Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Semi-supervised learning with natural language processing for right ventricle classification in echocardiography—a scalable approach

Eva Hagberg [a,b,*], David Hagerman [a,e], Richard Johansson [c], Nasser Hosseini [d], Jan Liu [e], Elin Björnsson [e], Jennifer Alvén [a,e], Ola Hjelmgren [a,b]

[a] Department of Molecular and Clinical Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
[b] Region Västra Götaland, Sahlgrenska University Hospital, Department of Clinical Physiology, Gothenburg, Sweden
[c] Department of Computer Science and Engineering, University of Gothenburg, Gothenburg, Sweden
[d] Department of Medical Physics and Biomedical Engineering, Sahlgrenska University Hospital, Gothenburg, Sweden
[e] Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden

ARTICLE INFO

Keywords:
Natural language processing
Scalable
Machine learning
Echocardiography
Right ventricle
Semi-supervised
Algorithm
Automation

ABSTRACT

We created a deep learning model, trained on text classified by natural language processing (NLP), to assess right ventricular (RV) size and function from echocardiographic images. We included 12,684 examinations with corresponding written reports for text classification. After manual annotation of 1489 reports, we trained an NLP model to classify the remaining 10,651 reports. A view classifier was developed to select the 4-chamber or RV-focused view from an echocardiographic examination ($n = 539$). The final models were two image classification models trained on the predicted labels from the combined manual annotation and NLP models and the corresponding echocardiographic view to assess RV function (training set $n = 11,008$) and size (training set $n = 9951$. The text classifier identified impaired RV function with 99% sensitivity and 98% specificity and RV enlargement with 98% sensitivity and 98% specificity. The view classification model identified the 4-chamber view with 92% accuracy and the RV-focused view with 73% accuracy. The image classification models identified impaired RV function with 93% sensitivity and 72% specificity and an enlarged RV with 80% sensitivity and 85% specificity; agreement with the written reports was substantial (both $\kappa = 0.65$). Our findings show that models for automatic image assessment can be trained to classify RV size and function by using model-annotated data from written echocardiography reports. This pipeline for auto-annotation of the echocardiographic images, using a NLP model with medical reports as input, can be used to train an image-assessment model without manual annotation of images and enables fast and inexpensive expansion of the training dataset when needed.

1. Introduction

Right ventricular (RV) dysfunction, a syndrome associated with poor clinical outcomes independently of the mechanism of disease [1–3], is most commonly initiated by increases in RV afterload and pulmonary artery pressure in patients with chronic left ventricular (LV) disease [4]. RV assessment by cardiac echocardiography, the most common cardiovascular diagnostic test besides electrocardiography, can be challenging, mainly because RV anatomy is complex and acquisition windows are limited. Other noninvasive cardiac imaging such as computed tomography and magnetic resonance imaging are increasing in popularity, but echocardiography remains important because it has unique capabilities, is free from radiation, and can be done at the patient's bedside.

According to guidelines for RV assessment by two-dimensional transthoracic echocardiography (2DE) from the American Society of Echocardiography and the European Association of Cardiovascular Imaging, an RV diameter >41 mm at the base and >35 mm at the midlevel in the RV-focused apical four-chamber view (RV-focused view) indicates RV dilatation [5]. Guidelines further suggest that RV function should be

Abbreviations: 2DE, two-dimensional transthoracic echocardiography; 3D, three-dimensional; 4C, four-chamber; BERT, bidirectional encoder representations from transformers; CNN, convolutional neural network; GPU, graphics processing unit; IQ, interquartile range; LVEF, left ventricular ejection fraction; NLP, natural language processing; RV, right ventricular; sPAP, systolic pulmonary artery pressure; TAPSE, tissue Doppler–derived tricuspid lateral annular systolic velocity.

* Corresponding author. Klinisk fysiologi, Sahlgrenska University Hospital, Blå Stråket 5, 41345, Gothenburg, Sweden.
E-mail address: eva.hagberg@wlab.gu.se (E. Hagberg).

https://doi.org/10.1016/j.compbiomed.2022.105282
Received 2 September 2021; Received in revised form 30 January 2022; Accepted 30 January 2022
Available online 15 February 2022

evaluated in multiple views, including the RV-focused view, by visual assessment and by measuring at least one of the following: tissue Doppler–derived tricuspid lateral annular systolic velocity (S′), fractional area change, RV index of myocardial performance, or tricuspid annular plane systolic excursion (TAPSE) [5]. More accurate values of RV ejection fraction from echocardiography can be attained only with 3-dimensional (3D) techniques. The complex evaluation also includes preload, afterload, and valvular conditions.

Years of training are required to interpret video loops and measurements from a full 2DE. Written echocardiography reports can be long and complex, containing both qualitative and quantitative information. A table of selected measurements is often included with the written report. However, data on functional parameters is often incomplete, leaving the explicit diagnostics of the RV embedded qualitatively in the text. A previous study aiming to extract data from echocardiography reports with natural language processing (NLP) found TAPSE in <5% of cases and could not find RV base diameter in their test set of 50 reports [6].

Supervised training of a deep learning model for image classification typically requires large numbers of images with high-quality annotations. Since annotations of RV size and function are often lacking in 2DE reports, this information must be extracted from the text. NLP is a branch of artificial intelligence that focuses on automatic interpretation of written and spoken language. Modern transformer-based NLP models excel at text comprehension [7]. Thus, an NLP model is an obvious candidate approach for extracting RV labels from the text of the 2DE echocardiography reports.

The majority of machine learning approaches in echocardiography mainly focus on LV segmentation [8,9]. To our knowledge, machine learning models for RV assessment in echocardiography have been used only for automatic tracking of the tricuspid annulus directly from videos [10] and for 3DE segmentation based on magnetic resonance images [11].

In this study, we sought to create a deep learning model, trained on auto-annotated image labels, to determine RV size and RV function automatically from 2DE video loops. We also wanted to show that NLP applied to the text of 2DE reports can be used to create auto-annotated image labels. In this way, large amounts of unlabeled examinations with free text reports can be converted to labeled images suitable for developing image classification models. This approach would be versatile and applicable to medical diagnostic fields with limited access to labeled image data.

2. Materials and methods

The study was a retrospective register study. All data was anonymized before use and informed consent was not obtained from the study subjects. This protocol was approved by the Clinical Medical Research Ethics Board of Sweden (ref. number: 818-18).

2.1. Model overview

A schematic overview of the pipeline we used and the models is presented in Fig. 1. Briefly, as described in detail below, a subset of reports was manually annotated and used to train two NLP text classification models, one for RV size and one for RV function, to automatically extract RV labels. Next, we developed an in-house view classifier to identify the four-chamber (4C) or RV-focused view from the echocardiographic examination. Then, we used the manually and automatically extracted RV size and function labels as “ground truth” to train a 3D convolutional neural network, using the 4C or RV-focused video loops as input. This resulted in two image classification models, one for RV size and one for RV function. Characteristics for the image classification training and test sets are shown in Table 1.

2.2. Data and inclusion criteria

For the manual classification and the NLP text classification models, we included 12,684 2DE examinations with corresponding written reports. Since we aimed to create an unbiased model, all examinations, even those with less-than-optimal image quality (reverberations, artefacts), were included that met following criteria: (1) 2DE was done at the Department of Clinical Physiology, Sahlgrenska University Hospital, Gothenburg, Sweden between January 1, 2007 and December 3, 2017; (2) report was written by an experienced physician, defined as >200 written reports within the time period; (3) examination done during regular office hours; (4) was first examination of an included patient during the time period; and (5) examination was done with a GE Healthcare ultrasound system.

To apply criterion 5, we first exported all examinations that met inclusion criteria 1 to 4 and in which left ventricular ejection fraction (LVEF) was reduced (n = 9456) or supranormal (n = 264) (males <52% or >72%, females <54% or >74%) [5] to include all examinations with impaired LV function. We also randomly selected 9156 examinations with normal LVEF to get a balanced sample. Application of criterion 5 to

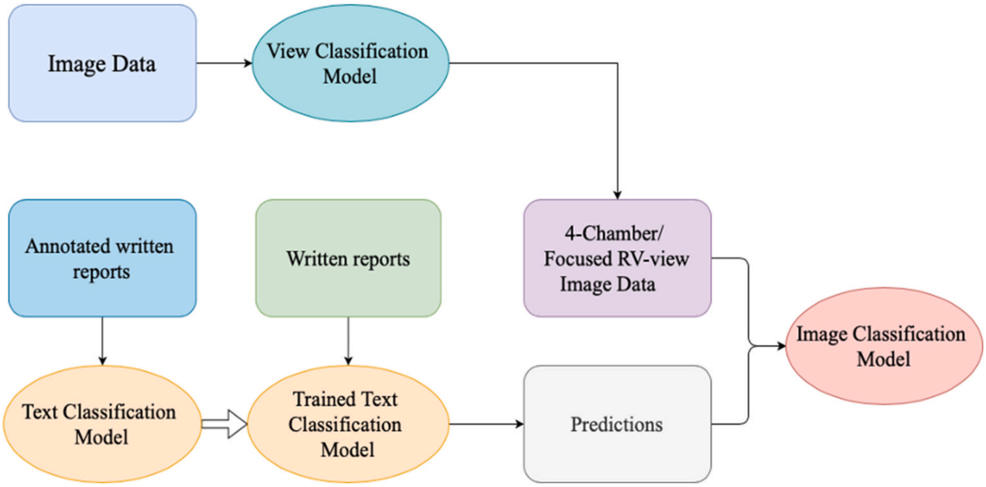
Image Data
View Classification Model
Annotated written reports
Written reports
4-Chamber/ Focused RV-view Image Data
Image Classification Model
Text Classification Model
Trained Text Classification Model
Predictions

Fig. 1. Schematic of model development. Manually annotated written reports were used to train the text classification model. The remaining written reports were classified by the trained text classification model. Examinations were processed by the view classifier, and all 4-chamber or RV-focused views were selected. Classifications from written reports plus 4-chamber images or RV-focused views were used as training data for the image classification models.

Table 1
Characteristics of the examinations used in the NLP text classification dataset and the image classification training and test sets.

	All examinations	Human expert annotated reports	Image classification training set		Image classification test set	
			RV function	RV size	RV function	RV size
Examinations (n)	12,684	1489	11,008	9561	300	300
Age (years), mean ± SEM	64.9 ± 16.3	70.5 ± 15.3	64.5 ± 17.8	70.5 ± 15.3	82.5 ± 12.4	82.0 ± 12.4
Female, % (*n*)	38% (4781)	37% (549)	37% (5084)	38% (3633)	40% (120)	39% (118)
Length (cm), mean ± SEM	173 ± 10.1 $n = 11,923$	172 ± 9.7 $n = 1355$	173 ± 10.1 $n = 10,347$	173 ± 17.5 $n = 10,348$	170 ± 10.0 $n = 270$	171 ± 10.1 $n = 280$
Body mass index (kg/m^2)	26 ± 4.9 $n = 11,923$	26 ± 4.9 $n = 1233$	26 ± 7.2 $n = 10,363$	27 ± 7.6 $n = 8605$	25 ± 3.8 $n = 267$	25 ± 4.2 $n = 282$
Heart rate (min^1), mean ± SEM	71 ± 21 $n = 11,440$	78 ± 19 $n = 1310$	74 ± 17 $n = 10,004$	74 ± 17 $n = 8490$	80 ± 20 $n = 270$	79 ± 19 $n = 276$
RV function classification, % (*n*)						
Impaired	15% (1941)	40% (592)	15% (1,604)	13% (1,212)	48% (144)	41% (122)
Normal	80% (10,096)	46% (692)	85% (9404)	86% (8246)	50% (151)	52% (155)
No information	4% (467)	14% (205)	0%	1% (103)	2% (5)	8% (23)
Excluded	1% (180)	0%	0%	0%	0%	0%
RV size classification, % (*n*)						
Impaired	10% (1310)	28% (420)	9% (1038)	10.5% (1005)	37% (113)	50% (149)
Normal	72% (9165)	47% (699)	77% (8485)	89.5% (8555)	49% (148)	49% (148)
No information	16% (2029)	25% (370)	14% (1485)	0% (1)	13% (39)	1% (3)
Excluded	1% (180)	0%	0%	0%	0%	0%
LVEF *						
Median (IQR)	50 (40–60)	45 (30–60)	51 (40–60)	55 (43–66)	40 (30–55)	45 (33–55)
Severely impaired, % (*n*)	12% (1504)	27% (397)	11% (1221)	10% (992)	32% (97)	28% (83)
Moderately impaired, % (*n*)	40% (5132)	38% (569)	40% (4472)	39% (3747)	44% (131)	46% (137)
Normal, % (*n*)	48% (6048)	35% (523)	48% (5315)	50% (4822)	24% (72)	27% (300)
sPAP >40 mm Hg, % (*n*)	26% (2219)	47% (563)	24% (1768)	23% (1459)	58% (175)	57% (150)

*Severely impaired:<30%; moderately impaired: 30–52% (females), 30–54% (males); normal >52% (females),>54%(males).
Measurements with missing data indicated as ($n =$ cases) with valid data in each variable. RV: right ventricular, LVEF: left ventricular ejection fraction, IQ: interquartile range, sPAP: systolic pulmonary artery pressure.

these 18,876 examinations resulted in a final dataset of 12,684 examinations (done with GE Vivid 6, Vivid 9, and Vivid E95 ultrasound systems).

For the view classification model, we selected from the 18,876 examinations all those from 2015. Of those, 630 met criteria described in the Supplement (section S4). Application of criterion 5 resulted in a final dataset of 539 examinations. A flowchart of the included patients for development of the view classifier is presented in Fig. 2.

An examination includes both image and text data, but these were never combined as one concatenated input. Instead, the text data was transformed by the text classification model into a label that can be used for supervised training of the image classification model. Thus, the text classification model extracts labels from text reports to generate training data, whereas the image classification model classifies images. The labels are only used for supervision, never as input data, for the two different models.

2.3. NLP text classification models

A subset of examinations was selected for manual annotation. To achieve a balanced model exposed to different RV conditions, we randomized patients from different ranges of pulmonary artery and central venous pressures and LVEF, as described in the Supplement (section S3.1). The text of the report was manually annotated with Prodigy (Explosion AI) by a single physician (EH) with 2 years of echocardiographic experience. RV function was classified as normal, reduced, or no information. Two NLP text classifiers, one for RV size and one for RV function, were trained to classify a written report. In the final model, we used a 75%/15%/10% split for the training/testing/validation sets. The RV size and RV function models were 12-layer BERT models, pre-trained on a large Swedish dataset [12,13] and fine-tuned on the annotated data. During training, the validation and test sets were hidden from the model. The validation set was used to optimize the hyperparameters and the architecture of the final model. Each data point in the validation set was given to the trained model, and the predictions were compared to the corresponding ground truth. Data selection and model development are described in detail in the Supplement (sections S3.1 and S3.2). Characteristics of the NLP text classification dataset are shown in Table 1.

2.4. View-classification model

A complete echocardiographic exam can consist of >25 2DE videos, but only some are feasible for RV assessment [14]. In recent years, several view-classification models for 2 DE have been published [15,16]. We developed an in-house view-classification model, using examinations selected as described in the Supplement (section S4.) The dataset was split into 70%/20%/10% training/testing/validation sets. The training set was manually annotated in equal shares by two physicians (EH, OH) with labels indicating the echocardiographic view. In the test set, all examinations were separately annotated by both physicians, who were blinded to each other.

The image data from the echocardiographic exams was extracted from the original DICOM file format with Pydicom Python library. The original frame size of 484 × 636 pixels was kept without rescaling, and data augmentation was not used. The model architecture was the ResNet50 model [17], imported from TensorFlow and initialized with pre-trained ImageNet weights [18]. The model was trained on individual frames with a batch size of four for five epochs with the lower layers frozen and a learning rate of 1e-3 and then for another five epochs with all layers unfrozen and a learning rate of 2e-5. The Adam optimizer [19] and a regularization factor of 1e-2 was used for all epochs. The hyperparameters were based on the most common hyperparameters for finetuning ResNet models and then tweaked slightly after multiple rounds of iteration. Categorical cross-entropy loss was used as the loss function. The final performance was evaluated on the unseen test set

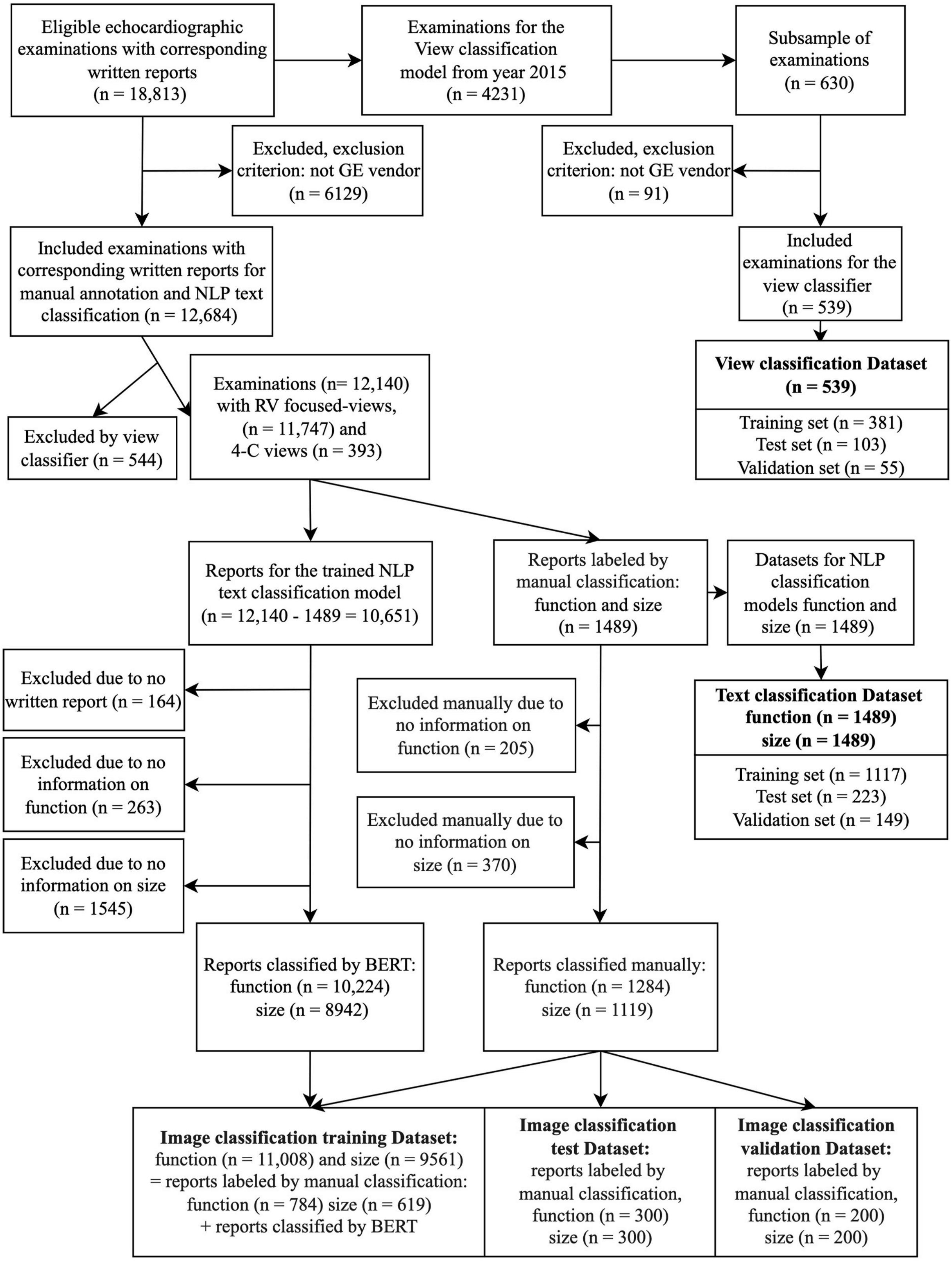
Eligible echocardiographic examinations with corresponding written reports (n = 18,813)
Examinations for the View classification model from year 2015 (n = 4231)
Subsample of examinations (n = 630)
Excluded, exclusion criterion: not GE vendor (n = 6129)
Excluded, exclusion criterion: not GE vendor (n = 91)
Included examinations with corresponding written reports for manual annotation and NLP text classification (n = 12,684)
Included examinations for the view classifier (n = 539)
View classification Dataset (n = 539)
Training set (n = 381)
Test set (n = 103)
Validation set (n = 55)
Excluded by view classifier (n = 544)
Examinations (n= 12,140) with RV focused-views, (n = 11,747) and 4-C views (n = 393)
Reports for the trained NLP text classification model (n = 12,140 - 1489 = 10,651)
Reports labeled by manual classification: function and size (n = 1489)
Datasets for NLP classification models function and size (n = 1489)
Text classification Dataset function (n = 1489) size (n = 1489)
Training set (n = 1117)
Test set (n = 223)
Validation set (n = 149)
Excluded due to no written report (n = 164)
Excluded due to no information on function (n = 263)
Excluded due to no information on size (n = 1545)
Excluded manually due to no information on function (n = 205)
Excluded manually due to no information on size (n = 370)
Reports classified by BERT: function (n = 10,224) size (n = 8942)
Reports classified manually: function (n = 1284) size (n = 1119)
Image classification training Dataset: function (n = 11,008) and size (n = 9561) = reports labeled by manual classification: function (n = 784) size (n = 619) + reports classified by BERT
Image classification test Dataset: reports labeled by manual classification, function (n = 300) size (n = 300)
Image classification validation Dataset: reports labeled by manual classification, function (n = 200) size (n = 200)

Fig. 2. Flowchart with step-by-step description of the echocardiographic examinations used.

using majority vote over all frames in a video sequence; that is, the most frequent class over all frames was assigned to the video. For detailed performance on different classes, see Supplemental Table 2.

2.5. Image classification models

The RV size and RV function labels predicted by the text classifier and the selected 4C or RV-focused videos formed the datasets used to develop the image classification models. The trained view classifier was applied to the corresponding video loops to find the videos of these views.

We trained two image classification models, one for RV size and one for RV function. For each model, a 4-layer, 3D convolutional neural network (CNN) was trained by using the 4-C or RV-focused video loops as input data and the RV text classifications as ground truth. Separate versions of the models were trained on (1) the dataset of manually labeled written reports, (2) the auto-annotated dataset, and (3) both datasets combined. The echocardiograms used as training data for the image classifier were pre-processed in several steps, and data augmentation was done before and during training. Characteristics of the image classification training dataset are shown in Table 1.

First, the image data was extracted from the original DICOM files and converted from RGB to grayscale. The frames per second rate for each video was rescaled to 12, lower than the median of 25, to reduce the size of the input data and to ensure that each frame corresponded to the same time span. Next, the videos were truncated to the first 20 frames, and each frame was rescaled from the original size (484 × 636 pixels) to the final frame size (169 × 222 pixels). To prevent class imbalance, the data was duplicated to balance the training data with respect to the most common class. In a pre-processing step, each frame was roughly centered on the RV by removing the bottom 20% of rows and the right 35% of columns of pixels from each frame, resulting in a frame size of 136 × 145 pixels. Centering removed the black outline in echocardiograms that contains no medical information, thereby lower input size and speeding up the training. The grayscale value of each pixel in every frame was normalized to values between 0 and 1. To reduce overfitting, data augmentation was used for each video (i.e., additive Gaussian noise, random brightness shifts, and random vertical and horizontal translations).

The model has a custom CNN architecture built with PyTorch (Fig. 3). It uses four convolutional blocks in series followed by a final fully connected classification layer. Each convolutional block consists of four layers in a fixed order: a 3D convolutional layer, a batch normalization layer, a leaky rectified linear unit activation layer, and a max pooling layer. The four convolutional blocks were identical except for the input and output size of the 3D convolutional layer. The architecture was developed through multiple rounds of iteration and testing; four layers proved to be optimal. An increase in the number of filters, especially in the later layers, improve the performance. However, the number of filters was limited by the memory of the graphics processing unit (GPU). Other choices of architecture (*e.g.*, design of batch normalization, rectified linear activation function, and max pooling) follow common best practices in computer vision architecture. Variants of this architecture and other existing architectures such as ResNext [20] were tested and evaluated but proved inferior to this custom 3D architecture (Table 2).

Table 2
Trained and evaluated networks for the image classification model.

Problem	Model	Average test accuracy
Size	ResNet18 (flattened video)	71%
Size	ResNet18 (single frame)	79%
Size	ResNext 3D	80%
Size	**Custom 3D CNN**	**83%**
Size	Custom 2D CNN	73%
Mobility	ResNet18 (flattened video)	72%
Mobility	ResNet18 (single frame)	76%
Mobility	ResNext 3D	81%
Mobility	**Custom 3D CNN**	**82%**
Mobility	Custom 2D CNN	74%

3D: 3-dimensional; CNN: convolutional neural network.

Several models based on 2D architectures were trained and evaluated but were outperformed by the 3D models. The best performing 2D model, a ResNet18 architecture trained on single frames, had 79% accuracy in classifying RV size and 76% accuracy in classifying RV function. The 2D models used the same video compression as the 3D models, and the difference is therefore most likely due to inclusion of the temporal dimension.

Because of the 3D structure of the data, GPU memory was a limiting factor for both model and input. Experiments were done both with more-complex models and low-resolution images and with less-complex models and high-resolution images to find the best performing configuration.

The final model converged on the validation set after 30 epochs with a batch size of 32 using cross-entropy loss. The Adam optimizer was used with a learning rate of 1e-3 and a weight decay of 1e-4. The hyperparameters were selected after an iterative process that compared the final validation results of models trained on different sets of parameters. After the training, the final performance was evaluated on the unseen test set.

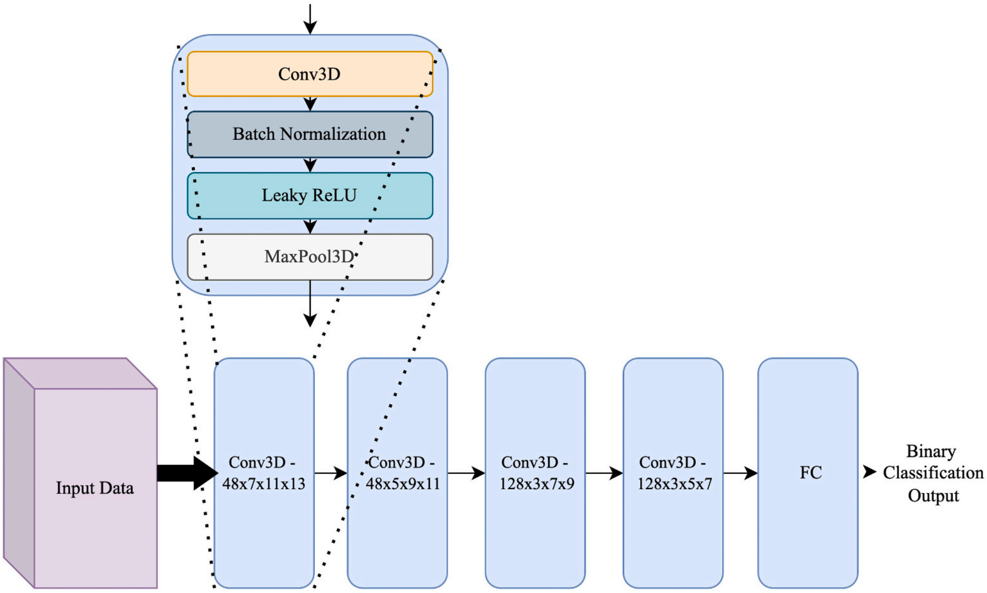

Fig. 3. Architecture of the convolutional neural network used for image classification. It consists of four convolutional blocks, each starting with a convolutional (Conv) layer followed by a batch normalization layer, a leaky rectified linear unit (ReLU) activation layer, and a max pooling layer. The number and size of the filters used in the convolutional layers are denoted on each block. The output from the last block is fed through a fully connected (FC) layer to produce the final classification result.

2.6. Final test set and interobserver study

To test the final image classification models for RV size and function assessment from 2DE video loops, we formed a test dataset that was unavailable during model training. We randomly selected two datasets, each consisting of 300 examinations, 50% with normal RV function and 50% with reduced RV function or dilation (Table 1). We then tested the accuracy of the two human interpreters against each other and against the model. For this interobserver study, we randomized a subset of the test set ($n = 100$). This subset was randomized and balanced with respect to RV function (50% normal/50% impaired) and roughly balanced with respect to RV size (60% normal/40% enlarged). The subset was selected from the echocardiograms previously labeled in written reports as described above (section 2.3). From these 2DE 4C or RV-focused video loops, RV size and function were classified by two physicians (EH, OH). The classifications were done blinded and independently.

3. Statistics

Agreement was assessed primarily with Cohen's kappa coefficient. The confidence intervals were computed by the method of Clopper and Pearson. Statistical analyses were done with SPSS statistics v. 25, R studio 1.2.5001 and GraphPad Prism 8.

4. Results

4.1. Material

Of 91,561 echocardiographic examinations done at the Department of Clinical Physiology, Sahlgrenska University Hospital during the study period, 52,628 met inclusion criteria 1 to 4. Of those, 9470 met criterion 5, and an additional 9156 examinations with normal LVEF were randomly selected. Thus, 12,684 2DE examinations were included for the NLP text classification model (Table 1). A subsample of 539 examinations was included for the view classifier as described above (section 2.2). Sample preprocessing included separating video loops from still images, extracting meta-data and written reports, and anonymizing all examinations.

4.2. NLP text classification models

We applied the view classifier to all 12,684 examinations to identify 4C or RV-focused views. If no such view was found, the examination was excluded from the dataset. This approach resulted in 12,140 4C or RV-focused views that could be used in the NLP text classification models.

Initially, a subset of the examinations, selected as described in the Supplement (section 3.1) was annotated ($n = 1197$). A second set, selected with the help of a separate NLP model trained on the initial dataset to identify RV dysfunction, was annotated to increase the ratio of reports indicating RV dysfunction ($n = 367$). The dataset then consisted of reports from both rounds of annotation ($n = 1564$). Seventy-two reports were excluded for lack of a 4C or RV-focused view, leaving 1489 for training of the final BERT model. In the flowchart (Fig. 2), the selections described above are simplified, and the 72 reported excluded after the second round of annotations are included in the total number excluded by the view classifier. The final characteristics of the annotated examinations are presented in Table 1.

Table 3
Performance of the best NLP text classification models.

Class	Precision	Sensitivity	Specificity	Accuracy
RV function				
Normal	0.97	0.97	0.98	0.97
Reduced	0.95	0.99	0.96	0.97
No information	0.90	0.79	0.99	0.84
RV size				
Normal	0.98	0.95	0.98	0.97
Increased	0.94	0.98	0.98	0.96
No information	0.92	0.92	0.97	0.92

NLP: natural language processing, RV: right ventricular.

RV size was classified as normal, enlarged, or no information. RV function was classified as normal, reduced, or no information. Several architectures for the text classifier were trained and evaluated, as described in the Supplement (section S3.2). The best-performing model was based on the BERT architecture and pre-trained on a large Swedish cohort [12]. This model had a sensitivity of 99% and a specificity of 98% in classifying impaired function and a sensitivity and specificity of 98% in classifying an enlarged RV (Table 3).

4.3. View classification model

The view classifier was trained on a subset of the dataset, selected as described in the Supplement (section S4). The best-performing view model was a ResNet50 model. Other architectures was evaluated during development, such as VGG16 [21] and MobileNet [22]. A model trained using a custom-built 3D architecture was tested but performed significantly worse than the 2D models with 76% average accuracy over all views. This could be attributed to the fact that the videos had to be compressed significantly in size and length to fit in GPU memory. It is also possible that classifying the view of a video loop is a relatively time-independent problem.

The view classification model found the 4C video loop with 92% accuracy and the RV-focused view with a 73% accuracy (Supplemental Fig. S1). Because of this discrepancy, the 4C view was selected as the primary view to be used in the image classification set; the RV-focused view was used only when a 4C view was not found in an examination.

4.4. Image classification models

In total, 10,651 written reports were fed into the NLP text classification models (Fig. 2). In 164 cases, the report turned out to be missing. No information on RV function could be extracted from 263 reports, and no information on RV size could be extracted from 1545 reports. Thus, after the view classifiers and text classifiers were applied, and manually annotated reports were added, the training dataset for image classification included 11,008 examinations for RV function (10,224 from BERT annotation, 784 from manual annotation) and 9561 examinations for RV size (8942 from BERT annotation, 619 from manual annotation) (Fig. 2).

The model was a custom CNN architecture built with PyTorch as described above in section 2.4. In the final test set ($n = 300$), the RV function model had a sensitivity of 93%, specificity of 72%, and accuracy of 82% (95% CI 0.78–0.86) and showed substantial agreement with written reports ($\kappa = 0.65$; 95% CI 0.56–0.73) (Table 4). McNemar's test showed a significant systematic difference between model prediction and manually annotated written report ($P < 0.001$), indicating that the

Table 4
Performance of image classification models.

Ground Truth	Model: RV function (n)	
	Normal	Reduced
Normal function	108	42
Impaired function	11	139
	Model: RV size (N)	
	Normal	Dilated
Normal Size	127	23
Enlarged	30	120

In the final test, the RV function model had an accuracy of 82% (95% confidence interval [CI] 0.78–0.86) and kappa of 0.65 (95% CI 0.56–0.73). The RV size model had an accuracy of 82% (95% CI 0.79–0.87) and kappa of 0.65 (95% CI 0.56–0.73). RV: right ventricular.

Table 5
Image classification model accuracy and dataset size.

Dataset	Dataset size (n)	Model accuracy (%)
RV size		
Manually annotated	819	80%
Model-annotated	8942	80%
Combined	9561	83%
RV function		
Manually annotated	984	82%
Model-annotated	10224	81%
Combined	11008	82%

The size of the combined datasets is not the exact sum of the manually annotated datasets and the model-annotated datasets, since the final model was also trained on their validation data. RV: right ventricular.

model tends to classify normal cases as impaired. In the test set, the RV size model had a sensitivity of 80%, specificity of 85% (Table 4), and accuracy of 82% (95% CI 0.79–0.87) and showed substantial agreement with written reports ($\kappa = 0.65$; 95% CI 0.56–0.73). McNemar's test ($P = 0.41$) showed no systematic differences.

We also compared manually annotated versus auto-annotated medical reports as ground truth when training image classification models. Both alternatives resulted in comparable accuracies (Table 5). Using both automatically and manually annotated examinations slightly improved the accuracy of the RV size model but not of the RV function model.

Saliency maps were used to verify that the models focused on relevant parts of the video loop (Fig. 4). Each pixel in a saliency map corresponds to the sum of gradient magnitudes generated by that pixel when feeding an image to the CNN model. The gradient is a multidimensional derivate that describes how much a function changes if the inputs are changed. In this case, the function is the whole network. This sum of gradient magnitudes is visualized in the form of a heatmap, where a stronger color intensity indicates a stronger gradient for the corresponding input pixel. The saliency map can help interpret how each pixel affects the classification results.

4.5. Human interobserver agreement

Agreement of RV function assessments by the two readers was substantial ($\kappa = 0.72$; 95% CI 0.59–0.85) and accuracy was 86% (95% CI 0.78–0.92; $P \leq 0.05$, McNemar's test). Agreement of RV size assessments by the two readers was also substantial ($\kappa = 0.65$; 95% CI 0.49–0.81) and accuracy was 85% (95% CI 0.76–0.91; $P = 0.30$, McNemar's test). Comparisons of the agreement between human readers, RV image classification models, and human annotated written reports are summarized in Table 6. The highest accuracy was seen in the reader-to-reader comparisons for RV function (85%) and RV size (86%). For RV function, the accuracy of the model vs ground truth (written reports) and that of reader 1 vs ground truth were both 82%. For RV size, the accuracy of the model vs ground truth was higher than that of reader 1 (82% vs 72%).

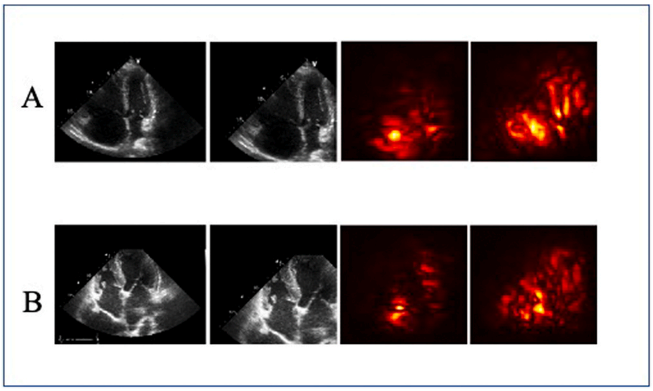

Fig. 4. Saliency maps for model relevance verification. **(A)** Dilated right ventricle (RV) with reduced function. **(B)** RV with normal size and function. Column 1: Original 2D images. Column 2: 2D images as seen by the models after preprocessing (down-sampling and centered on the right side). Column 3: Saliency maps indicating areas of interest for the RV function model. Column 4: Saliency maps indicating areas of interest for the RV size model.

Table 6
Interobserver agreement.

Comparison	Accuracy	Kappa
RV function		
Human-annotated written report vs model	82% (0.78–0.86)	0.65 (0.56–0.73)
Reader 1 vs. reader 2	86% (0.78–0.92)	0.72 (0.59–0.85)
Human-annotated written report vs. reader 1	82% (0.73–0.89)	0.64 (0.49–0.79)
RV size		
Human-annotated written report vs. model	82% (0.78–0.86)	0.65 (0.56–0.73)
Reader 1 vs. reader 2	85% (0.76–0.91)	0.65 (0.49–0.81)
Human-annotated written report vs. reader 1	72% (0.62–0.81)	0.44 (0.27–0.61)

Values in parentheses are 95% confidence intervals. RV: right ventricular.

5. Discussion

Large volumes of data are continuously generated from diagnostic studies such as echocardiographic examinations. This study shows that NLP can be used to generate labels for deep learning training sets and that auto-annotation can be integrated in NLP applications, making the approach scalable. We found that such models can be trained on existing medical data (*i.e.*, text reports) without need for extensive manual annotation. Do note that the text and image classification models are separate entities, with separate inputs (text reports and images respectively) and designed to be used in a development pipeline; a text model produces labeled image data from the clinical archives, and the labeled data is used to train two image classification models.

The use of both automatically and manually annotated examination reports improved the accuracy of the RV size image classification model only slightly over use of manually annotated reports alone, and did not improve not the accuracy of the RV function image classification model, perhaps because the model was close to saturated and did not need more labeled data. The performance of the final model was very close to the interobserver rate for experts. Another possibility is that incorrect labeling in the initial annotation, a manual task performed by a human, affected the model negatively. Nevertheless, we show that the auto-annotated dataset is suitable for model training. Large-scale model training on auto-annotated datasets would probably have generated better results if the dataset were smaller and without annotation errors.

Our results show that 3D models were superior to 2D models for classifying the 3D data. The difference was smaller for assessment of RV size. The single-frame 2D model performed surprisingly well on the size-classification task, and 2D models for echo assessment should be considered, depending on the classification problem at hand. During training, the models extract the image features they find relevant.

The saliency maps (Fig. 4) show that the RV models mostly seem to focus on what we generally believe are the most important heart structures for the assessment tasks. In assessing RV size, the model seems to focus the whole area of the RV. In assessing RV function, the model seems to focus on the horizontal movement close to the basal free lateral wall, the septum, and the free RV wall, but it also focuses on the intra-atrial septum in some images. Evidently, the model extracts relevant information, mirroring the physician's assessment of RV function as judged from TAPSE or S′, but the model also seems to be sensitive to some image features that would not be noted by a human. Humans typically interpret an image by assessing meaningful subregions, whereas a CNN processes image features rather than image semantics. Thus, image patterns deemed semantically irrelevant by humans can contribute to the CNN classification results [23].

Clinical echocardiographers often describe RV abnormalities as mild, moderate, or severe. It would have been interesting to train a model for these functional classes. However, cut-off values for these classes are not internationally standardized, and the results would have been hard to interpret [5]. Further, the reports were written by experienced physicians according to international guidelines implemented in the clinic between 2007 and 2017. During this period, echocardiographic technology and knowledge evolved, and the recommendations for echocardiographic chamber quantification from ASE from 2015 [5] differ slightly from those in guidelines from 2005 [24]. We still believe that the use of real-life data as ground truth makes the models reliable and relevant to a clinical setting.

The 2DE video loops and reports we used were collected directly from hospital records. Our automatic models for classifying RV function and RV size performed as well as an experienced physician. However, the physicians responsible for the interobserver test had access only to the single 4C or RV-focused view, whereas the authors of the echocardiographic reports had access to the full examination. Thus, the physician's performance may have been closer to ground truth if they had access to the full examination.

A further developed RV model could be integrated in a decision support "alert" model. This decision support would be of true value and give a recommendation for further evaluation of the RV, which is important because RV dysfunction can be missed, especially by a less experienced user.

6. Limitations

Our study had several limitations. First, we did not test the model on a second set of 2DE video loops from an external cohort [25]. Second, the videos in our dataset did not have view labels, so we had to train an additional model to generate them for the dataset. This step may have biased selection toward higher-quality examinations, as examinations where the view classifier could not find a relevant view were excluded from the final dataset used for image classification. Moreover, only the 4C or RV-focused view was used to train the image classification models. It might have been preferable to use the RV-focused view as the main view even though the view classifier performed better on the 4 C view.

Since the shape and wall motion of the RV can vary considerably, particularly at the apex, caution is needed when diagnosing an abnormal RV from a single tomographic plane. Adding a few selected views, such as the parasternal long-axis view and the short-axis views, could improve the models. The use of input from several views for model training is a topic for further research.

With respect to the NLP text model, the BERT model was pre-trained on a Swedish text dataset, but the actual language in the reports is mostly medical jargon, which is less abundant in the pretraining text dataset. Using a BERT model pre-trained on a large text dataset of generic medical texts and reports would probably have improved performance. The model architecture is not dataset specific, but the pre-trained weights are specifically Swedish. Someone with an English dataset could still use the same BERT architecture but with a set of pre-trained English weights. This would most likely lead to slightly better results, as the English weights are trained on an even larger corpus.

In a pre-processing step for the image classification model, the borders of the frames were removed to decrease the runtime and memory footprint of the training. The frames were not aligned to a specific point in the image, and the border thickness was chosen very carefully after manual inspection of several videos to ensure minimal information loss. Nevertheless, some information might have been lost. Model performance might be slightly improved by better alignment of each video and a more refined cropping method.

7. Conclusion

We developed a deep learning model to automatically assess RV size and function from 2DE video loops, solving a task of true patient value in a field where annotated qualitative data is sparse and often incomplete. To this end, we propose a pipeline for auto-annotation of the 2DE video loops with an NLP model, using the medical reports as input. This pipeline can be used to train a video-loop-assessment model without manual image annotation, enabling fast and inexpensive expansion of the training dataset when needed. Training an image assessing model with auto-annotated training data is feasible for echocardiographic classification tasks. To be able to handle the full examination, we also propose a model for view classification of the 2DE video loops. The results of this study open the door for the development of image analysis tools that use existing medical records to auto-annotate the training data. That is, this idea can be extended to image assessment models all over the medical field.

Declation of competing interest

The authors report no conflict of interest.

Acknowledgements

We thank Anders Arvidsson for preparing the metadata and Stephen Ordway for editorial assistance. This study was supported by the Sahlgrenska Academy, University of Gothenburg; Sahlgrenska University Hospital; Gothenburg Medical Society and Chalmers University of Technology.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2022.105282.

References

[1] A. Pueschner, P. Chattranukulchai, J.F. Heitner, et al., The prevalence, correlates, and impact on cardiac mortality of right ventricular dysfunction in nonischemic cardiomyopathy, JACC Cardiovasc Imaging 10 (2017) 1225–1236.

[2] L.A. Zornoff, H. Skali, M.A. Pfeffer, et al., Right ventricular dysfunction and risk of heart failure and mortality after myocardial infarction, J. Am. Coll. Cardiol. 39 (2002) 1450–1455, https://doi.org/10.1016/s0735-1097(02)01804-1.

[3] S. Ghio, C. Raineri, L. Scelsi, M. Asanin, M. Polovina, P. Seferovic, Pulmonary hypertension and right ventricular remodeling in HFpEF and HFrEF, Heart Fail. Rev. 25 (2020) 85–91.

[4] F. Haddad, R. Doyle, D.J. Murphy, S.A. Hunt, Right ventricular function in cardiovascular disease, part II: pathophysiology, clinical importance, and management of right ventricular failure, Circulation 117 (2008) 1717–1731.

[5] R.M. Lang, L.P. Badano, V. Mor-Avi, et al., Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging, Eur Heart J Cardiovasc Imaging 16 (2015) 233–270.

[6] C. Nath, M.S. Albaghdadi, S.R. Jonnalagadda, A natural language processing tool for large-scale data extraction from echocardiography reports, PLoS One 11 (2016), e0153749.

[7] J. Devlin, M.-W. Chang, L. Kenton, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018. https://arxivorg/abs/181004805. (Accessed 2 February 2020).

[8] J. Zhang, S. Gajjala, P. Agrawal, et al., Fully automated echocardiogram interpretation in clinical practice, Circulation 138 (2018) 1623–1635.

[9] C. Chen, C. Qin, H. Qiu, et al., Deep learning for cardiac image segmentation: a review, Front Cardiovasc Med 7 (2020) 25.

[10] A.N. Beecy, A. Bratt, B. Yum, et al., Development of novel machine learning model for right ventricular quantification on echocardiography—a multimodality validation study, Echocardiography 37 (2020) 688–697.

[11] D. Genovese, N. Rashedi, L. Weinert, et al., Machine learning-based three-dimensional echocardiographic quantification of right ventricular size and function: validation against cardiac magnetic resonance, J. Am. Soc. Echocardiogr. 32 (2019) 969–977.

[12] The National Library of Sweden. Swedish BERT models, https://github.com/Kungbib/swedish-bert-models. Accessed March 12, 2020.

[13] Malmsten M, Börjeson L, Haffenden C. Playing with words at the National Library of Sweden—making a Swedish BERT, https://ui.adsabs.harvard.edu/abs/2020arXiv200701658M. Accessed March 13, 2020.

[14] C. Mitchell, P.S. Rahko, L.A. Blauwet, et al., Guidelines for performing a comprehensive transthoracic echocardiographic examination in adults: recommendations from the American Society of Echocardiography, J. Am. Soc. Echocardiogr. 32 (2019) 1–64.

[15] A. Madani, R. Arnaout, M. Mofrad, R. Arnaout, Fast and accurate view classification of echocardiograms using deep learning, NPJ Digit Med 1 (2018) 6.

[16] K. Kusunose, A. Haga, M. Inoue, D. Fukuda, H. Yamada, M. Sata, Clinically feasible and accurate view classification of echocardiographic images using deep learning, Biomolecules 10 (2020) 665.

[17] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition, https://ui.adsabs.harvard.edu/abs/2015arXiv151203385H. Accessed May 17, 2020.

[18] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[19] D. Kingba, J. Ba, Adam, A Method for Stochastic Optimization, 2014 arXiv: 14126980.

[20] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated Residual Transformations for Deep Neural Networks, 2016, p. 161105431, arXiv.

[21] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014, p. 1409556, arXiv.

[22] A. Howard, M. Zhu, B. Chen, et al., Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications, 2017, p. 170404861, arXiv.

[23] A. Alqaraawi, M. Schuessler, P. Weiss, E. Costanza, N. Berthouze, Evaluating Saliency Map Explanations for Convolutional Neural Networks: a User Study, 2020, p. 200200772, arXiv.

[24] R.M. Lang, M. Bierig, R.B. Devereux, et al., Recommendations for chamber quantification: a report from the American society of echocardiography's guidelines and standards committee and the chamber quantification writing group, developed in conjunction with the European association of echocardiography, a branch of the European society of cardiology, J. Am. Soc. Echocardiogr. 18 (2005) 1440–1463.

[25] X. Liu, L. Faes, A.U. Kale, et al., A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis, Lancet Digital Health 1 (2019) E271–E297.