Convolutional Neural Networks for Heart Sound Segmentation

Francesco Renna, Jorge Oliveira, Miguel T. Coimbra
Instituto de Telecomunicações, Faculdade de Ciências da Universidade do Porto, Portugal
Email: {frarenna, oliveira_jorge, mcoimbra}@dcc.fc.up.pt

***Abstract*—In this paper, deep convolutional neural networks are used to segment heart sounds into their main components. The proposed method is based on the adoption of a novel deep convolutional neural network architecture, which is inspired by similar approaches used for image segmentation. A further post-processing step is applied to the output of the proposed neural network, which induces the output state sequence to be consistent with the natural sequence of states within a heart sound signal (S1, systole, S2, diastole).**

The proposed approach is tested on heart sound signals longer than 5 seconds from the publicly available PhysioNet dataset, and it is shown to outperform current state-of-the-art segmentation methods by achieving an average sensitivity of 93.4% and an average positive predictive value of 94.5% in detecting S1 and S2 sounds.

I. Introduction

Cardiac auscultation is arguably the simplest, quickest, and most cost-effective first line of screening for a large number of heart conditions. On the other hand, heart sounds are difficult to identify and analyze by the human listener, as they are faint, significant events are closely spaced in time, and their frequency content is at the lower end of the audible frequency range [1]. These reasons have motivated recent research efforts in automatizing part or the entire process of analysis of the phonocardiogram (PCG) signal, in order to extract useful diagnostic information from it.

A key step required in the analysis of PCG signals is represented by the segmentation of heart sounds in their fundamental components. In fact, each heart cycle is normally divided into a first heart sound (S1), a systolic interval, a second heart sound (S2), and a diastolic interval. Extra sound components of interest are represented by the third and fourth heart sounds (S3 and S4), murmurs, ejection clicks, splits, etc.

Several solutions have been proposed in the literature to perform PCG segmentation (see [2] for a general overview). A first class of segmentation algorithms is based on the extraction of envelograms from the PCG signals and the use of peak-picking algorithms to estimate the principal heart sounds S1 and S2, as well as their boundaries. Such envelograms can be defined in the time domain (e.g., Shannon energy [3]), in the frequency domain (e.g. S-transform [4]), and other transform domains (e.g., wavelets [5], [6]).

A second class of segmentation algorithms leverages statistical models as the hidden Markov model (HMM) and the hidden semi-Markov model (HSMM) to include prior information about the sequential nature of PCG signals. In particular, the HMM has been adopted for heart sound segmentation by [7] and later by [8], which extended this approach by using Gaussian mixture model (GMM) priors to model the emission distributions. More recently, HSMMs have shown to outperform HMMs by introducing explicit modeling of the statistics of the time spent by the PCG signal in each state [9]. Then, refined modeling of the emission distributions have shown to improve the overall performance of HSMM-based segmentation algorithms: namely, support vector machines (SVMs) [10] were proposed and later emission distributions based on the logistic regression function were used [11], thus leading to state-of-the-art results.

A third class of segmentation algorithms is based on the extraction of features from the PCG, which are then assigned to the different heart sound states using a classifier. Some of the classifiers used for heart sound segmentation include SVMs [12], different kinds of artificial neural network (ANN) [13], and, more recently, deep neural networks [14].

In contrast with the variety of approaches presented in the literature, PCG segmentation still represents a challenging problem to solve when considering its application in real-world, noisy environments. In fact, the results of the 2016 PhysioNet Challenge on the classification of normal/abnormal heart sound recordings have shown that only a limited increase in classification performance can be achieved by using more sophisticated sound classifiers. On the other hand, improved segmentation algorithms are expected to be the best point of entry to obtain more significant improvements in heart sound classification [15].

This work proposes a novel heart sound segmentation approach, which is based on the use of a deep convolutional neural network (CNN). Although deep CNNs have been recently adopted for heart sound classification, i.e., to discriminate between normal and abnormal heart sounds [16], to the authors' knowledge, this work represents the first contribution using CNNs specifically for heart sound

This work is a result and funded by the project NanoSTIMA (NORTE-01-0145-FEDER-000016), supported by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF). It is also a result of the project DigiScope2 (POCI-01-0145-FEDER-029200), funded by Programa Operacional Competitividade e Internacionalização (POCI), in its FEDER component, and Fundação para a Ciência e Tecnologia, I.P., through national funds. It is also a result of the internal project SmartHeart in scope of project UID/EEA/50008/2013. This work was also funded by the FCT grant SFRH/BPD/118714/2016.

ISBN 978-90-827970-1-5 © EURASIP 2018

segmentation . The proposed method, unlike the deep learning solution in [14], does not require the adoption of a heart sound activity detection (HSAD) step to initially detect PCG segments before performing recognition, neither it relies on the extraction of *ad hoc* features from the signal. On the other hand, it can be applied directly to the PCG signal itself or to envelograms extracted from it. In this way, the sounds features that minimize segmentation errors are learnt directly from training data by the CNN.

These observations, jointly with the recent success achieved by CNN architectures in related tasks, e.g., image segmentation [17], motivate the use of deep CNNs for heart sound segmentation, which are expected to efficiently capture the characteristics of heart sounds corresponding to different PCG states, due to their ability in modeling complex signal behaviors and their robustness to high signal variability.

Therefore, this work proposes a novel segmentation approach that involves the following steps:

1) pre-processing of the PCG signal and extraction of envelograms from it;
2) application of a trained CNN to different portions of the envelograms extracted from the PCG signal;
3) combination of the CNN outputs corresponding to the different portions of the PCG, in order to produce the estimated state sequence.

The remainder of this paper is organized as follows: the proposed segmentation method is presented in Section II, and the experimental methodology and results are reported in Section III. Finally, conclusions are drawn in Section IV.

II. Methodology

In this section, the three main steps of the proposed segmentation algorithm are described in details. The proposed approach consists in a training phase and a testing phase. Signals involved in both training and testing are pre-processed according to the methods described in Section II-A. Labeled training data are used to determine the parameters (weights) that define the operations implemented by the CNN. In the testing phase, the trained CNN is applied to the pre-processed testing data, and the corresponding output undergoes a further post-processing stage in order to generate the estimated state sequence associated to each heart sound in the testing set (see Fig. 1).

A. Pre-processing

PCG signals are first filtered with high-pass and low-pass Butterworth filters of order two with cut-off frequencies equal to 25 Hz and 400 Hz, respectively. Then, the spike removal procedure described in [9] is applied to the filtered signals. The four envelograms/envelopes considered in [11] are extracted from the filtered signals: i) homomorphic envelogram, ii) Hilbert envelope, iii) wavelet envelope, and iv) power spectral density (PSD) envelope. Such envelograms are then downsampled at 50 Hz [11]. Finally, all the four envelograms/envelopes are normalized in order to have zero mean and unit variance.

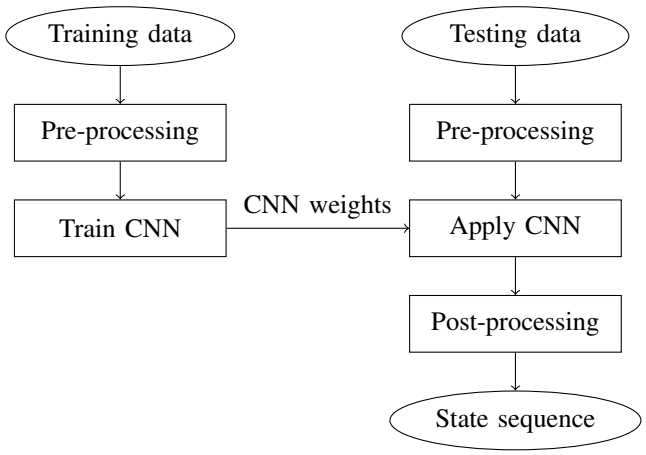


Fig. 1. Diagram of the proposed segmentation method: training phase and testing phase.

For each heart sound, the normalized envelograms are collected in the 4-dimensional signal $\mathbf{x}(t)$, where $\mathbf{x}(t) \in \mathbb{R}^4$ for $t = 0, \ldots, T-1$, and where t indicates the time instant. Then, $s(t)$ is defined as the sequence containing the state labels associated to each time instant, i.e., $s(t) \in \{0, 1, 2, 3\}$, where state 0 corresponds to an S1 sound, state 1 to a systole interval, state 2 to an S2 sound, and state 3 to a diastole interval.[1] Then, given a heart sound signal $\mathbf{x}(t)$, the objective of the segmentation algorithm is to provide an estimate of the corresponding state sequence $s(t)$.

Four-dimensional patches of fixed length n are extracted from the signal $\mathbf{x}(t)$, with a given stride τ. Such portions of the signal $\mathbf{x}(t)$ represent the inputs of the CNN. They are denoted by $\mathbf{X}_i \in \mathbb{R}^{n \times 4}$ and they are obtained as

$$\mathbf{X}_i = \begin{bmatrix} \mathbf{x}(i \cdot \tau) \\ \vdots \\ \mathbf{x}(i \cdot \tau + n - 1) \end{bmatrix}, \tag{1}$$

for $i = 0, \ldots, \left\lfloor \frac{T-1-n}{\tau} \right\rfloor$, where $\lfloor x \rfloor$ denotes the greatest integer lower than or equal to x.

B. Convolutional neural network architecture

Various CNN architectures have been presented in the quickly growing deep learning literature. This work proposes the use of a CNN architecture which is inspired by the U-net originally presented in [17] for image segmentation. The proposed architecture is reported in Fig. 2.

Such deep network contains convolutional layers which are followed by rectified linear unit (ReLU) activation functions [18]. The convolutional layers implement 1-dimensional convolutions of their inputs with different sets of filters (feature maps). Each filter in a feature map of a convolutional layer is defined by 3 weights. Moreover, the proposed architecture includes max pooling layers, which are responsible for downsampling (by a factor 2) the outputs of middle layers, as well as upsampling layers (by a

[1] In this work, we consider only the 4 main PCG components for segmentation, i.e., S1, systole, S2, and diastole.

ISBN 978-90-827970-1-5 © EURASIP 2018

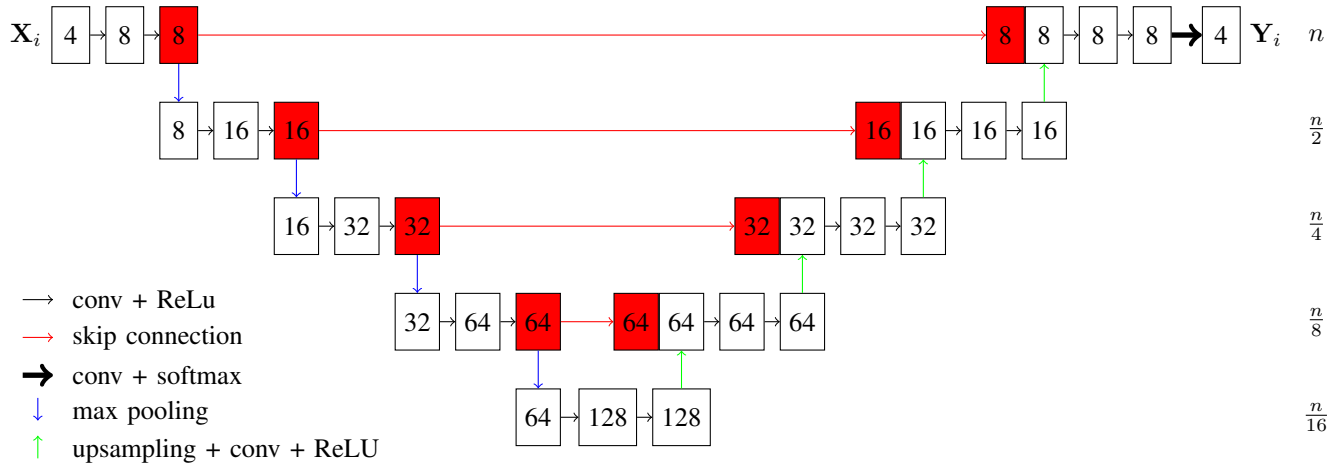


Fig. 2. Architecture of the deep CNN used in the proposed segmentation algorithm. The numbers inside the boxes in the diagram represent the number of feature maps in the corresponding convolutional layer. The numbers on the right hand side of the figure indicate the spatial dimension of the inputs and outputs of the layers contained in the corresponding row.

factor 2) which are interleaved with the late convolutional layers. This choice is fundamentally motivated by the fact that the adopted CNN structure implements a dyadic scale decomposition, which has a nice analog with a multiresolution wavelet decomposition, which has been shown to be effective for PCG segmentation [5], [6]. In fact, such CNN implements an encoder-decoder architecture, in the sense that the outputs of the middle layers offer a compact representation of the input signals in a low-dimensional manifold which contains the main information about the segmentation state of the PCG, thus reducing the impact of noise and signal variability. Moreover, it is possible to note that the overall receptive field of the proposed CNN is large, which means that information coming from neighboring portions of the input signal is effectively combined when deciding the state of a particular sample drawn from the PCG signal under test. Finally, skip connections are also inserted in the network, in order to allow direct transfer of information from the first layers to the late layers.

The inputs of convolutional layers are zero-padded so that the corresponding outputs have the same spatial dimensions as the input. Finally, the last convolutional layer implements 4 feature maps with a single weight each, and it is followed by a softmax activation function [18]. Therefore, on denoting by $\Phi_\theta(\cdot)$ the function implemented by the proposed CNN, where θ represents the set of weights that define the convolutional layers in the network, the CNN outputs are given by

$$\mathbf{Y}_i = \Phi_\theta(\mathbf{X}_i). \tag{2}$$

Then, each row of the matrix $\mathbf{Y}_i \in \mathbb{R}^{n\times 4}$ contains the probability that the corresponding sample of the PCG signal belongs to state 0, 1, 2, or 3.

C. Post-processing

Depending on the values of the patch size n and the stride τ, for a given time instant t of the heart sound signal there can be different output matrices $\mathbf{Y}_i$ containing the probabilities that such time instant belongs to one of the 4 available states. In fact, overlapping patches are used in order to minimize the impact on segmentation of data samples near the border of each patch. Therefore, the information obtained from different overlapping patches is combined by simply averaging the state probabilities associated to different output matrices $\mathbf{Y}_i$. Then, a first coarse estimate of the state sequence $s(t)$, which is denoted by $\tilde{s}(t)$, is obtained by choosing the state corresponding to the maximum probability among S1, systole, S2, or diastole.

On the other hand, it is straightforward to note that such state sequence $\tilde{s}(t)$ does not incorporate any prior information about the fact that only few transitions are available between different states, as an S1 event is always followed by a systole event, a systole event is always followed by an S2 event, etc. Therefore, in order to force the output sequence to contain only admissible transitions among states, a further post-processing step is performed on $\tilde{s}(t)$, which leads to the definition of the output sequence $\hat{s}(t)$ as follows: $\hat{s}(0) = \tilde{s}(0)$, and

$$\hat{s}(t) = \begin{cases} \tilde{s}(t) & , \text{ if } \tilde{s}(t) = (\hat{s}(t-1)+1) \mod 4 \\ \hat{s}(t-1) & , \text{ otherwise} \end{cases}, \tag{3}$$

for $t > 0$. Note that the output state sequence $\hat{s}(t)$ contains only admissible state transitions, thus providing a consistent way to localize the different segments of the heart sounds.

III. Experiments

The proposed segmentation algorithm is compared with the method described in [11], which is currently considered as the state-of-the-art PCG segmentation algorithm. Such method was also adopted as the segmentation standard for the 2016 PhysioNet Challenge on the classification of normal/abnormal heart sound recordings [19].[2]

[2]The HSMM-based segmentation algorithm in [11] was implemented using the code made available online by the authors at https://physionet.org/physiotools/hss/.

ISBN 978-90-827970-1-5 © EURASIP 2018

The performance of the proposed CNN-based segmentation algorithm and of the HSMM-based method in [11] is tested via 10-fold cross-validation on the available heart sound data. Namely, the available heart sound dataset is divided into 10 subsets. Then, tests are performed by selecting iteratively 1 out of this 10 subsets as the testing sets, while using the remaining 9 subsets for training. Note that subsets are divided so as to guarantee that sounds from patients contained in the testing set are not contained in the training set, in order to avoid overfitting.

The CNN used in the proposed segmentation method is trained with patches of dimension $n = 256$ that are extracted from the training recordings with a stride of $\tau = 32$ samples. The weights of the CNN are learnt using the categorical cross-entropy loss function [18] and the Adam optimizer [20] with learning rate equal to 10^{-4}. The maximum number of training epochs is fixed to 15, and early stopping is adopted by extracting 10% of the training data and using them for cross-validation, thus retaining the weights corresponding to the minimum loss function on cross-validation data. Note that the data used for cross-validation are not used to train the network.

A. Performance metrics

A first performance metric used in this work is the sample accuracy (A), which represents the fraction of samples in the output sequence $\hat{s}(t)$ that are correctly allocated to the corresponding state in the ground truth sequence $s(t)$.

Two further metrics are used to evaluate the performance of the proposed algorithm in determining the position of the fundamental heart sounds S1 and S2: positive predictive value (P_+) and sensitivity (S). Such metrics are computed according to the description in [9], i.e., a true positive (T_p) is counted when the center of an S1 (S2) sound in the estimated sequence $\hat{s}(t)$ is closer than 60 ms from the center of the corresponding S1 (S2) sound in the ground truth sequence $s(t)$. All other S1 and S2 sounds in the estimated state sequence are considered as false positives (F_p). Then, the positive predictive value P_+ is given by:

$$P_+ = \frac{T_p}{T_p + F_p}. \tag{4}$$

On the other hand, the sensitivity S is defined as:

$$S = \frac{T_p}{T_{tot}}, \tag{5}$$

where T_{tot} represents the total number of S1 and S2 sounds in the ground truth state sequence $s(t)$.

All performance metrics are computed for each recording in the test set and then averaged over the test set. Finally, the values corresponding to the different 10 test subsets are reported in Section III-C.

B. Materials

The heart sounds used for the experiments presented in this work were taken from the ensemble of sounds that were made publicly available for the PhysioNet/CinC challenge 2016. In particular, we considered 427 heart sounds recorded from 130 patients in different clinical and non-clinical environments.[3] From those, 181 sounds are collected from patients with pathological heart lesions (most commonly mitral valve prolapse), as assessed by echocardiography. The remaining 246 sounds are collected from healthy patients. Sound recordings have variable durations in the range from 5.12 to 35.5 seconds and they are sampled at 1 kHz. They are collected from several spots over the chest and they are possibly corrupted by different sources and noise levels. The annotations provided with the dataset are computed via the analysis of synchronous ECG recordings, based on the agreement between five different automatic R-peak and end-T-wave detectors [2].

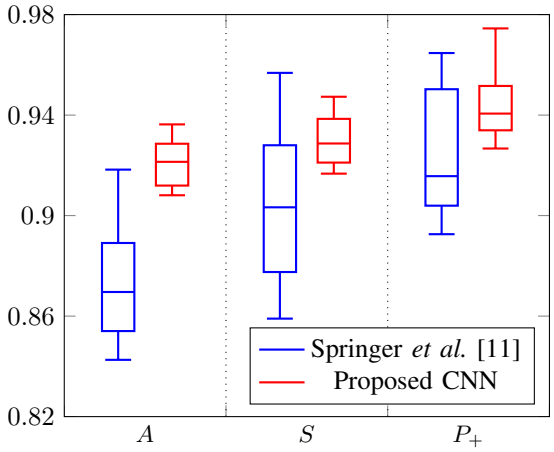


Fig. 3. Segmentation performance of the HSMM-based algorithm presented in [11] (blue boxes, on the left) and the proposed CNN-based approach (red boxes, on the right), when tested on heart sound recordings from the PhysioNet datasets. Boxplots of the sample accuracy (A), sensitivity (S), and positive predictive value (P_+) obtained from 10-fold cross-validation on the dataset.

C. Results and discussion

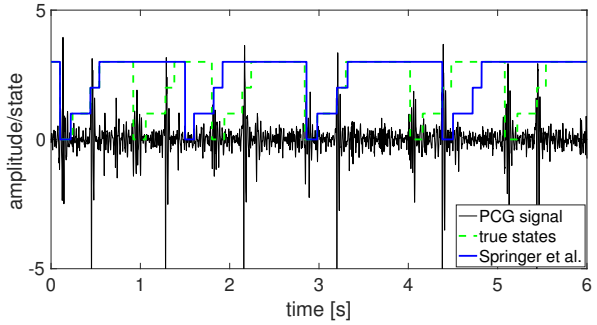
In Fig. 3 are reported the boxplots corresponding to the values of sample accuracy (A), sensitivity (S), and positive predictive value (P_+) achieved by the HSMM-based segmentation algorithm presented in [11] and by the proposed method.

It is possible to observe that the proposed CNN-based algorithm guarantees better performance than the HSMM-based algorithm for all the considered metrics. These results hint to the fact that the proposed CNN can capture more effectively the inter- and intra-sound state variability than the logistic regression model used in [11]. Moreover, the proposed CNN incorporates effectively information coming from neighboring time samples, due to the large receptive field guaranteed by the network architecture described in Section II-B. At the same time, the proposed solution allows for a more flexible modeling of the time spent by the PCG in

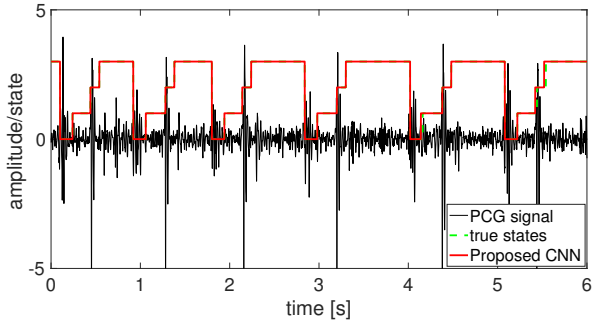
[3]The sounds are available online at https://PhysioNet.org/physiotools/hss/.

ISBN 978-90-827970-1-5 © EURASIP 2018

each state (sojourn time) with respect to the HSMM-based solution.



(a) HSMM-based algorithm by Springer *et al.* [11]



(b) Proposed CNN-based segmentation algorithm

Fig. 4. Segmentation example: PCG signal (black lines), true states derived from the dataset annotations (green, dashed lines), states assigned by the HSMM-based algorithm in [11] ((a), blue line), and states assigned by the proposed CNN-based segmentation algorithm ((b), red line).

This behavior can be clearly observed in Fig. 4, where it is possible to notice how the HSMM-based algorithm returns an estimated state sequence with larger diastole duration than the true state sequence. This means that a poor initialization of the HSMM-based algorithm, and in particular of the sojourn time statistics, can lead to significant errors, due to the lack of flexibility of this approach in modeling sojourn times. On the other hand, the proposed CNN-solutions identify almost perfectly the true state sequence of the noisy heart sound signal reported in Fig. 4.

IV. Conclusion

In this paper, the use of deep convolutional neural networks for heart sound segmentation was proposed. In particular, a novel network architecture, which is inspired by a network used for image segmentation, has been shown to guarantee improved performance in recovering the exact position of fundamental heart sounds in a phonocardiogram with respect to the current state-of-the-art. This behavior can be related to the enhanced capacity of the proposed method in discriminating among sound portions corresponding to different states as well as its flexibility in modeling the time evolution of the heart sound signal.

Future work arising from the initial results reported in this paper will involve more extensive testings with different heart sound datasets and a more thorough exploration of different CNN architectures and their associated parameter spaces. Moreover, more sophisticated post-processing procedures will be explored.

References

[1] A. Pease, "If the heart could speak," *Pictures of the Future*, pp. 60–61, 2001.

[2] C. Liu *et al.*, "An open access database for the evaluation of heart sound algorithms," *Physiological Measurement*, vol. 37, no. 12, pp. 2181–2213, 2016.

[3] H. Liang, S. Lukkarinen, and I. Hartimo, "Heart sound segmentation algorithm based on heart sound envelogram," in *Computers in Cardiology*, 1997, pp. 105–108.

[4] A. Moukadem, A. Dieterlen, N. Hueber, and C. Brandt, "A robust heart sounds segmentation module based on s-transform," *Biomedical Signal Processing and Control*, 2013.

[5] L. Huiying, L. Sakari, and H. Iiro, "A heart sound segmentation algorithm using wavelet decomposition and reconstruction," in *Engineering in Medicine and Biology Society, 1997. Proceedings of the 19th Annual International Conference of the IEEE*, vol. 4, Oct 1997, pp. 1630–1633 vol.4.

[6] A. Castro, T. T. V. Vinhoza, S. S. Mattos, and M. T. Coimbra, "Heart sound segmentation of pediatric auscultations using wavelet analysis," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, July 2013, pp. 3909–3912.

[7] D. Gill, N. Gavrieli, and N. Intrator, "Detection and identification of heart sounds using homomorphic envelogram and self-organizing probabilistic model," in *Computers in Cardiology*, 2005, pp. 957–960.

[8] Y.-J. Chung, *Pattern Recognition and Image Analysis, Iberian Conference*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, ch. Classification of Continuous Heart Sound Signals Using the Ergodic Hidden Markov Model, pp. 563–570.

[9] S. Schmidt, C. Holst-Hansen, C. Graff, E. Toft, and J. J. Struijk, "Segmentation of heart sound recordings by a duration-dependent hidden Markov model," *Physiological measurement*, vol. 31, no. 4, pp. 513–529, 2010.

[10] D. B. Springer, L. Tarassenko, and G. D. Clifford, "Support vector machine hidden semi-markov model-based heart sound segmentation," in *Computing in Cardiology Conference (CinC), 2014*, Sept 2014, pp. 625–628.

[11] ——, "Logistic regression-HSMM-based heart sound segmentation," *IEEE Trans. Biomed.l Eng.*, vol. 63, no. 4, pp. 822–832, 2016.

[12] J. Vepa, "Classification of heart murmurs using cepstral features and support vector machines," in *IEEE EMBC*, 2009, pp. 2539–2542.

[13] T. Leung, P. White, W. Collis, E. Brown, and A. Salmon, "Classification of heart sounds using time-frequency method and artificial neural networks," in *IEEE EMBC*, 2000, pp. 988–991.

[14] T.-E. Chen, S.-I. Yang, L.-T. Ho, K.-H. Tsai, Y.-H. Chen, Y.-F. Chang, Y.-H. Lai, S.-S. Wang, Y. Tsao, and C.-C. Wu, "S1 and S2 heart sound recognition using deep neural networks," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 2, pp. 372–380, 2017.

[15] 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.

[16] J. Rubin, R. Abreu, A. Ganguli, S. Nelaturi, I. Matei, and K. Sricharan, "Classifying heart sound recordings using deep convolutional neural networks and mel-frequency cepstral coefficients," in *Computing in Cardiology Conference (CinC)*, Sept 2016, pp. 813–816.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[18] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1.

[19] C. Liu, D. Springer, and G. D. Clifford, "Performance of an open-source heart sound segmentation algorithm on eight independent databases," *Physiological measurement*, vol. 38, no. 8, pp. 1730–1745, 2017.

[20] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

ISBN 978-90-827970-1-5 © EURASIP 2018