2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Do we really need a segmentation step in heart sound classification algorithms?

Jorge Oliveira, Diogo Nogueira, Francesco Renna *IEEE Senior Member*,
Carlos Ferreira, Alípio M. Jorge and Miguel Coimbra *IEEE Senior Member*.

***Abstract*— Cardiac auscultation is the key screening procedure to detect and identify cardiovascular diseases (CVDs). One of many steps to automatically detect CVDs using auscultation, concerns the detection and delimitation of the heart sound boundaries, a process known as segmentation. Whether to include or not a segmentation step in the signal classification pipeline is nowadays a topic of discussion. Up to our knowledge, the outcome of a segmentation algorithm has been used almost exclusively to align the different signal segments according to the heartbeat. In this paper, the need for a heartbeat alignment step is tested and evaluated over different machine learning algorithms, including deep learning solutions. From the different classifiers tested, Gate Recurrent Unit (GRU) Network and Convolutional Neural Network (CNN) algorithms are shown to be the most robust. Namely, these algorithms can detect the presence of heart murmurs even without a heartbeat alignment step. Furthermore, Support Vector Machine (SVM) and Random Forest (RF) algorithms require an explicit segmentation step to effectively detect heart sounds and murmurs, the overall performance is expected drop approximately 5% on both cases.**

I. Introduction

Cardiovascular diseases (CVDs) are the leading cause of death in developed and developing countries and one of the major causes of hospitalization. By 2030, almost 23.6 million people will die from CVDs, according to the world health organization [1]. One possible solution, is a cost-effective screening of the population, not only to identify risk groups but also to follow up those who need emergency care. In this sense, heart sound auscultation represents a key exam, due to its simplicity and low cost. Although collecting heart sounds represents a relatively straightforward task, their interpretation is a challenging task for human listeners, since heart sounds are faint, their frequency content is located at the lower end of the audible spectrum. For these reasons, the design of an autonomous heart sound system can play an important role in boosting the accuracy and the pervasiveness of screening for CVDs. Standard approaches to extract useful information from heart sound recordings include an initial signal pre-processing step, which aims at reducing the impact of noise. After that, the following two phases are usually implemented: i) signal segmentation, and ii) signal classification. In particular, heart sound segmentation consists in the detection of the position and the boundaries of the fundamental heart sounds, S1 and S2. Heart sound segmentation plays a fundamental role, as it allows for the extraction of targeted features corresponding to the different segments of the signal but it also allows the detection of extra sounds (S3 and S4), murmurs, clicks, etc. Recently, deep learning solutions had become the state-of-art of heart sound classification. A particularly successful approach is CNN, which could be applied either to the segmented signal in the time domain, or to its time-frequency transform. In particular, [2] leverages the segmentation of the PCG and divides each heart sound recording into segments of 3 seconds long, all starting in correspondence of an S1 sound. From such segments, 2-dimensional heat maps containing the corresponding MFCCs are computed and used as input of a deep CNN. The intrinsic sequential nature and periodic behavior of heart sounds has also suggested the use of deep learning models able to keep track of the time evolution of time series. In particular, Latif *et al.* [3] studied and compared the performance of several RNNs in classifying heart sound signals. The methods proposed in [3] leverage the segmentation step by dividing PCG signals into segments of 2, 5, and 8 complete heartbeats. Then, for each segment, MFCCs are extracted and fed to the different RNN classifiers. Other recent approaches using deep learning algorithms skip completely the segmentation stage and apply directly classifiers to parts of the PCG signal. In particular, in [4], segments of duration 5 seconds are extracted sequentially from the PCG signals, with a constant stride of 1 second. Then, such segments are used to compute spectrograms and MFCCs, which are further used as input of a CNN classifier. Also RNNs have been recently embedded into end-to-end approaches, i.e., without recurring to signal segmentation step first. For example, Thomae *et al.* [5] proposed an end-to-end deep neural network combining 1-dimensional convolutional layers and gated recurrent unit (GRU) layers, where the input signal is entirely fed into the network.

A. Motivation and Contributions

In 2016, the PhysioNet community has organized a challenge with focus on the development of heart sound classification algorithms able to discriminate between normal and abnormal sounds. One of the interesting results and

J. Oliveira is with Universidade Portucalense Infante D. Henrique, Rua Dr. António Bernardino de Almeida, 541 4200-072 Porto, Portugal (email: jorgefisicomat@gmail.com). F. Renna is with the Instituto de Telecomunicações, Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre 1021/1055, 4169-007 Porto, Portugal (email: frarenna@dcc.fc.up.pt. M. Nogueira, C. Ferreira, A. Jorge and M. Coimbra are with INESC TEC, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal. A. Jorge and M. Coimbra are also with Faculdade de Ciências da Universidade do Porto (email: diogo.m.nogueira@inesctec.pt, amjorge@fc.up.pt, cgf@isep.ipp.pt, mcoimbra@dcc.fc.up.pt). This work is financed by national funds through FCT - Fundação para a Ciência e Tecnologia, within the scope of the project DSAIPA / AI / 0083/2020, and by national funds, through Fundação para a Ciência e Tecnologia (FCT); and UID/EEA/50008/2019.

conclusions drawn by the organizers of such challenge was that improved segmentation algorithms should be expected to be the best point of entry to obtain more significant improvements in abnormal heart sound detection tasks [6]. Motivated by this remark and by the different approaches recently appeared in the literature for abnormal heart sound detection, some of which are skipping a segmentation step, this work aims at assessing and quantifying the impact of a heart sound segmentation step on the performance of different families of heart sound classification algorithms. Note that the focus of this work is on the evaluation of classifiers that simply determine if a heart sound recording is normal or abnormal, since a more refined classification of the particular heart condition affecting a given patient is expected to rely on a segmentation step, for example, in order to determine the exact source of the murmur. The rest of the paper is organized as follow s. In Section II, the experimental methodology is described. In Section III, configuration and setup are presented for each model. In Section IV results are showed. Finally, conclusions are withdrawn in Section V.

II. Methodology

A. Materials

In this work, the database from the 2016 PhysioNet/ Computing in Cardiology Challenge [7] is used. The database provides a large collection of heart sound recordings, divided into eight different training sets. The Physionet/Cinc Challenge provided 3153 heart sounds. From this, 363 heart sounds were discarded by the following reasons: 87 records from Folder E are not heart sounds; 276 records do not have annotations or they are not properly annotated, therefore not considered in this study. Although signals were collected at different sampling frequencies (800Hz, 1000Hz, 2000 Hz, 3000Hz, 4000 Hz, 8000 Hz or 22050 Hz) no details concerning aliasing and imaging effects generated are provided.

B. Training and Testing Data

Aiming to get statistically significant results, a large dataset is created. In order to do so, the different constituent datasets of the 2016 Physionet/Computers in Cardiology Challenge (discussed in Section II-A) are merged. A total of 764 subjects and 3153 recordings are at disposal in this study. In order to increase the size of our dataset each heart sound is split into continuous segments of three seconds, through the provided annotations. Moreover, each segment starts at S1 state sample in order to ensure alignment during the classification process. As a result, our final dataset is composed by 17089 heart sound segments. When no explicit segmentation step is considered, heart sound segments are free to start in any state sample. In this case, our final dataset is composed by 19565 heart sound segments. During the process, 70% of healthy and unhealthy patients (and their corresponding audio records) are randomly used for training and the remaining ones are used for testing purposes. During the testing phase, segments from the same patient are analysed individually by one of the tested classifiers. As a result, a percentage of segments are classified as abnormal and the remaining segments as normal. In order to outcome a detection decision, i.e. the presence or not of murmur waves in the recording, a decision is made by setting an upper-bound limit over the distribution, i.e. if at least *m%* of the segments are classified as abnormal then the entire audio signal is classified as an abnormal heart sound signal, otherwise it is classified as a normal heart sound signal. Finally, in order to extract statistically significant results, the aforementioned procedure is repeated ten times. Note, that the generated training and testing sets at each run have different ratios of healthy and unhealthy records.

C. Pre-processing and Feature Extraction

The majority of the frequency components of heart sounds and murmurs are between 50 and 500Hz and higher frequencies are of little clinical significance, hence a band-pass Butterworth filter with cut-off frequencies at $50-850$ Hz was used. Furthermore, heart sound signals are downsampled from 2000Hz to 1000Hz and then normalized between -1 and +1. From the pre-processed PCG signal, MFCCs are extracted, as they represent the preferred features used by the majority of heart sound classification algorithms present in the literature (see, for example, [8] and [2]). During the windowing stage, overlapping sliding windows of 25 ms ran over segments of three seconds with strides of 10 ms. A total of 13 MFCC filterbanks per sliding window are computed, i.e, 300 MFCCs for each input signal of three seconds are extracted.

III. Models

A. GRU Model

Given a sequence of input vectors X = $(x_1,...,x_T)$ of length T, a standard GRU network processes sequentially each input vector x and generates a sequence of hidden state vectors $H=(h_1,...,h_T)$. Afterwards, h_T is feed into a fully connected (FC) layer, where a soft-max function is used in order to compute the output probability distribution.

1) Configuration and Setup: Regarding the GRU net design, the weight matrices are set to $\Re^{16\times13}$. FC weight's are set to $\Re^{2\times16}$. The bias vectors and the state memory vector are set to $\Re^{16\times1}$, respectively. The weight matrices of GRU, FC layers and the bias vectors are initialized using a uniform random distribution. Furthermore, h_0 is the zero vector, i.e., every component is set to zero.

2) Objective Function and Optimizer: As for the loss function, the binary cross entropy is used and minimized using the Adam optimizer algorithm. The step size is set to 0.01, the exponential decay rate for the first and second moment are set to 0.9 and 0.999, respectively. In order to avoid any division by zero during the computation of the cost function, a small scalar perturbation (10^{-6}) is added in the computation.

3) Training, Cross-validation and Model Selection: In the training phase, heart sound segments are shuffled randomly according to a uniform random distribution at the beginning of each epoch, and each segment is processed individually and sequentially (batches of size one). A cross-validation

dataset composed by 5% of balanced data is used. The training phase lasts for 25 epochs, and only the model that achieves the highest score (an average of sensitivity and specificity performances) in the cross-validation dataset is saved and further used during the testing phase.

B. CNN Model

1) Configuration and Setup: The CNN architecture was empirically selected. Several network sizes were tested, ranging from millions of weight parameters up to thousands of weight parameters, but without any significant score difference. Therefore, the simplest CNNs architecture was selected, consisting of 39426 parameters. In order to classify each heart sound segment, the processed MFCCs image of size 300×13 is fed into the network. In the first layer, 32 feature maps are generated. These are further convolved by another set of 32 independent filters. Since no zero padding techniques are being used, the image shrinks along the network, leading to feature maps of size 298×11, in the output of the second layer. Before feeding the output of the second layer into the third layer, a two-dimensional pooling operation is applied, in order to compress the most relevant features on the image. As a result, feature maps of size 59×5 are outputted at the third layer. Finally another set of 32 filters is applied leading to an output feature map of size 57×3. The outputs of all convolutional layers are followed by a ReLU activation function, which operates entry-wise. Furthermore, the last output feature map is compressed again by applying a two-dimensional pooling operation of (5×2), leading to tensor of shape $(11 \times 1 \times 32)$. The tensor is then flattened, creating a vector composed by 352 elements. This vector is then fed into a multi-layer perceptron, where the input layer is made of 352 neurons, the hidden layer is made of 32 neurons and the last layer is made of 2 neurons, one for each possible class. A soft-max function is also used in order to compute the output probability distribution.

2) Objective Functions or Optimization Functions: The filter weights are randomly initialized. These are further optimized, using the binary cross entropy as a loss function and the Adam optimizer algorithm is used to search for the optimal solution. The learning rate and ε are set to 10^{-5}, 10^{-6} respectively during the entire learning phase.

3) Cross-validation and Model Selection: During the training stage, all the samples of the training data were shuffled, and batches of size 128 were extracted from the shuffled training data. In order to evaluate the performance of the model, a validation dataset composed by 10% of the training data is used. The CNN is trained over 50 epochs, and at the end of each epoch, the CNN is evaluated and saved. Finally, at the end of the training phase, only the model that obtains the lowest loss value in the validation data is retained and used in the testing phase.

C. SVM Model

1) Configuration and Setup: An SVM with a radial basis function (RBF) kernel is used. The data is fed into an SVM, by first transforming an MFCCs image of size 300×13 into a column vector of size 3900, a process known as vectorization. To adjust the SVM parameters, we use a generic function that tunes hyperparameters of statistical methods using a grid search over supplied a parameter ranges.

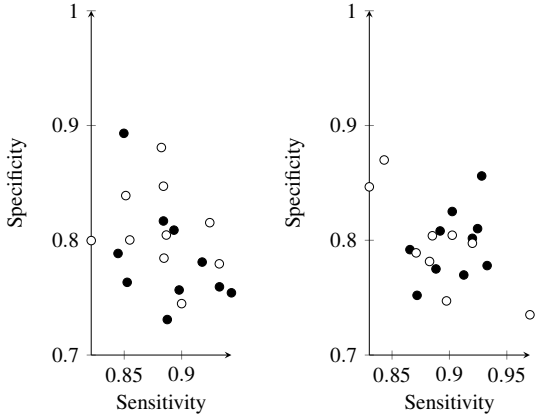

Fig. 1. Sensitivity and Specificity results provided by a GRU (on the left side) and CNN net (on the right side) based algorithm. Ten independent trials have been made, each symbol depicts a result computed from a independent test set. The PCG classifications are made based upon the best predefined threshold decision. The black circles display the results using the annotations provided by Physionet/Cinc Challenge. The white circles display the results without the support of any kind of annotation.

D. RF Model

1) Configuration and Setup: The data is fed into an RF model, by first transforming an MFCCs image of size 300×13 into a column vector of size 3900. To find the best values for the hyperparameters, we train the model several times, each time using different model hyperparameters. After training, the model is evaluated on a validation dataset composed by 10% of the training data. We then compare the performance in the validation dataset, and choose the hyperparameters with which the best accuracy was obtained. After a first phase of hyperparameter tunning, we use the same hyperparameters on all experiments performed in this work.

IV. Results

In this section, the impact of an explicit segmentation is measured on four different heart sound classifiers: GRU, CNN, SVM and RF. Two different avenues are compared: in the first one, the expert annotations are used to split the PCG signal into continuous and non-overlapping segments of three seconds, and starting at the beginning of an S1 state sample (black circles in Figures 1 and 2). In the second avenue, algorithms are asked to accomplish the same aforementioned task but without the support of any kind of human-made annotation. The PCG input signal is again split into continuous and non-overlapping segments of three seconds, starting immediately from the first sample. As a result, segments are free to start in any state sample (S1, Sys, S2, Dias). These results are displayed in Figures 1 and 2 with white circles. In the first experiment, the GRU net based algorithm explained in Section III-A is used to detect

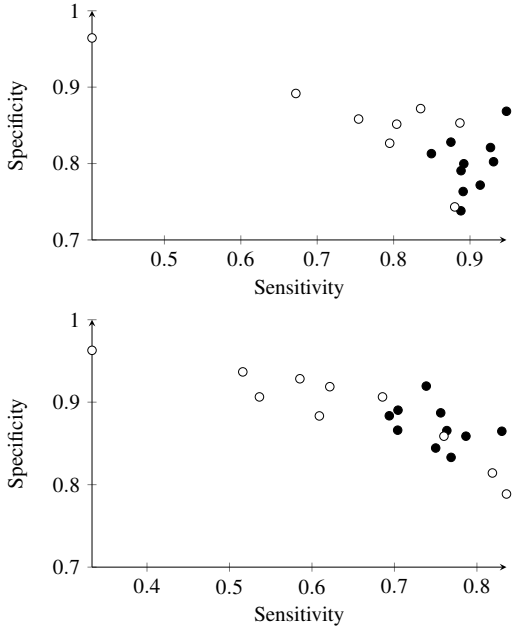

Fig. 2. Sensitivity and Specificity results provided by a SVM (on the top) and RF based algorithm (on the bottom), respectively. Ten independent trials have been made, each symbol depicts a result computed from a independent test set. The PCG classifications are made based upon the best predefined threshold decision. The black circles display the results using the annotations provided by Physionet/Cinc Challenge. The white circles display the results without the support of any kind of annotation.

abnormal heart sounds. The results show an unexpected similarity between the two strategies, see Figure 1. A GRU net with and without a segmentation step achieved an average overall performance of 0.838 ± 0.002 and 0.845 ± 0.001 with a m = 30%, respectively. Therefore, the standard explicit segmentation step (heart beat alignment), is not necessary or needed to train efficiently a GRU net algorithm to detect abnormal heart sounds, more specifically heart murmurs. This is in part justified by the reset and update gates of the GRU net which aim to keep a sort of consistent memory over time. Using such gates, the GRU net memorizes important events on the three seconds long MFCC segments and forget irrelevant facts, which are not considered important for the classification task. Furthermore, splitting the PCG signal into shorter audio segments of three seconds is made in order to surpass the lack of long-term memory of these networks, and also by the fact that murmur waves are quasi-periodic events, therefore likely to exist in every heartbeat of an unhealthy patient. In our second experiment, the CNN based algorithm, explained in Section III-B is used to detect abnormal heart sounds. The results show an unexpected similarity in terms of sensitivity and specificity over the two different scenarios, see Figure 1. A CNN net with and without a segmentation step achieved an average overall performance of 0.850 ± 0.001 and 0.841 ± 0.002 with a m = 10%, respectively. Therefore, the standard explicit segmentation step (heart beat alignment) is not necessary or needed to train efficiently a CNN algorithm to detect abnormal heart sounds, at least for murmur detection tasks. This is in part explained by the fact that CNNs are well known to be invariant to audio time shifts. Therefore, it is not important to have audio alignment, i.e. audio segments starting at the beginning of an S1 state sample, as long as murmur events are present in the input audio segment. From our previous analysis, no expensive human made annotations are needed to train effectively GRU and CNN nets. Or perhaps a more refined usage of the segmentation outcome is needed in order to boost significantly the score of GRU and CNN nets. In our third and forth experiment, the SVM and the RF based algorithms, explained in Section III-C and III-D respectively are used to detect abnormal heart sounds. A SVM with and without a segmentation step achieved an average overall performance of 0.850 ± 0.001 and 0.80 ± 0.02 with a m = 10%, respectively. Furthermore, a RF with and without a segmentation step achieved an average overall performance of 0.810 ± 0.001 and 0.76 ± 0.01 with a m = 10%, respectively. This is in part explained by the fact that such models, do not explore efficiently the temporal dependencies among events in a PCG signal.

V. Conclusion

In this paper, the impact of an explicit segmentation step on the classifier's ability to detect abnormal heart sounds is tested and measured experimentally. First, it is observed that the standard explicit heart sound segmentation step is not needed when GRU net or CNN classifiers are used to detect cardiac murmurs. Perhaps, a more robust segmentation step is needed and so it should be proposed by the scientific community, in order to completely address this important thematic. Secondly, SVMs and RFs are shown to be more dependent of an explicit segmentation step, in our experiments, the overall score dropped on average 5% on both algorithms.

References

[1] N. B. Mendis. Shanthi, Puska. Pekka, *Global atlas on cardiovascular disease prevention and control*. World Health Organization, World Heart Federation, 2011.

[2] J. Rubin, R. Abreu, A. Ganguli, S. Nelaturi, I. Matei, and K. Sricharan, "Recognizing abnormal heart sounds using deep learning," *CoRR*, vol. abs/1707.04642, 2017.

[3] S. Latif, M. Usman, R. Rana, and J. Qadir, "Phonocardiographic sensing using deep learning for abnormal heartbeat detection," *IEEE Sensors Journal*, vol. 18, pp. 9393–9400, 2018.

[4] T. Nilanon, J. Yao, J. Hao, S. Purushotham, and Y. Liu, "Normal/abnormal heart sound recordings classification using convolutional neural network," in *2016 Computing in Cardiology Conference (CinC)*, pp. 585–588, IEEE, 2016.

[5] C. Thomae and A. Dominik, "Using deep gated rnn with a convolutional front end for end:to:end classification of heart sound," *CinC*, 09 2016.

[6] G. D. Clifford, C. Liu, B. Moody, J. Millet, S. Schmidt, Q. Li, I. Silva, and R. G. Mark, "Recent advances in heart sound analysis," *Physiological measurement*, vol. 38, p. E10, 2017.

[7] C. Liu, D. Springer, Q. Li, and et. al., "An open access database for the evaluation of heart sound algorithms," *Physiological Measurement*, vol. 37, no. 12, p. 2181, 2016.

[8] D. M. Nogueira, C. A. Ferreira, and A. M. Jorge, "Classifying heart sounds using images of MFCC and temporal features," in *Progress in Artificial Intelligence - 18th EPIA 2017*, pp. 186–203, 2017.