2019).
- 534 6 A. Østvik, E. Smistad, S. A. Aase, *et al.*, “Real-time standard view classification in transtho-
535 racic echocardiography using convolutional neural networks,” *Ultrasound in medicine & bi-
536 ology* **45**(2), 374–384 (2019).
- 537 7 S. K. Zhou, J. Park, B. Georgescu, *et al.*, “Image-based multiclass boosting and echocardi-
538 ographic view classification,” in *2006 IEEE Computer Society Conference on Computer Vision
539 and Pattern Recognition (CVPR’06)*, **2**, 1559–1565, IEEE (2006).
- 540 8 J. Stoitsis, I. Valavanis, S. G. Mougiakakou, *et al.*, “Computer aided diagnosis based on med-
541 ical image processing and artificial intelligence methods,” *Nuclear Instruments and Meth-
542 ods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated
543 Equipment* **569**(2), 591–595 (2006).

- 544 9 K. Doi, “Computer-aided diagnosis in medical imaging: historical review, current status and
545 future potential,” *Computerized medical imaging and graphics* **31**(4-5), 198–211 (2007).
- 546 10 A. Coates, B. Huval, T. Wang, *et al.*, “Deep learning with cots hpc systems,” in *International
547 conference on machine learning*, 1337–1345 (2013).
- 548 11 M. I. Razzak, S. Naz, and A. Zaib, “Deep learning for medical image processing: Overview,
549 challenges and the future,” *Classification in BioApps* , 323–350 (2018).
- 550 12 A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolu-
551 tional neural networks,” in *Advances in neural information processing systems*, 1097–1105
552 (2012).
- 553 13 G. Litjens, T. Kooi, B. E. Bejnordi, *et al.*, “A survey on deep learning in medical image
554 analysis,” *Medical image analysis* **42**, 60–88 (2017).
- 555 14 B. Zoph and Q. V. Le, “Neural architecture search with reinforcement learning,” in *5th Inter-
556 national Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26,
557 2017, Conference Track Proceedings*, OpenReview.net (2017).
- 558 15 H. Pham, M. Guan, B. Zoph, *et al.*, “Efficient neural architecture search via parameters shar-
559 ing,” in *International Conference on Machine Learning*, 4095–4104, PMLR (2018).
- 560 16 H. Liu, K. Simonyan, and Y. Yang, “Darts: Differentiable architecture search,” in *Interna-
561 tional Conference on Learning Representations*, (2018).
- 562 17 S. Xie, H. Zheng, C. Liu, *et al.*, “SNAS: stochastic neural architecture search,” in *Interna-
563 tional Conference on Learning Representations*, (2019).
- 564 18 F. Hutter, L. Kotthoff, and J. Vanschoren, *Automated machine learning: methods, systems,
565 challenges*, Springer Nature (2019).

- 566 19 I. Bello, B. Zoph, V. Vasudevan, *et al.*, “Neural optimizer search with reinforcement learn-
567 ing,” in *International Conference on Machine Learning*, 459–468, PMLR (2017).
- 568 20 B. Zoph, V. Vasudevan, J. Shlens, *et al.*, “Learning transferable architectures for scalable
569 image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern
570 recognition*, 8697–8710 (2018).
- 571 21 E. Real, A. Aggarwal, Y. Huang, *et al.*, “Regularized evolution for image classifier architec-
572 ture search,” in *Proceedings of the aaai conference on artificial intelligence*, **33**, 4780–4789
573 (2019).
- 574 22 E. Strubell, A. Ganesh, and A. McCallum, “Energy and policy considerations for deep learn-
575 ing in nlp,” in *Proceedings of the 57th Annual Meeting of the Association for Computational
576 Linguistics*, 3645–3650 (2019).
- 577 23 S. Ebadollahi, S.-F. Chang, and H. Wu, “Automatic view recognition in echocardiogram
578 videos using parts-based representation,” in *Proceedings of the 2004 IEEE Computer So-
579 ciety Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, **2**, II–II,
580 IEEE (2004).
- 581 24 D. Agarwal, K. Shriram, and N. Subramanian, “Automatic view classification of echocardi-
582 ograms using histogram of oriented gradients,” in *2013 IEEE 10th International Symposium
583 on Biomedical Imaging*, 1368–1371, IEEE (2013).
- 584 25 H. Wu, D. M. Bowers, T. T. Huynh, *et al.*, “Echocardiogram view classification using low-
585 level features,” in *2013 IEEE 10th International Symposium on Biomedical Imaging*, 752–
586 755, IEEE (2013).
- 587 26 R. Kumar, F. Wang, D. Beymer, *et al.*, “Cardiac disease detection from echocardiogram using

- 588 edge filtered scale-invariant motion features,” in *2010 IEEE Computer Society Conference on*
589 *Computer Vision and Pattern Recognition-Workshops*, 162–169, IEEE (2010).
- 590 27 M. Otey, J. Bi, S. Krishna, *et al.*, “Automatic view recognition for cardiac ultrasound images,”
591 in *International Workshop on Computer Vision for Intravascular and Intracardiac Imaging*,
592 187–194 (2006).
- 593 28 D. Beymer, T. Syeda-Mahmood, and F. Wang, “Exploiting spatio-temporal information for
594 view recognition in cardiac echo videos,” in *2008 IEEE Computer Society Conference on*
595 *Computer Vision and Pattern Recognition Workshops*, 1–8, IEEE (2008).
- 596 29 R. Kumar, F. Wang, D. Beymer, *et al.*, “Echocardiogram view classification using edge fil-
597 tered scale-invariant motion features,” in *2009 IEEE Conference on Computer Vision and*
598 *Pattern Recognition*, 723–730, IEEE (2009).
- 599 30 X. Gao, W. Li, M. Loomes, *et al.*, “A fused deep learning architecture for viewpoint classifi-
600 cation of echocardiography,” *Information Fusion* **36**, 103–113 (2017).
- 601 31 R. C. Deo, J. Zhang, L. A. Hallock, *et al.*, “An end-to-end computer vision pipeline for
602 automated cardiac function assessment by echocardiography,” *CoRR* (2017).
- 603 32 A. Madani, R. Arnaout, M. Mofrad, *et al.*, “Fast and accurate view classification of echocar-
604 diograms using deep learning,” *NPJ digital medicine* **1**(1), 1–8 (2018).
- 605 33 O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical im-
606 age segmentation,” in *International Conference on Medical image computing and computer-*
607 *assisted intervention*, 234–241, Springer (2015).
- 608 34 A. Madani, J. R. Ong, A. Tibrewal, *et al.*, “Deep echocardiography: data-efficient supervised

- 609 and semi-supervised deep learning towards automated diagnosis of cardiac disease,” *NPJ*
610 *digital medicine* **1**(1), 1–11 (2018).
- 611 35 C. Szegedy, V. Vanhoucke, S. Ioffe, *et al.*, “Rethinking the inception architecture for computer
612 vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*,
613 2818–2826 (2016).
- 614 36 G. Huang, Z. Liu, L. Van Der Maaten, *et al.*, “Densely connected convolutional networks,” in
615 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708
616 (2017).
- 617 37 H. Vaseli, Z. Liao, A. H. Abdi, *et al.*, “Designing lightweight deep learning models for
618 echocardiography view classification,” in *Medical Imaging 2019: Image-Guided Procedures,
619 Robotic Interventions, and Modeling*, **10951**, 109510F, International Society for Optics and
620 Photonics (2019).
- 621 38 J. P. Howard, J. Tan, M. J. Shun-Shin, *et al.*, “Improving ultrasound video classification: an
622 evaluation of novel deep learning methods in echocardiography,” *Journal of medical artificial
623 intelligence* **3** (2020).
- 624 39 K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image
625 recognition,” in *3rd International Conference on Learning Representations, ICLR 2015, San
626 Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, (2015).
- 627 40 K. He, X. Zhang, S. Ren, *et al.*, “Deep residual learning for image recognition,” in *Proceed-
628 ings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
- 629 41 E. Real, S. Moore, A. Selle, *et al.*, “Large-scale evolution of image classifiers,” in *Interna-
630 tional Conference on Machine Learning*, 2902–2911, PMLR (2017).

- 631 42 M. Botvinick, S. Ritter, J. X. Wang, *et al.*, “Reinforcement learning, fast and slow,” *Trends in*
632 *cognitive sciences* **23**(5), 408–422 (2019).
- 633 43 T. Elsken, J. H. Metzen, F. Hutter, *et al.*, “Neural architecture search: A survey.,” *J. Mach.*
634 *Learn. Res.* **20**(55), 1–21 (2019).
- 635 44 D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *3rd International*
636 *Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015,*
637 *Conference Track Proceedings*, (2015).
- 638 45 A. Paszke, S. Gross, S. Chintala, *et al.*, “Automatic differentiation in pytorch,” (2017).
- 639 46 K. Cheng, M. Monaghan, A. Kenny, *et al.*, “3d echocardiography: Benefits and steps to wider
640 implementation,” *Br J Cardiol* **25**, 63–68 (2018).
- 641 47 D. Loeckx, J. Ector, F. Maes, *et al.*, “Spatiotemporal non-rigid image registration for 3d
642 ultrasound cardiac motion estimation,” in *Medical Imaging 2007: Ultrasonic Imaging and*
643 *Signal Processing*, **6513**, 65130X, International Society for Optics and Photonics (2007).
- 644 48 P. Carnahan, J. Moore, D. Bainbridge, *et al.*, “Multi-view 3d echocardiography volume com-
645 pounding for mitral valve procedure planning,” in *Medical Imaging 2020: Image-Guided*
646 *Procedures, Robotic Interventions, and Modeling*, **11315**, 1131510, International Society for
647 Optics and Photonics (2020).
- 648 49 E. S. Lane, N. Azarmehr, J. Jevsikov, *et al.*, “Multibeat echocardiographic phase detection
649 using deep neural networks,” *Computers in Biology and Medicine* , 104373 (2021).
- 650 50 N. Azarmehr, X. Ye, F. Janan, *et al.*, “Automated segmentation of left ventricle in 2d echocar-
651 diography using deep learning,” in *International Conference on Medical Imaging with Deep*
652 *Learning – Extended Abstract Track*, (London, United Kingdom) (2019).

653 **Neda Azarmehr** is a Research Associate at the University of Sheffield. In May 2017, Neda
654 was awarded a PhD scholarship at the University of Lincoln, in collaboration with Imperial College
655 London to develop automated models using deep learning, computer vision algorithm to assess the
656 left ventricle function which enables physicians to analyse cardiac echo images more precisely.
657 Her research focuses on developing models using Artificial Intelligence, Deep Learning, Computer
658 Vision to support clinicians in decision-making.

659 **Xujiong Ye** is a Professor of Medical Imaging and Computer Vision in the School of Computer
660 Science, University of Lincoln, UK. Prof. Ye has over 20 years of research and development
661 experience in medical imaging and computer vision from both academia and industry. Her main
662 research is to develop computational models using advanced image analysis, computer vision, and
663 artificial intelligence to support clinicians in decision-making.

664 **James P Howard** is currently undertaking his PhD, “Machine Learning in Cardiovascular
665 Imaging”, at Imperial College London, UK. His research interests include the applications of
666 constitutional neural networks in the processing of echocardiograms, cardiac magnetic resonance
667 imaging, and coronary physiology waveform analysis.

668 **Elisabeth Sarah Lane** completed an MSc in Software Engineering in 2019 and is currently
669 a PhD candidate at The University of West London. Her current research focus is the application
670 of Deep Learning algorithms for automatic phase detection in echocardiograms. Beth’s interests
671 include Machine Learning, Computer Vision and Artificial Intelligence for the analysis and inter-
672 pretation of clinical imaging.

673 **Robert Labs** is currently undertaking his PhD, Artificial Intelligence for automated quality

674 assessment of 2D echocardiography at University of West London, UK. His research interests
675 include the clinical application of machine learning for accurate quantification and diagnosis of
676 cardiac infarction.

677 **Matthew J Shun-shin** is a Clinical Lecturer in Cardiology at National Heart and Lung Insti-
678 tute, Imperial College London. His research interests include Artificial Intelligence, Echocardiog-
679 raphy, valvular heart disease, and improving device therapies for heart failure.

680 **Graham D Cole** is a Consultant Cardiologist primarily based at Hammersmith Hospital, part
681 of Imperial College Healthcare NHS Trust. Graham qualified from Gonville and Caius College,
682 University of Cambridge in 2005. He subsequently completed a four-year fellowship in cardiac
683 MRI at Heart Hospital Imaging Centre, the London CT fellowship and a PhD in echocardiography.
684 He has interests in the optimal use of cardiac imaging and research reliability.

685 **Luc Bidaut** has worked with and on most aspects of biomedical imaging and technology in
686 highly multidisciplinary research, clinical and translational international environments, always in
687 direct collaboration with all relevant stakeholders from scientific, technical and medical disci-
688 plines. His active involvement in the development, implementation and actual deployment of re-
689 lated technologies and applications was and remains primarily focused on maximizing the utility
690 and actionability of the information collected through imaging modalities and other sensors, at
691 various stages of the translational pipeline or clinical workflow.

692 **Darrel P Francis** is a Professor of Cardiology at the National Heart and Lung Institute, Impe-
693 rial College London. He specialises in using quantitative techniques, derived from mathematics,
694 engineering, and statistics, to problems that affect patients with heart disease.

695 **Massoud Zolgharni** is a Professor of Computer Vision at the School of Computing and En-
696 gineering, University of West London. He is also an Honorary Research Fellow at the National
697 Heart and Lung Institute, Imperial College London. His research interests include Computer Vi-
698 sion, Medical Imaging, Machine Learning, and Numerical Simulations.

699 **List of Figures**

- 700 1 The 14 cardiac views in transthoracic echocardiography: apical two-chamber (A2CH),
701 apical three-chamber (A3CH), apical four-chamber left ventricle focused (A4CH-
702 LV), apical four-chamber right ventricle focused (A4CH-RV), apical five-chamber
703 (A5CH), parasternal long-axis (PLAX-Full), parasternal long-axis tricuspid valve
704 focused (PLAX-TV), parasternal long-axis valves focused (PLAX-Valves), paraster-
705 nal short-axis aortic valve focused (PSAX-AV), parasternal short-axis left ventricle
706 focused (PSAX-LV), subcostal (Subcostal), subcostal view of the inferior vena
707 cava (Subcostal-IVC), suprasternal (Suprasternal), and apical left atrium mitral
708 valve focused (LA/MV).
- 709 2 Distribution of data in the training, validation and test dataset; values show the
710 number of frames in a given class.
- 711 3 Schematic of a DARTS cell. Left: a computational cell with four nodes C^0 - C^3 .
712 Edges connecting the nodes represent some candidate operations (e.g., 5×5 con-
713 volution, 3×3 convolution, and max-pooling represented in Figure 3 by red, blue,
714 and green lines, respectively). Right: the best-performing cell learnt from retaining
715 the optimal operations. Figure inspired by Elsken et al.⁴³

- 716 4 Optimal normal and reduction cells for the input image size of 128×128 pixels,
717 as suggested by the DARTS method, where 3×3 and 5×5 dilated separable con-
718 volutions, 3×3 max-pooling, and skip-connection operations have been retained
719 from the candidate operations initially included. Each cell has 2 inputs which are
720 the cell outputs in the previous two layers. The output of the cell is defined as the
721 depth-wise concatenation of all nodes in the cell. A schematic view of the "2-cell-
722 DARTS", formed from a sequential stack of 2 cells, is also displayed on the left.
723 Stem layer incorporates a convolution layer and a batch normalisation layer.
- 724 5 Confusion matrix for the 2-cell-DARTS model and input image resolution of 128×128
725 pixels.
- 726 6 t-Distributed Stochastic Neighbor Embedding (t-SNE) visualisation of 14 echo
727 views from the 2-cell-DARTS model (128×128 image size). Each point repre-
728 sents an echo image from the test dataset, and different colored points represent
729 different echo view classes.
- 730 7 Three different misclassified examples predicted by the 2-cell-DARTS model for
731 the image resolution of 128×128 pixels.
- 732 8 Comparison of accuracy for different classification models and different image res-
733 olutions; image width of 32 correspond to the image resolution of 32×32 pixels.
- 734 9 Accuracy of the 2-cell-DARTS model for various input image resolutions. Up-
735 per: class-wise prediction accuracy. Lower: relative confusion matrix showing
736 improvement associated with using image resolution of 96×96 versus 32×32 pix-
737 els.

- 738 10 Comparison of accuracy of different classification models for image size of 128×128
739 versus different fragments of training dataset used when training the models. For
740 each sub-dataset, all models were retrained from scratch.
- 741 11 Correlation between the classification accuracy and the image quality (judged by
742 the expert cardiologist) of A4CH-LV view in the test dataset. Area of the bubbles
743 represent the relative frequency of the images in that quality score category. Re-
744 sults correspond to the the 2-cell-DARTS model and image resolution of 128×128
745 pixels. Here, p-value is the probability that the null hypothesis is true; i.e., the
746 probability that the correlation between image quality and classification accuracy
747 in the sample data occurred by chance.

748 **List of Tables**

- 749 1 Experimental results on the test dataset for input sizes of (32×32) , (64×64) , (96×96)
750 and (128×128) and different network topologies. Accuracy is ratio of correctly
751 classified images to the total number of images; precision and recall are the macro
752 average measures (average overall views of per-view measures); F1 score is the
753 harmonic mean of precision and recall. The values in bold indicate the best perfor-
754 mance for each measure.* For these experiments, a maximum batch size of <64
755 could be fitted on the GPU.

756 2 The dependence of overall accuracy on the number of echo views; experimental
757 results on the test dataset with 5, 7, and 14 classes for different network topolo-
758 gies, and image resolution of 64×64 pixels. The 7-class study included A2CH,
759 A3CH, A4CH-LV, A5CH, PLAX-full, PSAX-LV, Subcostal-IVC, and a total of
760 24464 images. The 5-class study included A4CH-LV, PLAX-full, PSAX-AV, Sub-
761 costal, Suprasternal, and a total of 18896 images. Accuracy is ratio of correctly
762 classified images to the total number of images; precision and recall are the macro
763 average measures (average overall views of per-view measures); F1 score is the
764 harmonic mean of precision and recall.