Exploring a Segmentation-Classification Deep Learning-based Heart Murmurs Detector

Daniel Enériz[1], Antonio J Rodríguez-Almeida[2], Himar Fabelo[2,3], Samuel Ortega[4], Francisco Balea-Fernandez[2,5], Nicolás J Medrano[1], Belén Calvo[1], Gustavo M Callicó[2]

[1]Aragon Institute of Engineering Research (I3A), University of Zaragoza, Zaragoza, Spain
[2]Institute for Applied Microelectronics (IUMA), Univ. of Las Palmas de G. C., Las Palmas de G. C., Spain
[3]Fundación Canaria Instituto de Investigación Sanitaria de Canarias, Las Palmas de G. C., Spain
[4]Norwegian Institute of Food Fisheries and Aquaculture Research (Nofima), Tromsø, Norway
[5]Dept. of Psychology, Sociology and Social Work, ULPGC, Las Palmas de Gran Canaria, Spain

Abstract

This work presents the advances of the UZ-ULPGC team in the Heart Murmur Detection from Phonocardiogram Recordings: The George B. Moody PhysioNet Challenge 2022. As the 2016 challenge proved the success of the combination of a segmentation algorithm and a classifier, a deep learning-based murmur detector is developed using the sequence segmentation-classification. A U-Net-based segmentation model is used to extract each cardiac cycle from the PCG with state-of-the-art accuracy. Three deep models are tested for the classification: a model based on four independent 1D-convolutional feature extractors; its variation enabling combination of the features; and an autoencoder. Furthermore, to enable unique patient diagnostic, a decision model gathering all the patient-related cardiac cycles information is added. All classifiers show limited performance, probably due to the heavy class imbalance of the data at the cardiac cycle level and the minimal preprocessing chosen in the architecture. Note that our models have not been tested in the hidden challenge data and therefore we are not ranked. Hence, a 10-fold cross-validation over the training set is used to evaluate their performance, with the best model getting a weighted accuracy score in the presence task of 0.58±0.10 and 10735±2208 in Challenge cost score for the outcome.

1. Introduction

In 2019, 17.9 million people died due to cardiovascular diseases (CVDs), being the leading cause of death globally. Moreover, the cardiac auscultation, which is the fundamental method for first screening CVDs, is difficult to learn. These two factors have motivated the development of automatic Phonocardiogram (PCG) analysis, since a computer-aided decision system based on auscultation would lead to accessible and accurate screening, with shorter diagnostic times, facilitating the referral of patients to cardiology doctors. The 2022 George B. Moody PhysioNet Challenge [1][2] addresses this issue, pursuing the development of an open-source algorithm that performs two patient diagnostic tasks: heart murmur detection and clinical outcome identification; both using all the information available in multiple PCGs from several auscultation locations. Inspired by the success in the 2016 Physionet/CinC challenge [3] of PCG segmentation followed by cardiac cycle classification algorithms, we explore the performance of a deep learning-based segmentation-classification architecture followed by a global rule algorithm that combines all the patient-related data at the cardiac cycle level to give a single diagnostic. Three classifier candidates are tested: the deep-learning part of the best scoring entry in the 2016 Challenge [4], a variation of it enabling mixture between input features, and an autoencoder; all of them showing limited performance in both classification tasks. The challenge database is part of the CirCor Digiscope dataset [5]. The publicly available data is composed by the PCG recordings from 942 patients with the annotation of their heart states, the demographic data and the murmur presence and clinical outcome labels. Furthermore, part of the 2016 PhysioNet/CinC database has been used in this work to pretrain the segmentation algorithm.

2. Method

In Figure 1 the proposed architecture is shown, which is based on the combination of a segmentation model, a classifier and global rule algorithm, all of them with their preprocessing step. A detailed description of all the stages of the architecture is included in this section.

2.1. Segmentation

This stage is based on the work of Renna *et al.* 2019 [6], where a 1D U-Net model was presented, setting the current state-of-the-art accuracy in PCG segmentation. Firstly, each heart sound is band-pass filtered between 25 and 400 Hz. Then, the spike removal method described in [7] is applied. The next step is the generation of four different envelopes/envelograms: Hilbert envelope, Homomorphic envelogram, Power Spectral Density (PSD) envelope, and Wavelet envelope. Finally, the envelograms are downsampled to 50 Hz and normalized to have zero mean and unit variance. For each recording, the normalized envelograms are grouped in a 4-dimensional signal. Besides, each time instant has an associated state label (S1, systole, S2, diastole). Patches of fixed length N=64 are extracted from with a specific stride τ=8. These portions of the signal are the input to the segmentation algorithm.

The model has two stages: an encoder with four blocks and a decoder with another four blocks. In each encoding block the signal is compacted in time dimension while the number of channels is increased. This is done with two consecutive *Conv1D* layers with ReLU activation whose number of filters is the double of the block input channels. Then a *MaxPooling* layer is applied to halve the time dimension. The decoder blocks expand back the information in the time dimension while the number of channels is reduced. Additionally, there are skip connections between each pair of encoder-decoder blocks to allow direct transfer of information. Hence, a decoder block is based in one *UpSampling* layer that doubles the time dimension of the input, a *Conv1D* layer fed by the concatenation of the output of the *UpSampling* layer and the skip connection, and another *Conv1D* layer. Both *Conv1D* have the number of filter equal to the halve of the number of channels of the input, and ReLU activation. At the output of the model, the probabilities of being in each fundamental heart state in each time instant are given. Sequential max temporal modelling is used at this point to admit only allowable transitions between heart states, since it has a good balance between complexity and performance, as shown in [6]. Thus, the segmentation model is firstly trained in the 2016 challenge data with the procedure shown in [6], except for the epochs, which are fixed to 5. Then, it is trained in the 2022 challenge public data, with the same method.

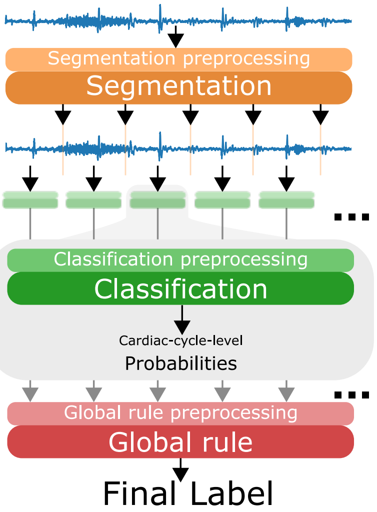

Figure 1. General scheme of the architecture. First, the arbitrary-lengthen PCG is divided into cardiac cycles using a segmentation model. Then, each cardiac cycle is passed through a classification model, specifically trained for the target task (outcome classification or murmur detection) outputting a label in the cardiac cycle level. Finally, all the patient-related labels are combined using a global rule to give a single diagnostic label.

2.2. Classification

In the classification step, the target is to individually classify in the selected task each cardiac cycle extracted from the segmentation. In this stage the preprocessing is the same as the used in [4]: a downsampling to 1 kHz, a 2nd-order Butterworth bandpass filter between 25 and 400 Hz, and the spike removal method from [7]. Then, the segmentation described above is applied to extract 2.5 s cardiac cycles. If the cycle has shorter duration is zero padded.

As commented in the Introduction, different models are tested. First, the C. Potes *et al.* convolutional model [4] is replicated. This model is based on four independent 1D-convolutional feature extractors, which are fed by four exclusive band-pass filtered signals of the PCG (with ranges of [25-45], [45-80], [80-200], and [200-400] in Hz). Each 1D-convolutional extractor is based on two combinations of a *Conv1D* layer, the ReLU activation function and a *MaxPooling* to halve the time dimension. Both convolutional layers have a kernel size of 5, stride 1, the padding method is set to keep the same dimensionality and there is no bias, but the first one has 8 filters and the last one 4. After the two *Conv1D*+ReLU+*MaxPooling* combinations, all the features are concatenated and flattened. Then a multilayer perceptron (MLP) with a hidden layer of 20 neurons is placed. The output layer has the same number of neurons as the number of classes in the selected task (3 for murmur presence and 2 for outcome), and SoftMax activation is used. A dropout of 25 % is applied after the flattened layer, and another dropout of 50 % is placed after the hidden layer of the MLP, where a 10^{-2} L2 activity regularization is also employed. The Adam optimizer with the Cross Entropy loss function is used to train the model during 100 epochs, with a batch size of 1024 and a learning rate of $7 \cdot 10^{-4}$. The weights obtained after the epoch with minimum loss function value in the validation set are saved. The hyperparameters employed here are the same as in [4], except for the epochs that are halved since the minimal validation loss is always in the first 100 epochs.

Then, as a second classifier, the C. Potes *et al.* [4] model is modified by joining the 4 independent feature extractors into a single CNN. This makes the model 4 times smaller approximately in terms of both feature maps and parameters. The same training procedure as the described in the previous classification model is applied for this one.

Finally, an autoencoder-based model is also tested. Inspired by the performance of the Renna *et al.* segmentation model [6], a variation of it is tested to be used as an autoencoder. The number of filters in each encoding/decoding block is changed to compress/decompress the information in both the time and the channels dimension. Also, skip connections between the encoding and the decoding parts are removed to avoid direct signal transfer. Table 1 contains all the details of the architecture. It is trained to replicate the input signal, and only the normal samples (absence of murmur in the detection task, normal diagnostic for the outcome) are used in the training, with the purpose that the anomalies present in the abnormal samples could not be replicated. Then, an error function, as the Mean Absolute Error (MAE), computed between the model input and its output can be used to set a threshold to identify the normal classes vs. the abnormal ones. The Adam optimizer with the Mean Squared Error (MSE) loss function is used to train the autoencoder during 200 epochs with batch size of 64 and learning rate of 10^{-3}. As in the previous models, the weights corresponding to the minimum validation loss are saved. The selection of these hyperparameters is motivated by the necessity of a larger training process, due to the higher complexity of the signal replication problem compared to the segmentation task.

2.3. Global rule

For the final diagnostic algorithm, a MLP is in charge of analyze statistical metrics extracted from the arbitrary-lengthen set of classifier outputs per patient in the cardiac-cycle level. These outputs are the probabilities of being in each class, thus having a total shape per patient of (n_{cycles}, n_{classes}). The chosen statistic metrics are three per class: the mean and the standard deviation of the probabilities of the class and the number of cycles whose class probability is maximal divided by the total number of cycles, n_{cycles}. Thus, the MLP has an input layer of $3n_{\text{classes}}$, that is followed by a hidden layer with a fixed number of neurons of 5 with ReLU activation. Finally, the output layer has the number of neurons equal to n_{classes} and SoftMax activation. It is trained with the Adam optimizer using the Cross Entropy loss function during 10 epochs with a batch size of 16 and a learning rate of 10^{-3}.

Block/*Layer*	**# filters**	**# params.**	**Output shape**
Input	-	-	(2496, 1)
Encoder_0	64	12480	(1248, 64)
Encoder_1	32	9216	(624, 32)
Encoder_2	16	2304	(312, 16)
Encoder_3	8	576	(156, 8)
Conv1D	4	96	(156, 4)
Conv1D	4	48	(156, 4)
Decoder_0	8	288	(312, 8)
Decoder_1	16	1152	(624, 16)
Decoder_2	32	4608	(1248, 32)
Decoder_3	64	18432	(2496, 64)
Conv1D	1	192	(2496, 1)

Table 1. Autoencoder architecture. The rest of the details are the same as the segmentation architecture. Note that the length of the input signal has been shortened to 2.496 s to enable the iterative halve of this dimension.

3. Results

In this section the performance results of each architecture stage are presented. The segmentation model is evaluated in the 2022 challenge public data with 10-fold cross validation (CV), and its results are similar to the ones presented in [6] for the N=64, τ=8 model: 90.2±0.9 % in accuracy, 96.0±1.1 % in sensitivity, and 96.0±1.0 % in positive predictive value.

Same validation procedure is applied for the CNN-based classifiers, whose results are available in Table 2. Figure 2 contains the histogram of the MAE distribution for the presence and absence classes. Note that there is a high overlapping between these distributions, making impossible to set a threshold to differentiate the classes. Same phenomena appear for the outcome. Thus, autoencoder performance metrics could not be obtained.

The results of the final global rule, and therefore of the entire model, on the public data using 10-fold CV are available in Tables 2, 3 and 4. Unfortunately, no model could be evaluated with the hidden data.

4. Discussion and Conclusions

The different classification architectures assessed in this work achieved modest performance. Since the segmentation algorithm replicated the state-of-art results [6], this fact might be related with the need of a more complex architecture in the classification stage of our approach. Neither the CNN-based classifiers nor the autoencoder have been capable of extract features that clearly identify a cardiac cycle with a murmur or a patient that presents an abnormal clinical outcome. This is probably due to the complexity both tasks involve, where respiration noises, frictions and another external noise present in a realistic scenario affect the heart sound

		Classification				Global					
	Model	Acc.	AUROC	Recall	Specifity	AUROC	AUPRC	F-measure	Acc.	W. Acc.	Cost
Pres.	Original	0.58±0.02	0.81±0.02	0.46±0.03	0.89±0.01	0.89±0.05	0.79±0.09	0.41±0.12	0.79±0.07	0.54±0.14	22042±4082
	Variation	0.56±0.02	0.80±0.02	0.41±0.03	0.91±0.01	0.79±0.09	0.65±0.09	0.45±0.09	0.80±0.06	0.58±0.10	21446±3777
Outc	Original	0.60±0.03	0.64±0.04	0.60±0.03	0.60±0.03	0.69±0.03	0.67±0.03	0.78±0.07	0.79±0.06	0.74±0.09	10735±2208
	Variation	0.60±0.03	0.64±0.04	0.60±0.03	0.60±0.03	0.63±0.03	0.61±0.03	0.70±0.05	0.70±0.05	0.64±0.06	12718±2176

Table 2. 10-fold CV results over the public data of the CNN classifiers for both tasks, presence (up) and outcome (down).

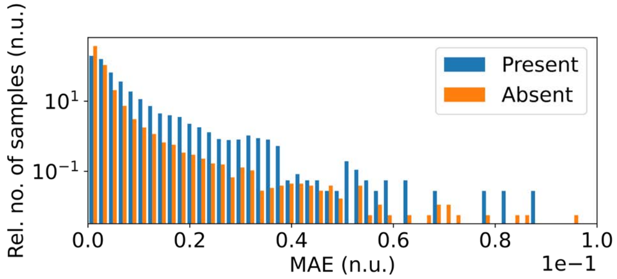

Figure 2. Distribution of the Mean Absolute Error computed between the input and the output of the autoencoder for the murmur detection classes.

recording, as the CirCor dataset is. For this reason, we believe that is necessary a more sophisticated preprocessing, extending the input up to entire murmur frequency range (0-600 Hz) and further reducing the impact of these noises' sources. Furthermore, the usage of more than one cardiac-cycle to feed the classifier would probably increase its performance, since the low-frequency noise would have lower impact. Additionally, the heavily imbalance of classes that appears at the cardiac cycle level in the murmur detection problem hardens the correct class separation, being the justification of the better performance at the outcome.

Training	Validation	Test	Ranking
0.56±0.02	-	-	-

Table 3. Best weighted accuracy metric score (official Challenge score) for the murmur detection task. Note that the architecture is not evaluated over the validation set nor the test set, and therefore our entry is not ranked. We used 10-fold cross validation over the training set.

Training	Validation	Test	Ranking
10735±2208	-	-	-

Table 4. Best cost metric scores (official Challenge score) for the murmur detection task. Note that the architecture is not evaluated over the validation set nor the test set, and therefore our entry is not ranked. We used 10-fold cross validation over the training set.

Acknowledgments

This work has been supported by the "Ministerio de Ciencia e Innovación", through the PID2019-106570RB-I00, the "Gobierno de Aragón" PhD grant BOA20201210014 and the "Universidad de Zaragoza", "Fundación Bancaria Ibercaja" and "Fundación CAI" research stray grant IT 9/21. Also, this work was completed while Antonio Rodríguez was beneficiary of a pre-doctoral fellowship by the "Agencia Canaria de Investigación, Innovación y Sociedad de la Información (ACIISI)" of the "Consejería de Economía, Conocimiento y Empleo" of the "Gobierno de Canarias", which is part-financed by the European Social Fund (FSE) (POC 2014-2020, Eje 3 Tema Prioritario 74 (85%)).

References

[1] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 2000;101(23):e215–e220.

[2] Reyna MA, Kiarashi Y, Elola A, Oliveira J, Renna F, Gu A, et al. Heart murmur detection from phonocardiogram recordings: The George B. Moody PhysioNet Challenge 2022. medRxiv 2022; URL https://doi.org/10.1101/2022.08.11.22278688.

[3] Clifford GD, Liu C, Moody B, Springer D, Silva I, Li Q, Mark RG. Classification of normal/abnormal heart sound recordings: The PhysioNet/Computing in Cardiology Challenge 2016. In 2016 Computing in Cardiology. 2016 Sep 11 (pp. 609-612). IEEE.

[4] Potes C, Parvaneh S, Rahman A, Conroy B. Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds. In 2016 Computing in Cardiology. 2016 Sep 11 (pp. 621-624). IEEE.

[5] Oliveira J, Renna F, Costa PD, Nogueira M, Oliveira C, Ferreira C, et al. The CirCor DigiScope dataset: from murmur detection to murmur classification. IEEE Journal of Biomedical and Health Informatics 2021;26(6):2524–2535.

[6] Renna F, Oliveira J, Coimbra MT. Deep convolutional neural networks for heart sound segmentation. IEEE Journal of Biomedical and Health Informatics. 2019 Jan 21;23(6):2435-45.

[7] Schmidt SE, Holst-Hansen C, Graff C, Toft E, Struijk JJ. Segmentation of heart sound recordings by a duration-dependent hidden Markov model. Physiological measurement. 2010 Mar 5;31(4):513.

Address for correspondence:

Daniel Enériz Orta
Pedro Cerbuna, 12. Facultad de Ciencias. 50009. Zaragoza. Spain
eneriz@unizar.es