Computerized Medical Imaging and Graphics 99 (2022) 102085

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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

End-systole and end-diastole detection in short axis cine MRI using a fully convolutional neural network with dilated convolutions

Manuel Pérez-Pelegrí [a], José V. Monmeneu [b], María P. López-Lereu [b], Alicia M. Maceira [b], Vicente Bodi [c,*], David Moratal [a,**]

[a] Center for Biomaterials and Tissue Engineering, Universitat Politècnica de València, Valencia, Spain
[b] Cardiovascular Unit. Ascires Biomedical Group., Valencia, Spain
[c] Cardiology Department, Hospital Clínico Universitario de Valencia, Valencia, Spain

ARTICLE INFO

Keywords:
Cardiac magnetic resonance
Deep learning
Left ventricle
Dilated convolutions
Frame classification

ABSTRACT

The correct assessment and characterization of heart anatomy and functionality is usually done through inspection of magnetic resonance image cine sequences. In the clinical setting it is especially important to determine the state of the left ventricle. This requires the measurement of its volume in the end-diastolic and end-systolic frames within the sequence trough segmentation methods. However, the first step required for this analysis before any segmentation is the detection of the end-systolic and end-diastolic frames within the image acquisition. In this work we present a fully convolutional neural network that makes use of dilated convolutions to encode and process the temporal information of the sequences in contrast to the more widespread use of recurrent networks that are usually employed for problems involving temporal information. We trained the network in two different settings employing different loss functions to train the network: the classical weighted cross-entropy, and the weighted Dice loss. We had access to a database comprising a total of 397 cases. Out of this dataset we used 98 cases as test set to validate our network performance. The final classification on the test set yielded a mean frame distance of 0 for the end-diastolic frame (i.e.: the selected frame was the correct one in all images of the test set) and 1.242 (relative frame distance of 0.036) for the end-systolic frame employing the optimum setting, which involved training the neural network with the Dice loss. Our neural network is capable of classifying each frame and enables the detection of the end-systolic and end-diastolic frames in short axis cine MRI sequences with high accuracy.

1. Introduction

Cardiovascular diseases are one of the main causes of death in developed countries (Townsend et al., 2016; Lopez et al., 2021). In the clinical setting, the most reliable and accurate imaging technique for correctly assessing cardiac function is cardiac magnetic resonance (CMR), specifically with cine sequences which can show the motion of the entire heart. The cardiac short-axis sequences have shown to be accurate and reproducible for the assessment of the main biomarkers to characterize the function of the heart (Childs et al., 2011).

In the clinical setting, several parameters are used to characterize the left ventricle (LV), specifically its volume in end-diastole (ED), which corresponds to the state of maximum relaxation, and in end-systole (ES) which corresponds to the state of maximum contraction. Using these two values the ejection fraction can then be derived. In this context, some work has been done in the automation of LV segmentation in the end-systolic and end-diastolic frames, where convolution neural networks have shown especially good results (Poudel et al., 2017; Abdelmaguid et al., 2018; Tao et al., 2019; Tong et al., 2019; Chen et al., 2020; Perez-Pelegri et al., 2020; Pérez-Pelegrí et al., 2021). Still, the previous step of automatically detecting these frames, which is a required prior step, has not been studied so extensively and it is usually done manually, increasing the time required for the whole diagnosis.

Several works have addressed the issue of ED and ES detection in echocardiography imaging (Dominguez et al., 2005; Gifani et al., 2010; Shalbaf et al., 2011; Zolgharni et al., 2017; Meidellfiorito et al., 2018), but little work has been done with CMR. In the work of Kong et al. (2016) a convolutional neural network was used to extract spatial

* Correspondence to: Department of Cardiology, Hospital Clínico Universitario de Valencia, Avda. Blasco Ibanez 17, 46010 Valencia, Spain.
** Correspondence to: Center for Biomaterials and Tissue Engineering, Universitat Politècnica de València, Camí de Vera, s/n, 46022 Valencia, Spain.
E-mail addresses: vicente.bodi@uv.es (V. Bodi), dmoratal@eln.upv.es (D. Moratal).

https://doi.org/10.1016/j.compmedimag.2022.102085
Received 25 January 2022; Received in revised form 17 March 2022; Accepted 30 May 2022
Available online 3 June 2022

features from the images followed by the Long-Short-Term-Memory layers (LSTM) to encode the temporal information. Other works have focused on segmenting the entire LV and measuring its main parameters, such as volume (Hsin and Danner, 2016) or its location with respect a reference point (Yang et al., 2017) to determine ED and ES.

We now aim to design a fully convolutional neural network capable of detecting the ED and ES frames in a short-axis stack of cine CMR sequences with both an arbitrary number of frames and of slices per frame. The network uses dilated convolutions which have been applied to multiple deep learning tasks, especially in segmentation problems (Yu and Koltun, 2015; Chen et al., 2017, 2018). We based their use on the paradigm shift caused by the development of wavenet described by Oord et al. (2016) which demonstrated that the use of dilated convolutions could be used to encode temporal information and can surpass the different recurrent neural network layers usually employed to tackle temporal sequences.

2. Materials and methods

2.1. Image dataset

All patients gave written informed consent and the study was approved by the Institutional Review Board of our hospital. Our dataset consisted of 397 short-axis stacks of CMR cine sequences covering both the right and left ventricles along the whole cardiac cycle. This dataset comprised a total of 397 patients (270 men, 127 women), with age 64.53 $\pm$ 12.35 years (63.27 $\pm$ 11.98 years for men, 67.42 $\pm$ 12.75 years for women) (mean $\pm$ standard deviation), with both cardiac patients and healthy subjects. Main CMR findings of the patients were presence of myocardial fibrosis, necrosis, ischemia and LV systolic dysfunction (ejection fraction lower than normal and/or regional wall motion abnormalities). This is summarized in Table 1. CMR imaging was performed using a 1.5 T MRI scanner (Sonata Magnetom Siemens, Erlangen, Germany). All the acquisitions were done in end-inspiration during a breath-hold with the following typical sequence parameters: flip angle: 58°, repetition time: 52.92 ms, echo time: 1.25 ms. The in-plane resolution varied across the cases, ranging from 0.57 $\times$ 0.57 mm^2 to 1.09 $\times$ 1.09 mm^2. The slice thickness and spacing between slices was constant in all cases, 7 mm and 3 mm respectively. The resulting image sizes varied from 144 $\times$ 144–256 $\times$ 256 and the number of slices ranged from 8 to 14. The number of temporal frames in each sequence also varied in the dataset, the great majority included 35 frames (364 cases, 95% of the dataset). The remaining cases ranged from 14 to 25 frames. In all cases the sequence encompassed a single cardiac cycle. The time resolution between frames was of 0.023 s for the cases with 35 frames and varied from 0.062 to 0.078 s for the remaining cases.

Table 1
Classification of the dataset in categories according to its clinical diagnosis.

Categories	Number of cases
Normal cases, no pathology	48
Presence of necrosis	14
Presence of fibrosis	12
Presence of ischemia	10
Functional affection of LV (ejection fraction lower than normal and/or affected segmental contractility)	23
Functional affection of RV (ejection fraction lower than normal and/or affected segmental contractility)	2
Functional affection of LV and RV	135
Functional affection of LV and presence of fibrosis/necrosis/ischemia	45
Functional affection of RV and presence of fibrosis/necrosis/ischemia	4
Functional affection of RV and LV and presence of fibrosis/necrosis/ischemia	95
Other cases that do not fall in any other category	9

The dataset included the labels for the frames corresponding to ES and ED. This classification was done manually by mutual consensus of two expert cardiologists with more than 15 years of experience. For each image acquisition the two cardiologist discussed until reaching an agreement on the best labeling. For the majority of cases the time required for the labelling took around 2–3 min, which included the exploration of the images and the selection of the ES and ED, with little discussion and easy agreement. However, we note that some cases required a more profound discussion among them with longer decision times, this happened only for ES frames and the disagreement in frame distance was not larger than 1 frame. For these harder cases the additional required time was variable reaching up to around 5 min.

Every case of the entire dataset was categorized in one of the 11 categories corresponding to the diagnosis (see Table 1) in order to ensure that the split between training, validation and test sets had similar distribution with respect to diagnosis. Finally, all cases were randomly grouped in a training set (259 cases, 65%), validation set (40 cases, 10%) and test set (98 cases, 25%).

2.2. Image pre-processing

Prior to training the network the dataset was modified to reduce the problem complexity and to normalize the dataset. All the images were resampled to a constant in-plane resolution of 1 mm^2 and the image size was set to 176 $\times$ 176 pixels. Some images were cropped around the borders and others were zero-padded but in all cases the entire heart remained within the central region of the image. The z-axis and the time axis were left untouched. The intensity values for every volume stack was also normalized to a range of 0 and 1 using min-max normalization.

Additionally, the 4D stacks were not directly used as inputs for the network, due to the excessive memory consumption that this would require. Instead, we converted each 3D volume within the sequence to a single image. In order to do this, we applied the median value between the second and penultimate slice along the z-axis to generate a single image that represented the entire volume at each time frame. As a general rule, we did not include the first and last short axis slices since, usually, the LV is not fully present in all the frames along the cardiac cycle and also because these slices increased noise in the final image. This process modified our dataset from 4D stacks to 3D stacks, where the slices are the median representation of the volumes at specific frames and the z-axis represents the frames of the sequence. With this, memory requirements were reduced significantly while the information regarding cardiac contraction remained within the data. An example of the median images obtained in a sequence can be observed in Fig. 1 where the contraction and motility of the tissue is visually perceptible.

2.3. Neural network architecture

The architecture employed takes as input arrays of 176 $\times$ 176 $\times$ n being n the number of frames (time axis) and a variable number. The neural network initially applies several 2D convolutions using ReLU activation functions and max-pooling operations to extract the spatial features of the images and reduces their dimensions. These operations are implemented as 3D convolution of size 3 $\times$ 3 $\times$ 1 (equivalent to a 2D convolution) since the inputs are 3D arrays. Each convolution is always followed by batch-normalization (Ioffe and Szegedy, 2015; Santurkar et al., 2018) to improve the training speed and performance. After 4 steps of convolutions and downsampling, the result is a stack of channels of size 11 $\times$ 11 which are then collapsed to a single channel (using a 1 $\times$1 $\times$1 convolution). We then added different convolution paths which apply the operation to the time dimension as well. This is done with variable kernel sizes using 3D convolutions (3 $\times$ 3 $\times$ 3, 3 $\times$ 3 $\times$ 5 and 3 $\times$ 3 $\times$ 7) combined with different dilation rates of 1, 2 and 4 in the third axis. This makes a total of 9 paths, each one giving one channel as output. The combinations ensure that the field of view for the time dimension ranges from short term (minimum field of view of 3 frames)

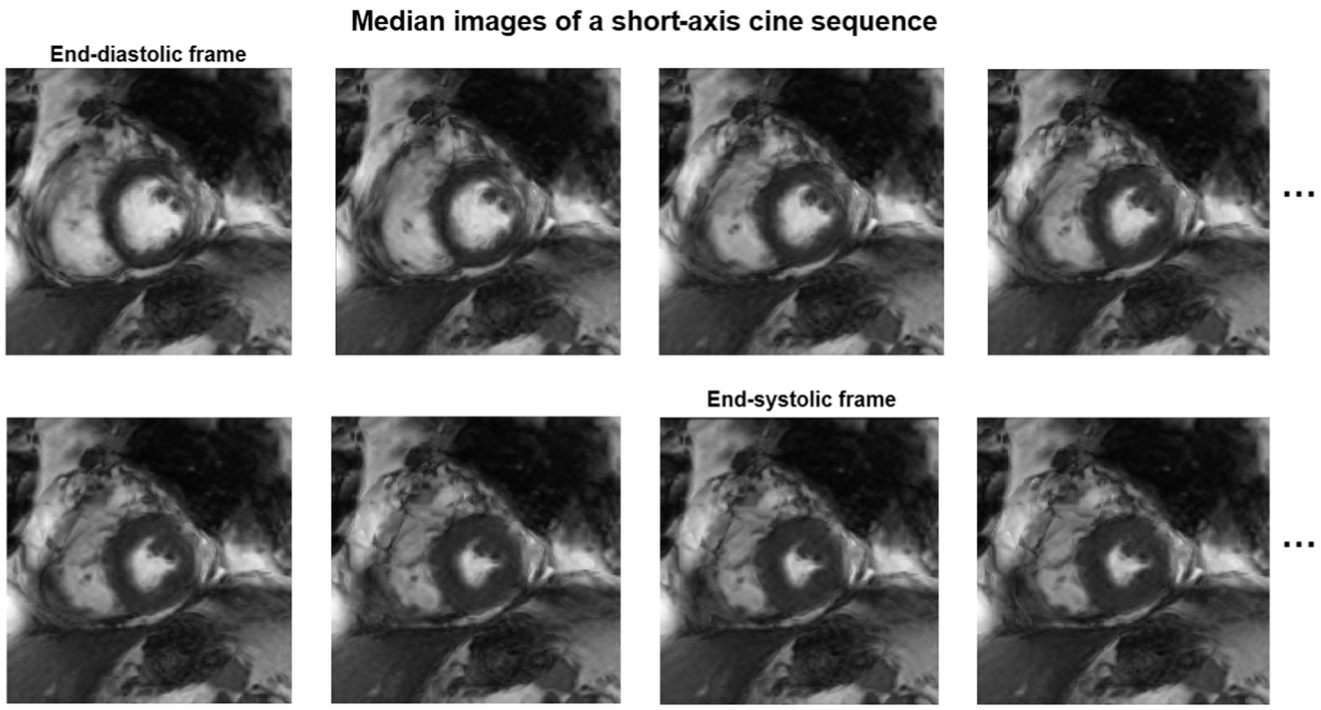

Fig. 1. Example of a sequence of the median representations of the volumes in each frame. The order is from left to right and from top to bottom. It can be seen that the contractility of the left ventricle is clearly visible. The ellipsis points at the right side indicate that the sequence continues. The end-systolic and end-diastolic frame are indicated above the images.

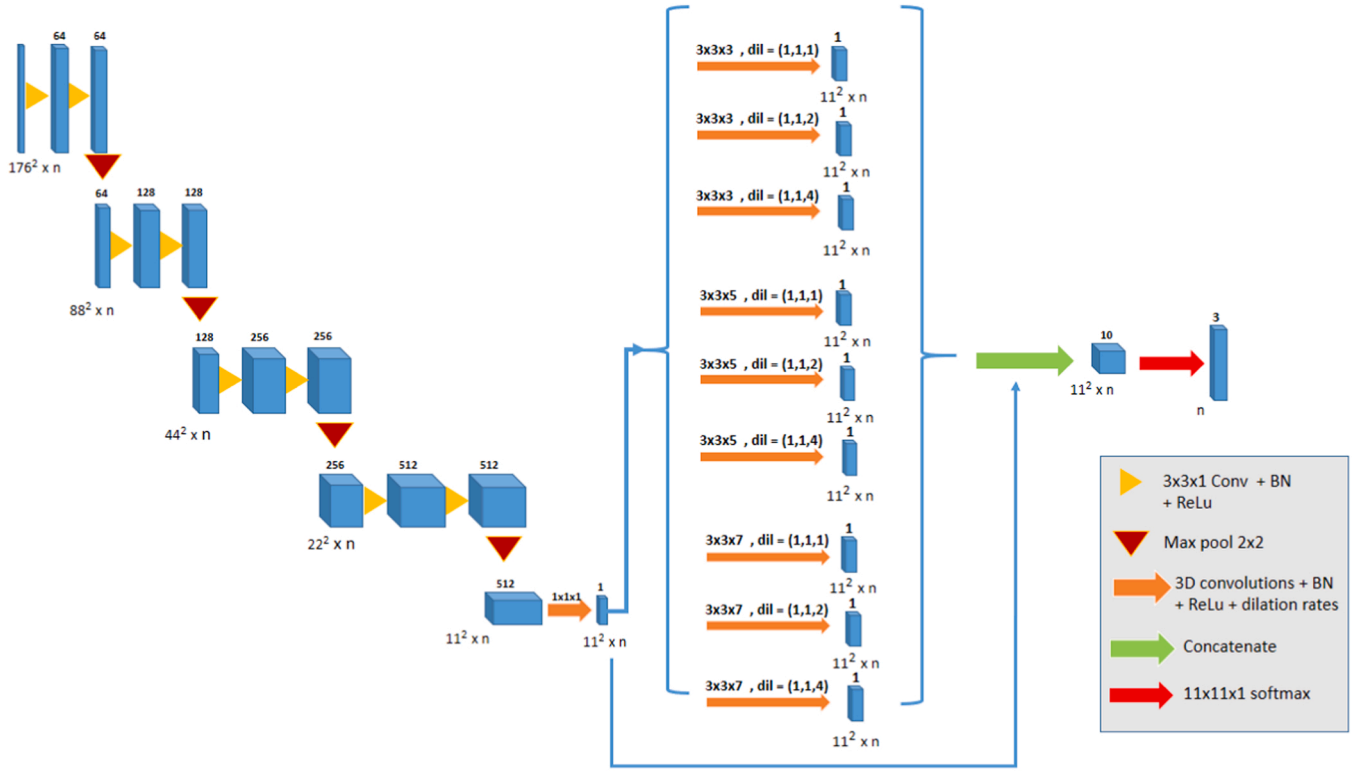

Fig. 2. Neural network architecture design. The first part of the network consists of several convolution layers and max pooling operations that process the spatial information. After this section a $1 \times 1 \times 1$ convolution collapses the previous feature maps to one single feature map. This is then passed to the second section with different dilated convolution paths. The dilation is applied only to the temporal dimension (third dimension) and each path produces one single feature map. The resulting feature maps are concatenated and a final convolution layer collapses the size to a single value per frame which is finally given a probability through a softmax activation function.

to long term (maximum field of view of 25 frames).

The core idea of using as much as 4 downsampling steps prior to the time-axis analysis is based on the fact that the heart occupies a large amount of space within the image in the more relaxed states (see Fig. 1) and thus the more downsampling steps applied the larger the spatial field of view the network can use in the temporal analysis. Additionally, we applied up to 4 steps in order for the feature maps to remain with an even matrix size and avoid using padding or cropping steps in the architecture.

Finally, the resulting outputs are concatenated and a last $11 \times 11 \times 1$ convolution followed by a softmax activation function is applied. The softmax outputs 3 classes which gives the probability of each frame being the end-systole frame, end-diastole frame or none of them (a background frame). These three probabilities always add up to 1. The entire architecture is schematized in Fig. 2. This architecture had a total of 4.7 million parameters and occupied 54 MB in HDF5 format.

2.4. Frame classification

The neural network outputs a list containing for all the frames a probability associated to each category: ES, ED or none. However, the neural network tended to give outputs where there were multiple frames with high probabilities of being ES and ED near those frames. This is explained by the fact that the differences in the contraction between adjacent frames is very subtle, and as such it is easy for the network to give high probabilities to frames that are close to the ED and ES. However, only one frame must be classified as ED or ES. We tested two classification methods to select one frame as ED and another for ES. The first one classified as ED or ES the frames with the highest probability assigned to those categories (naïve method). The second method took all the frames with a high probability of being ED or ES and then classified the central frame amongst those as the ED or ES (central method). The central method used a probability threshold of 90% to select candidate frames to be classified as either ES or ED. Fig. 3 shows a schematic of how the central method works.

2.5. Data augmentation

The training dataset employed was increased through data augmentation methods. We increased the total amount of cases by a factor of 7 (comprising a total of 1813 cases to train with). The dataset was increased applying random rotations around the image center (between +20 and −20 degrees), random shear (between 10 and −10 degrees) and random translations in both x and y axis (between 44 and −44 pixels). When applying these transformations, the images were zero-padded or cropped as needed in order to maintain the image dimensions to feed the neural network. We also applied a random delay in the time sequence to displace the location of the end-systolic and end-diastolic frames within the sequence (the delay was randomly applied between 0% and 40% of the number of frames within the sequence).

2.6. Choice of loss functions

The network was trained in two settings employing two different loss functions. The first one was a weighted sum of the categorical cross entropy of the classified frames (Eq. 1). In order to achieve acceptable results, we gave a big weight to the cross entropy given by the frames corresponding to systole and diastole. All the frames were given a weight of 1 and the systolic and diastolic frames were given a weight of 100 to help the network focus on correctly classifying those 2 frames. We tested with different weights but only achieved satisfactory results with these settings. This loss function is presented in Eq. 1. In the formulae Y is the hot-encoded categorical vector for the frame (with a value of 1 for the category it pertains and 0 for the remaining), P is the probability vector predicted by the network for the frame, NF is the number of frames of the sequence, and C is the number of classes (in our case 3 classes for background, ED and ES). w is the weight associated to each frame, where it has a value of 100 for ED and ES and 1 for the rest.

$$WCCE = -\sum_{i=1}^{NF} w_i \sum_{c=1}^{C} Y_{ic} \log\left(P_{ic}\right) \tag{1}$$

The second loss function used was the generalized Dice loss introduced by Sudre et al. (2017) which employs the Dice coefficient (Eq. 2), a parameter that measures the degree of overlap between two sets. The Dice coefficient is usually employed as a means of evaluating segmentation (Zou et al., 2004; Crum et al., 2006) and has also been employed as a loss function for segmentation problems with great results in the work of Milletari et al. (2016). The Dice loss has been speculated to be of use in this type of time series classification by Roald (2018) for its ability to target both sensibility and specificity. Furthermore, it has shown good results in problems like natural language processing (Li et al., 2019) which is a type of time-sequence problem. In this case the Dice is calculated between two vectors with three possible labels rather than two images. In the case of the Dice loss we applied a weight of 0.45 for the ES and ED categories, and 0.1 for the background category in order to force the network focus more in correctly detecting the ES and ED.

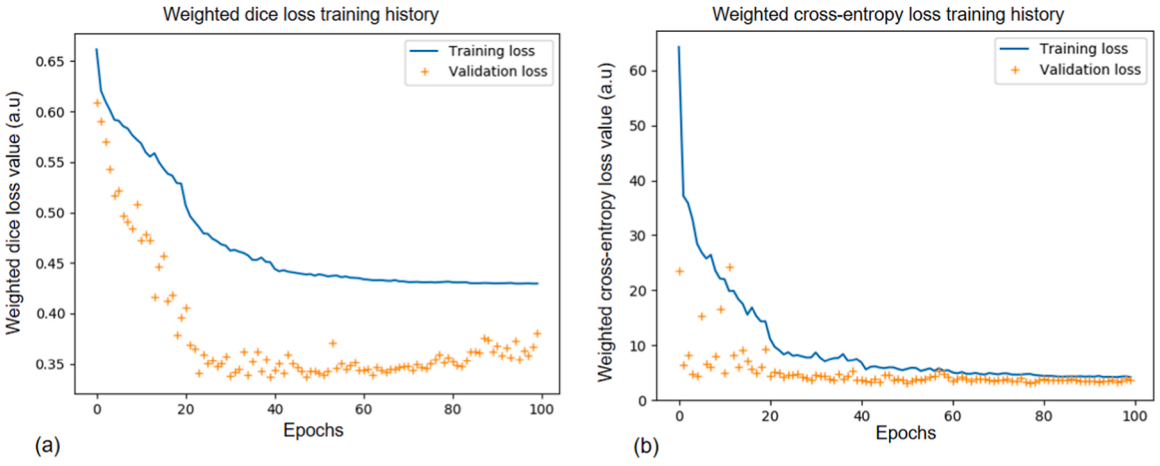

Fig. 3. Training and validation losses across epochs for the two tested loss functions. Both reached a plateau mid-training in the training loss. For the weighted cross entropy (b) the validation loss started with some fluctuation before stabilizing and finally reached a plateau without showing signs of overfitting. In contrast the weighted Dice validation loss (a) showed a more stable behavior throughout all the training, but started to get higher values around epoch 60, which indicate the beginning of a slight overfitting.

This loss function is presented in Eq. 2 where all variables have the same meaning as in Eq. 1 except w. In this case w represents the weight associated for each class, with fixed value of 0.45 for ED and ES and 0.1 for the background.

$$GDSC \quad loss = 1 - 2\sum_{c=1}^{C} w_c \frac{\sum_{i=1}^{NF} P_{ic} Y_{ic}}{\sum_{i=1}^{NF} P_{ic} + Y_{ic}} \tag{2}$$

2.7. Implementation details

The network was implemented in Python 3.7.6 using tensorflow 2.1 (www.tensorflow.org, Google Brain, Mountain View, CA) using its Keras API. The hardware used comprised a GPU RTX 2080 Ti with 11 Gb of RAM (Nvidia Corporation, Santa Clara, CA) CPU i9 9900 K (3.6 GHz) and 64 GB of RAM, running on Windows 10 operating system. The network was trained for 100 epochs using both the training dataset and the validation dataset. After some testing the training was set up using ADAM optimizer with a learning rate of $1e^{-5}$ and a batch size of 2. This learning rate was the one that showed best performance during training and the batch size was limited to 2 due to memory limitations.

3. Results

We present the results offered by the network trained with the two loss functions and using the two classification methods on the test set. Furthermore, we analyze the performance of the training by checking the validation and training loss across epochs.

3.1. Training performance

The training and validation loss records are depicted in Fig. 4. For the network trained with the weighted cross-entropy, the training loss decreased continually during all the epochs, starting to reach a plateau around epoch 60. In contrast, the validation loss was slightly erratic in the first epochs but it stabilized quickly around epoch 20. From this point the validation loss kept unchanged during the remaining epochs. It is important to note that the validation loss was lower than the training loss during all the training process (with some minor epoch points). The whole training process took 22 h to complete. The network we finally used was the one after the complete training, since we did not see any indications of overfitting at any point.

For the network trained with the weighted Dice loss, the training loss decreased continually during all the epochs, reaching a plateau around epoch 40. The validation loss followed a similar trend but with lower loss values. However, around epoch 60 a slight increase in the validation loss was detected that increased slowly across epochs, likely indicating a slight overfitting at that point. This training process took 22 h to complete, the same as with the cross-entropy loss setting. The network we finally used with this setting was the one obtained after epoch 60, since from that point the resulting networks seemed to lose quality.

3.2. Frame detection

We evaluated the quality of the network applying it to the test set comprising a total of 98 cases. The test set included 95 cases with 35 frames, 2 cases with 25 frames and 1 case with 22 frames. To evaluate the trained neural network, we employed the frame difference in the same way as Kong et al. (2016). Though in this previous work, the focus was on the average value, we now think it is important to know the distribution of the frame difference, so we provide the standard deviation as well. The frame difference for both ES and ED is the absolute difference in the position of the true frame to the one selected by the method (naïve or central method) after obtaining the probabilities by the neural network. We additionally computed the frame difference error normalized to the number of frames in the sequence, in order to have a better metric to compare against other methods which could employ sequences of different length.

The results for both methods applied to the two trained neural networks are presented in Table 2. The results clearly show that the central method and the Dice loss training perform better in all cases. The best results are those obtained by the network trained with the weighted Dice loss and using the central method. This scheme achieves a perfect result for ED, while for ES an average error of 1.24 frames (average relative error of 0.03) was obtained. We additionally observed that the great

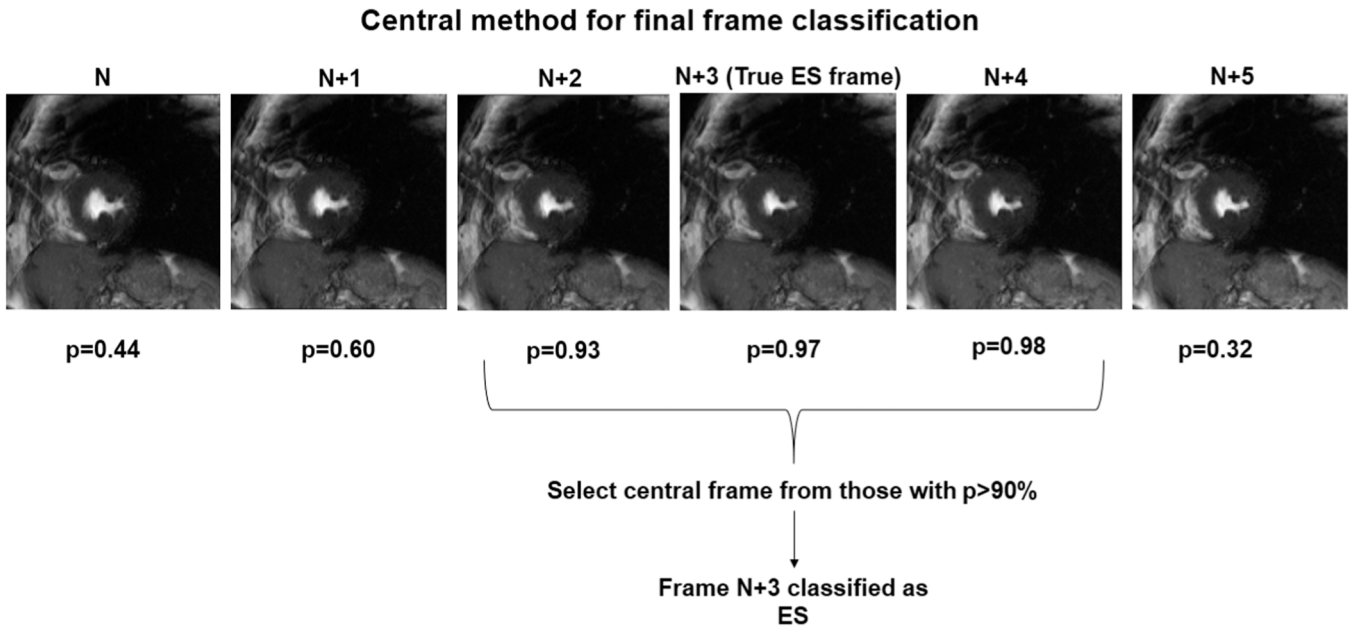

Fig. 4. Scheme of the central method for final ES and ED classification. The representation shows an example applied to the ES frame selection. The upper values of the images represent the frame number (being N an arbitrary location within the sequence), and the lower values represent the probabilities (p) of being ES given by the neural network. Only the frames with a probability higher than 90% are first selected, and then the final selection is done by taking the central frame among the selected ones even if its associated probability is lower than another.

Table 2
Frame difference error (number of frames) for the ES and ED detection with the different settings employed. The normalized values of the error are indicated below the absolute error values in parentheses. The best score is indicated in bold.

	ES (naïve method)	ES (central method)	ED (naïve method)	ED (central method)
Weighted cross-entropy	3.121 ± 3.500 (0.090 ± 0.101)	2.505 ± 2.249 (0.072 ± 0.065)	0.141 ± 0.619 (0.005 ± 0.018)	0.141 ± 0.619 (0.005 ± 0.018)
Weighted Dice	1.747 ± 1.849 (0.051 ± 0.053)	**1.242 ± 1.45 (0.036 ± 0.042)**	0 ± 0 (0 ± 0)	**0 ± 0 (0 ± 0)**

majority of cases had a frame difference error of either 0 or 1 (65% of the cases), with the remaining being between 2 and 4 with a decreasing number of cases the bigger the error. There were not any noticeable differences between the cases with 35 frames and the remaining ones (the cases with a lower number of frames had errors of 1 frame in 2 cases and one case had an error of 0 for the ES detection).

Finally, the time required for the network to process a single case using our hardware was on average 0.1 s, and the network provided the outputs for the entire test set in 10.2 s (employing a batch of 1 in order to average the time required for independent estimations by the network).

4. Discussion

We describe a methodology employing a fully convolutional neural network capable of classifying the frames within a short-axis stack of CMR cine sequences in order to detect ED and ES. The neural network is characterized by the use of dilated convolutions with different dilation rates in order to process temporal information.

Although the detection of the ES and ED frames is a prerequisite to measure several cardiac parameters, this has not been extensively addressed in CMR postprocessing compared to the ventricular segmentation. While it is true that manual segmentation is the more time-consuming task, the manual labeling to select the frames to apply the segmentation is still required. In this work the expert cardiologist that labeled the dataset needed around 2–3 min in easy cases, but the harder cases could take them around 5 min when agreement was not that clear between cardiologists. It is not far-fetched to think this time could also be slightly longer for less experienced cardiologists or radiologists. In a radiological setting when lots of patients require fast diagnosis, saving this time can further improve the workflow of the clinical experts.

When comparing our results to others we have that in the work of Kong et al. (2016) a convolutional neural network with a LSTM module was used to process temporal information. The authors obtained a frame difference average error of 0.38 and 0.44 for ED and ES respectively on a dataset comprising 420 sequences using 4-fold cross-validation. Noteworthy, the authors employed cine acquisitions with only one slice and a constant number of frames of 20, lower than our average number of frames of 35. Hsin and Danner (2016) reported a method for determining ES and ED with segmentation convolutional neural networks by selecting the frames with the lower and higher left ventricular segmentation area, no information regarding the error in this selection was reported. We note that this approach can be problematic as it would need a high quality segmentation in all frames, requiring a prior manual segmentation in all the frames prior to train the algorithm. Yang et al. (2017) used a neural network to segment the left ventricle in sequences of 1 slice with a higher duration. These sequences included a constant number of 84 frames. The authors used the relative location of the segmented left ventricle to determine ES and ED in 10 cases with 10 different sequences, each covering different regions of the heart which were used to validate their results. In this study an accuracy of 75% in the classification was obtained that increased to 95% when only mid region slices were analyzed.

Our approach is similar to the work of Kong et al. (2016), but instead of using recurrent layers like LSTM or GRU (gated recurrent unit) we used dilated convolutions to address the treatment of the temporal information. LSTM and GRU are types of layers that have been very successful at addressing time-related issues as described by Yu et al. (2019), however they are also difficult to train and are more unstable during training compared to convolutional layers (Pascanu et al., 2013; Hou et al., 2019). Additionally, the use of dilated convolutions has been shown to be very efficient in both quality and training performance, and has demonstrated to be capable of using longer-term information compared to recurrent layers (Oord et al., 2016).

The neural network was trained with two different loss functions: the classical weighted cross-entropy and the weighted Dice loss. The chosen weights for both loss functions were chosen based on our specific dataset, but we expect that with a dataset where the number of frames is dramatically different these weights would need to be adjusted in a different manner. Among the two loss functions, the Dice loss obtained the best results. The Dice loss tries to optimize the overlap between the real sequence and the predicted one, which means that the loss uses the information of the sequence as a whole, compared to the cross-entropy where each frame loss is computed independently and then averaged. This property and the results obtained in this work and in other works treating with sequential data as in Li et al. (2019) further indicate that using the Dice loss is more suited for this type of scenarios when training neural networks.

Due to memory limitations, the number of layers and parameters we used in the dilated convolutions was low (1 channel per path, with 9 different paths). Noteworthy, neural networks usually perform better with higher number of parameters. Memory limitations were caused by the large size of the inputs processed by the network (arrays of size 176 ×176 ×35 in most cases) which forced the design of the network to be of smaller size. Notwithstanding this limitation, the results offered by the network have shown to be highly accurate in the best training setting, with a perfect ED detection and a frame difference of 1.2 for ES in the test set. Additionally, even if there was a limited number of samples with a lower number of frames than the majority of 35, the test set included some cases with a significantly lower number of frames (22 and 25) and in these cases the frame difference for the ES was still either 0 or 1 and 0 for the ED. This shows that the network is capable of working with similar quality with different number of frames. Although the latter is true we also note that the number of samples with less than 35 frames is still small, and increasing the dataset with more of these cases could further prove these findings. The relative frame difference error employed helps to determine the true impact of the error in relation to the entire cardiac cycle. In this work the relative average difference error was of 0.03 for the ES which indicates that with respect to the entire cardiac cycle the error in frames corresponds to 3% of its length. The clinical impact of this error can be assumed to be low as adjacent frames in the sequences employed implied very small contraction differences.

An additional point of consideration is that our work focuses on finding the sequence volumes corresponding to ED and ES as a whole, but the different heart regions (basal, mid and apical regions) can slightly differ in their specific ED and ES as the motion is not perfectly simultaneous along these regions. The proposed neural network should be capable of working with lone heart slices or a limited number of slices covering a specific heart region. A future line of work is to test this methodology in such settings.

In conclusion, the results obtained are promising. The higher error obtained for ES detection indicate that this is more difficult to determine compared to ED. This was also true for the manual labelling done by the expert cardiologist, where some disagreements that required further discussion happened for some ES frames. By analyzing the images, it can

be seen how the left ventricle has a more regular shape (circular) compared to the ES. Additionally, the differences in contraction state are more apparent between adjacent frames close to the ED compared to the ES (see Fig. 1). These characteristics could explain why ED is harder to detect. These findings are consistent with the results reported by Kong et al. (2016). We hypothesize that employing a larger number of layers in the dilated convolution blocks could yield improved results. Additionally, newer architectures that work with temporal sequences have been developed in recent years. Specifically, the transformer architecture described in the work of Vaswani et al. (2017) has shown good results at generating text sequences as reported by Brown et al. (2020). Studying these novel types of architectures for the problem described in this work could provide a promising future line of work. Furthermore, considering the quality of the results obtained a full system including the first step for frame classification followed with another neural network to segment the regions of the heart (which is a well-studied and stablished problem with existing solutions) could be developed to make the whole analysis process automatic.

5. Conclusions

We have presented a fully convolutional neural network for the classification of the end-systolic and end-diastolic frames in short axis CMR cine sequences. The neural network employs dilated convolutions to encode the temporal features instead of the more widespread recurrent layers. This approach allows the network to require a lesser number of parameters and facilitates its training.

Our network has shown promising results that could allow its use in the clinical setting and save time in the diagnostic workflow. The detection of ED was perfect and the error for ES was very small when trained with the weighted Dice loss and applying the central method. Applying other settings like the weighted cross-entropy loss for training or selecting the frames with the highest probability of being ES and ED produced worse results.

We are aware of the limitations of our neural network, mainly the number of channels included in the dilated convolution operations was very limited (1 channel per path). This could be the reason why it has more difficulties in correctly classifying the end-systole. However, even in this situation, the classified frames are close to the right ones. This further proves that dilated convolution has great potential in time-sequence analysis. With higher hardware resources, a neural network with more depth in the dilated convolution paths could get even better results for classifying ES and ED employing the same type of architecture.

Funding sources

This work was partially supported by the Conselleria d'Innovació, Universitats, Ciència i Societat Digital, Generalitat Valenciana (grants AEST/2020/029 and AEST/2021/050).

CRediT authorship contribution statement

Manuel Pérez-Pelegrí: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Visualization. **José V. Monmeneu:** Conceptualization, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – review & editing, Visualization. **María P. López-Lereu:** Conceptualization, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – review & editing, Visualization. **Alicia M. Maceira:** Validation, Resources, Data curation, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Vicente Bodi:** Validation, Resources, Data curation, Writing – review & editing, Supervision, Project administration, Funding acquisition. **David Moratal:** Conceptualization, Methodology, Formal analysis, Investigation, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abdelmaguid, E., Huang, J., Kenchareddy, S., Singla, D., Wilke, L., Nguyen, M.H., Altintas, I., 2018. Left ventricle segmentation and volume estimation on cardiac MRI using deep learning. arXiv Comput. Vis. Pattern Recognit.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., Mccandlish, S., Radford, A., Sutskever, I., Openai, D.A., 2020. Language Models are Few-Shot Learners.

Chen, C., Bai, W., Davies, R.H., Bhuva, A.N., Manisty, C.H., Augusto, J.B., Moon, J.C., Aung, N., Lee, A.M., Sanghvi, M.M., Fung, K., Paiva, J.M., Petersen, S.E., Lukaschuk, E., Piechnik, S.K., Neubauer, S., Rueckert, D., 2020. Improving the generalizability of convolutional neural network-based segmentation on CMR images. Front. Cardiovasc. Med. 7, 105. https://doi.org/10.3389/fcvm.2020.00105.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40, 834–848. https://doi.org/10.1109/TPAMI.2017.2699184.

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. arXiv.

Childs, H., Ma, L., Ma, M., Clarke, J., Cocker, M., Green, J., Strohm, O., Friedrich, M.G., 2011. Comparison of long and short axis quantification of left ventricular volume parameters by cardiovascular magnetic resonance, with ex-vivo validation. J. Cardiovasc. Magn. Reson. 131 (13), 1–9. https://doi.org/10.1186/1532-429X-13-40.

Crum, W.R., Camara, O., Hill, D.L.G., 2006. Generalized overlap measures for evaluation and validation in medical image analysis. IEEE Trans. Med. Imaging 25, 1451–1461. https://doi.org/10.1109/TMI.2006.880587.

Dominguez, C.R., Kachenoura, N., Mulé, S., Tenenhaus, A., Delouche, A., Nardi, O., Gérard, O., Diebold, B., Herment, A., Frouin, F., 2005. Classification of segmental wall motion in echocardiography using quantified parametric images, in: Proceedings of the International Workshop on Functional Imaging and Modeling of the Heart. Springer Verlag, 477–486. ⟨https://doi.org/10.1007/11494621_47⟩.

Gifani, P., Behnam, H., Shalbaf, A., Sani, Z.A., 2010. Automatic detection of end-diastole and end-systole from echocardiography images using manifold learning. Physiol. Meas. 31, 1091–1103. https://doi.org/10.1088/0967-3334/31/9/002.

Hou, L., Zhu, J., Kwok, J.T., Gao, F., Qin, T., Liu, T.-Y., 2019. Normalization Helps Training of Quantized LSTM, in: Proceedings of the Neural Information Processing Systems (NeurIPS 2019). Vancouver, 7346–7356.

Hsin, C., Danner, C., 2016. Convolutional Neural Networks for Left Ventricle Volume Estimation.

Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift, in: Proceedings of the Thirty Second International Conference on Machine Learning, ICML 2015, International Machine Learning Society (IMLS), 448–456.

Kong, B., Zhan, Y., Shin, M., Denny, T., Zhang, S., 2016. Recognizing end-diastole and end-systole frames via deep temporal regression network, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016, Springer Verlag, 264–272. ⟨https://doi.org/10.1007/978-3-319-46726-9_31⟩.

Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., Li, J., 2019. Dice Loss for Data-imbalanced NLP Tasks, 465–476.

Lopez, E.O., Ballard, B.D., Jan, A., 2021. Cardiovascular Disease, StatPearls Publishing.

Meidellfiorito, A., Ostvik, A., Smistad, E., Leclerc, S., Bernard, O., Lovstakken, L., 2018. Detection of Cardiac Events in Echocardiography Using 3D Convolutional Recurrent Neural Networks. In: Proceedings of the IEEE International Ultrasonics Symposium, IUS, IEEE Computer Society, 1–4. ⟨https://doi.org/10.1109/ULTSYM.2018.8580137⟩.

Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the Fourth International Conference on 3D Vision, 3DV 2016, Institute of Electrical and Electronics Engineers Inc., 565–571. ⟨https://doi.org/10.1109/3DV.2016.79⟩.

Oord, A., van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K., 2016. WaveNet: a generative model for raw. Audio.

Pascanu, R., Mikolov, T., Bengio, Y., 2013. On the difficulty of training recurrent neural networks. In: Proceedings of the International Conference on Machine Learning, JMLR.org, Atlanta, 1310–1318.

Pérez-Pelegrí, M., Monmeneu, J.V., López-Lereu, M.P., Pérez-Pelegrí, L., Maceira, A.M., Bodí, V., Moratal, D., 2021. Automatic left ventricle volume calculation with explainability through a deep learning weak-supervision methodology. Comput. Methods Prog. Biomed. https://doi.org/10.1016/J.CMPB.2021.106275.

Perez-Pelegri, M., Monmeneu, J.V., Lopez-Lereu, M.P., Ruiz-Espana, S., Del-Canto, I., Bodi, V., Moratal, D., 2020. PSPU-Net for Automatic Short Axis Cine MRI Segmentation of Left and Right Ventricles, in: Proceedings of the IEEE Twentieth International Conference on Bioinformatics and Bioengineering (BIBE). Institute of Electrical and Electronics Engineers (IEEE), 1048–1053. ⟨https://doi.org/10.1109/bibe50027.2020.00177⟩.

Poudel, R.P.K., Lamata, P., Montana, G., 2017. Recurrent fully convolutional neural networks for multi-slice MRI cardiac segmentation. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 83–94. ⟨https://doi.org/10.1007/978–3-319–52280-7_8⟩.

Roald, M., 2018. Detecting Valvular Event Times from Echocardiograms Using Deep Neural Networks, University of Oslo.

Santurkar, S., Tsipras, D., Ilyas, A., Mit, A.M., A., 2018. How Does Batch Normalization Help Optimization? In: Proceedings of the Thirty Second International Conference on Neural Information Processing Systems, 2488–2498.

Shalbaf, A., Behnam, H., Gifani, P., Alizadeh-Sani, Z., 2011. Automatic detection of end systole and end diastole within a sequence of 2-D echocardiographic images using modified Isomap algorithm. In: Proceedings of the First Middle East Conference on Biomedical Engineering, MECBME, 2011, IEEE Computer Society, 217–220. ⟨https://doi.org/10.1109/MECBME.2011.5752104⟩.

Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M., 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 240–248. ⟨https://doi.org/10.1007/978–3-319-67558-9_28⟩.

Tao, Q., Yan, W., Wang, Y., Paiman, E.H.M., Shamonin, D.P., Garg, P., Plein, S., Huang, L., Xia, L., Sramko, M., Tintera, J., de Roos, A., Lamb, H.J., van der Geest, R. J., 2019. Deep learning–based method for fully automatic quantification of left ventricle function from cine MR images: a multivendor, multicenter study. Radiology 290, 81–88. https://doi.org/10.1148/radiol.2018180513.

Tong, Q., Li, C., Si, W., Liao, X., Tong, Y., Yuan, Z., Heng, P.A., 2019. RIANet: recurrent interleaved attention network for cardiac MRI segmentation. Comput. Biol. Med. 109, 290–302. https://doi.org/10.1016/j.compbiomed.2019.04.042.

Townsend, N., Wilson, L., Bhatnagar, P., Wickramasinghe, K., Rayner, M., Nichols, M., 2016. Cardiovascular disease in Europe: epidemiological update 2016. Eur. Heart J. https://doi.org/10.1093/eurheartj/ehw334.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: Proceedings of the Thirty First International Conference on Neural Information Processing Systems (NIPS'17), Neural Information Processing Systems Foundation, Long Beach, 6000–6010.

Yang, F., He, Y., Hussain, M., Xie, H., Lei, P., 2017. Convolutional neural network for the detection of end-diastole and end-systole frames in free-breathing cardiac magnetic resonance imaging. Comput. Math. Methods Med. 2017. https://doi.org/10.1155/2017/1640835.

Yu, F., Koltun, V., 2015. Multi-scale context aggregation by dilated convolutions. In: Proceedings of the Fourth Int. Conf. Learn. Represent, ICLR 2016 - Conf. Track Proc.

Yu, Y., Si, X., Hu, C., Zhang, J., 2019. A review of recurrent neural networks: LSTM cells and network architectures. Neural Comput. 31, 1235–1270. https://doi.org/10.1162/NECO_A_01199.

Zolgharni, M., Negoita, M., Dhutia, N.M., Mielewczik, M., Manoharan, K., Sohaib, S.M. A., Finegold, J.A., Sacchi, S., Cole, G.D., Francis, D.P., 2017. Automatic detection of end-diastolic and end-systolic frames in 2D echocardiography. Echocardiography 34, 956–967. https://doi.org/10.1111/echo.13587.

Zou, K.H., Warfield, S.K., Bharatha, A., Tempany, C.M.C., Kaus, M.R., Haker, S.J., Wells, W.M., Jolesz, F.A., Kikinis, R., 2004. Statistical validation of image segmentation quality based on a spatial overlap index. Acad. Radiol. 11, 178–189. https://doi.org/10.1016/S1076-6332(03)00671-8.