Detection of Heart Sound Murmurs and Clinical Outcome with Bidirectional Long Short-Term Memory Networks

Sofia Monteiro[1,2], Ana Fred[1,2], Hugo Plácido da Silva[1,2]

[1] Instituto Superior Técnico, Department of Bioengineering, Lisboa, Portugal
[2] Instituto de Telecomunicações, Lisboa, Portugal

Abstract

Heart sound recordings are a key non-invasive tool to detect both congenital and acquired heart conditions. As part of the George B. Moody PhysioNet Challenge 2022, we present an approach based on Bidirectional Long Short-Term Memory (BiLSTM) neural networks for the detection of murmurs and prediction of clinical outcome from Phonocardiograms (PCGs). We used the homomorphic, Hilbert, power spectral density, and wavelet envelopes as signal features, from which we extracted fixed-length segments of 4 seconds to train the network. Using the official challenge scoring metrics, our team SmartBeatIT achieved a murmur weighted accuracy score of 0.757 on the hidden test set (ranked 6th out of 40 teams), and an outcome cost score of 13815 (ranked 25th out of 39 teams). With 5-fold cross-validation on the training set, in the murmur detection task we obtained sensitivities of 0.827 and 0.312 for the Present and Unknown classes and a specificity of 0.801; and a sensitivity of 0.676 and a specificity of 0.544 in the outcome prediction task.

1. Introduction

The early detection of cardiovascular diseases is pivotal to prevent complications and premature deaths [1]. This is particularly challenging in developing countries, where there is an increased difficulty in the access to both specialized and primary health care. The use of automatic methods for detection of abnormalities based on the Phonocardiogram (PCG) could significantly facilitate the early diagnosis of both congenital and acquired heart conditions, as well as revolutionize how we approach health policies and disease management.

The goal of the George B. Moody PhysioNet Challenge 2022 was to identify, for each patient, whether any murmurs are discernible from heart sound recordings obtained from multiple auscultation locations and detect the clinical outcome [2, 3]. An in-depth description of the goals of the challenge and of the dataset can be found in [3, 4].

We present a deep learning approach based on Long Short-Term Memory networks (LSTM) for the classification of heart sounds. LSTMs are a type of recurrent neural network designed for the processing of sequential data and allow information to flow from one sample to next. Since PCG signals are a type of sequential data with a strong temporal correlation, they can be effectively processed by LSTMs for both segmentation and classification [5].

2. Methods

We implemented a Bidirectional LSTM (BiLSTM) neural network that classifies individual unsegmented PCG recordings based on temporal envelope features.

2.1. Temporal Features Extraction

PCG signals are high-dimensional sequence data, and as such processing them directly would incur a high computational cost. For this reason, instead of using the raw signal segments, we extracted four envelopes with a sampling frequency of 50 Hz. These not only provide a more compact description of the signals [6], but also reduce noise and other effects specific to the recording environment [7].

All the recordings were filtered with a 2^{nd} order Butterworth bandpass filter with cutoff frequencies of 25 Hz and 400 Hz, and normalized to have zero mean and unit variance. Then, using the method developed by Springer et al. [8], we calculated the homomorphic, Hilbert, power spectral density, and Wavelet envelopes, which have different trade-off levels between noise rejection and amplitude resolution [9]. These envelopes were also normalized to have zero mean and unit variance, and form a 4-dimensional multivariate time series for each heart sound segment.

Each recording was then decomposed into smaller fixed-length segments of 4-seconds. These segments still contain multiple cardiac cycles and enough information for the model to learn, and at the same time are small enough to allow us to significantly increase the amount of training data to build a more robust model [10].

A noteworthy aspect is that murmur classes are highly

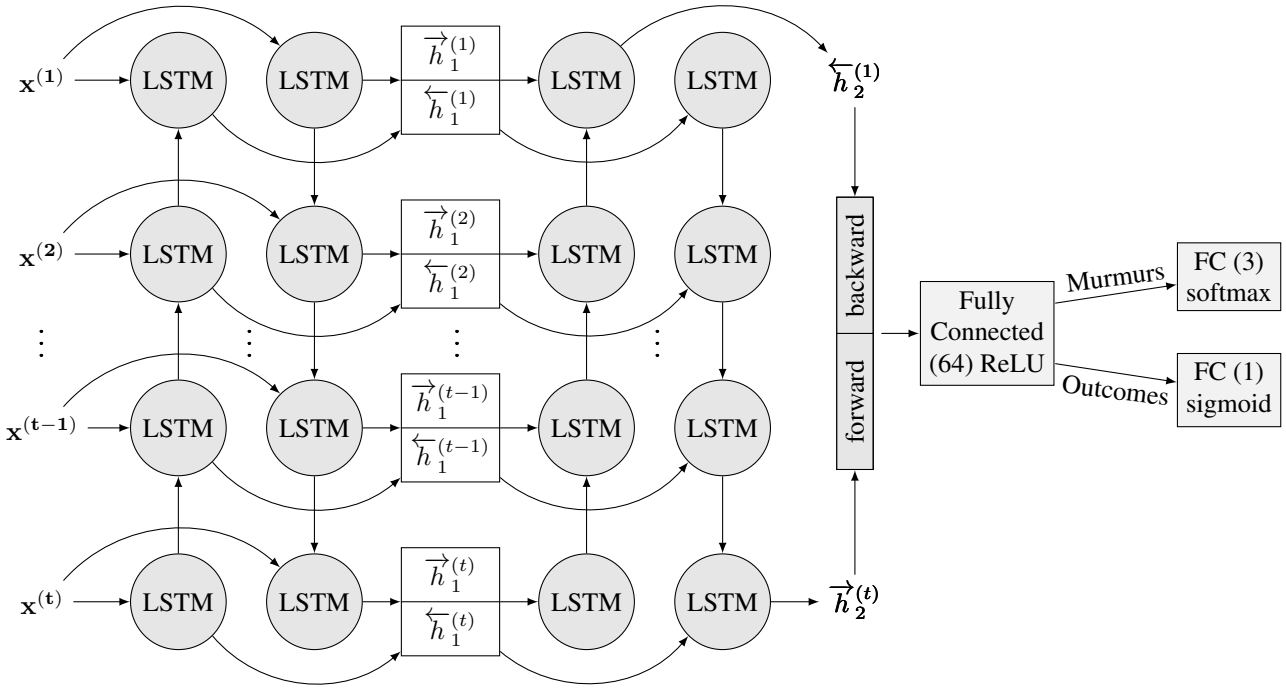

Figure 1. Bidirectional LSTM architecture. Each hidden layer has a number of cells equal to the number of timesteps, and $\overrightarrow{h}_i^{(t)}$ and $\overleftarrow{h}_i^{(t)}$ are, respectively, the hidden states of the forward and backward layers. $\mathbf{x(t)}$ represents the signal envelopes.

imbalanced, which presents a challenge for the training of deep learning models. The training set contains data from 942 patients; of those approximately 73.8% belong to the Absent class, 19% belong to the Present class, and only 7% belong to the Unknown class. On the other hand, the outcome classes are balanced between Normal (51.6%) and Abnormal (48.4%). To deal with the class imbalance in the murmur detection task, for the minority classes (Present and Unknown) the segments were extracted with 75% overlap. Recordings from patients in the Present class without an audible murmur were excluded from training.

2.2. BiLSTM Architecture

The architecture of the developed neural network is shown in Figure 1. It consists of stacked BiLSTM layers with 64 units. Unlike standard recurrent neural networks, LSTMs can remember or forget values over arbitrary periods of time, and thus are able to process long distance dependencies in the sequential data. This approach also processes the envelopes in both directions and can exploit both the past and future information of the signal.

The BiLSTM will read the entire sequence and then yield a single output, obtained by concatenating the hidden state of the last sample in the forward layer and the hidden state of the first sample in the backward layer. This is followed by fully connected layers for classification into one of the available classes. In the last layer, we used the softmax activation function for the murmur multiclass classification and the sigmoid activation function for the outcome binary classification [11].

2.3. Implementation

To select the number of stacked BiLSTM layers, we evaluated the local performance of our model for both tasks. The training set was split using 5-fold cross validation, ensuring that recordings from the same patient don't appear in two different folds, to avoid overfitting. The accuracy, sensitivity, specificity, and challenge scores for each model are presented in Tables 1 and 2.

In the murmur detection task, there is a significant increase in the performance of the model when the number of BiLSTM layers is increased from 1 to 2. However, when the number of layers increases from 2 to 3, the performance decreases. Even though deeper models should be able to learn more complex patterns in the signals, the small sample size can make the deeper network overfit to the training set and be unable to generalize to new data.

In the outcome detection task, the performance and scores of the model are not as correlated to the number of stacked BiLSTM layers. The best challenge score and highest sensitivity are obtained with a single layer, but the highest specificity is obtained with three layers.

Layers	Sen. % (Present)	Sen. % (Unknown)	Specificity %	Challenge Score %
1	77.6±4.0	24.7±15.1	77.6±5.2	60.0±4.8
2	**82.7±3.7**	**31.2±11.9**	**80.1±3.1**	**65.2±4.3**
3	78.4±5.0	25.2±11.7	78.4±5.0	60.2±7.4

Table 1. Training set results for the murmur detection task with different BiLSTM layers. Top results are in bold.

Layers	Sensitivity (%)	Specificity (%)	Challenge Score
1	**71.1±6.2**	53.4±9.3	**11532±635**
2	67.6±4.9	54.4±6.6	12434±401
3	67.6±5.3	**56.2±11.0**	12033±1092

Table 2. Training set results for the outcome prediction task with different BiLSTM layers. Top results are in bold.

The training was done using categorical cross-entropy as a loss function for murmur detection and binary cross-entropy for outcome prediction [11]. Given the class imbalance in the murmur detection task, we applied a weighing to the loss function to make the model pay more attention to the under-represented classes: each instance of the classes Present and Unknown is worth two instances of the class Absent. Table 3 shows the chosen parameters of the final BiLSTM implementation for both tasks.

To avoid overfitting, we used early stopping to monitor the loss of the model on a validation set obtained by randomly dividing the training data in each fold into 90% for training and 10% for validation. The networks were implemented with the Keras submodule from Tensorflow 2.6.1.

2.4. Multiple Instance Classification

The murmur and outcome classification tasks are cases of multiple-instance classification, given that each patient is represented by a set of instances (i.e. the recordings from the different auscultation locations), but it is the patient that carries the label [12]. Our approach was to train the model on individual recordings and then combine the instance-level decisions for the final patient label.

Network Parameters	
Hidden state dimension	64 units
Input segment size	200 samples
Nr. BiLSTM layers	2
Optimizer	SGD
Learning rate	10^{-3}
Momentum	0.9
Batch size	512
Max. number of epochs	300

Table 3. Selected parameters to train the BiLSTM.

The neural networks use 4-second segments as input so, to obtain one classification per recording, we split each signal into 4-second segments with 50% overlap, and combined the predictions by simply averaging the probabilities for each class and then selecting the class with the highest probability.

For the final patient labels, we assumed that a positive label contains at least one positive instance. In the murmur classification task, the final label and confidence scores for each patient were generated by selecting the recording with the highest probability for the Present class, since it is only necessary that the murmur is audible in one location to confirm its presence. Similarly, in the clinical outcome classification task, the final label and confidence scores were generated by selecting the recording with the highest probability for the Abnormal class.

3. Results

In Table 4 we present the sensitivity and specificity metrics for both tasks, as evaluated on the public training set with 5-fold cross-validation.

In the murmur detection task, even though the model could only reach an average sensitivity of 31.2% for the Unknown class, the sensitivity for the Present class and the specificity have high values, above 80%. In Table 5 we can see that the scores of the murmur detection task in the official hidden validation and test sets were superior to the scores obtained with cross-validation on the public data, which suggests that our model did not overfit to the training set. On the other hand, in the outcome prediction task, the model demonstrated low specificity and sensitivity, and a worse performance on the hidden test set (Table 6).

Murmur detection task		
Sensitivity (Present) %	Sensitivity (Unknown) %	Specificity %
82.7±3.7	31.2±11.9	80.1±3.1

Outcome detection task	
Sensitivity %	Specificity %
67.6±4.9	54.4±6.6

Table 4. Sensitivity and specificity metrics for both tasks on the public training set with 5-fold cross-validation.

Training	Validation	Test	Ranking
0.652±0.043	0.751	0.757	6/40

Table 5. Challenge weighted accuracy for the murmur detection task, with ranking on the hidden test set. We used 5-fold cross validation on the public training set, repeated scoring on the hidden validation set, and one-time scoring on the hidden test set.

Training	Validation	Test	Ranking
12434±401	11222	13815	25/39

Table 6. Challenge cost metric scores for the clinical outcome identification task, with ranking on the hidden test set. We used 5-fold cross validation on the public training set, repeated scoring on the hidden validation set, and one-time scoring on the hidden test set.

4. Discussion and Conclusions

In this work we proposed a model for automated murmur and outcome classification, and benchmarked its performance on the George B. Moody PhysioNet Challenge 2022 dataset. It is based on deep recurrent neural networks and temporal features, and demonstrated promising performance in the detection of murmurs.

In murmur detection, it is clear that the biggest difficulty of our model is the correct prediction of the Unknown class, which has a low sensitivity (Table 1). A possible explanation is the low sample size, which could make the instances of this class too varied for the model to learn in a way that can be generalized for new examples. Another possible reason is the fact that our model could be able to reliably identify the presence or absence of murmurs in these recordings, despite the inferior signal quality.

All the tested models had poor results in the outcome prediction task, with a low specificity and sensitivity. This could be due to the fact that the outcome labels result from an overall assessment of the patient's condition based on multiple examinations (such as clinical history, physical examination, analog auscultation, or echocardiogram), and not just from auscultation [3]. It is possible that some of the abnormalities cannot be diagnosed based solely on PCG data, meaning that the heart sounds do not contain all the information that is necessary for the model to learn.

Nonetheless, to improve these results, in future work we could supplement the envelope features with the provided demographic data, or try to adjust the decision thresholds to boost the score. We could also explore the use of attention mechanisms, which automatically learn the most relevant dependencies for each context, regardless of their distance in the sequence [13].

Acknowledgments

This work was partially funded by Fundação para a Ciência e Tecnologia (FCT)/Ministério da Ciência, Tecnologia e Ensino Superior through national funds and when applicable co-funded by EU funds under the project UIDB/50008/2020, by the European Regional Development Fund (FEDER) through the Operational Competitiveness and Internationalization Programme (COMPETE 2020), and by National Funds (OE) through the FCT under the LISBOA-01-0247-FEDER-069918 "CardioLeather".

References

[1] WHO. Cardiovascular Diseases (CVD). Accessed: 11 Aug, 2022; https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] 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.

[3] 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.

[4] Oliveira JH, Renna F, Costa P, Nogueira D, 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.

[5] Chen W, Sun Q, Chen X, Xie G, Wu H, Xu C. Deep Learning Methods for Heart Sounds Classification: A Systematic Review. Entropy 2021;23(6):667.

[6] Chen Y, Sun Y, Lv J, Jia B, Huang X. End-to-end Heart Sound Segmentation using Deep Convolutional Recurrent Network. Complex Intelligent Systems 2021;7(4):2103–2117.

[7] Ortiz JJG, Phoo CP, Wiens J. Heart Sound Classification Based on Temporal Alignment Techniques. In 2016 Computing in Cardiology (CinC). IEEE, 2016; 589–592.

[8] Springer DB, Tarassenko L, Clifford GD. Logistic Regression-HSMM-based Heart Sound Segmentation. IEEE Transactions on Biomedical Engineering 2016; 63(4):822–832.

[9] Renna F, Oliveira J, Coimbra MT. Deep Convolutional Neural Networks for Heart Sound Segmentation. IEEE Journal of Biomedical and Health Informatics 2019; 23(6):2435–2445.

[10] Latif S, Usman M, Rana R, Qadir J. Phonocardiographic Sensing Using Deep Learning for Abnormal Heartbeat Detection. IEEE Sensors Journal 11 2018;18(22):9393–9400.

[11] Bengio Y, Courville A, Goodfellow IJ. Deep Learning. Adaptive Computation and Machine Learning series. The MIT Press, 2016. 171-182.

[12] Alpaydin E, Cheplygina V, Loog M, Tax DM. Single- vs. Multiple-Instance Classification. Pattern Recognition 9 2015;48(9):2831–2838.

[13] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is All you Need. In Advances in Neural Information Processing Systems, volume 30. 2017;

Address for correspondence:

Sofia Monteiro

Instituto Superior Técnico, Dpt. of Bioengineering, Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal

sofia.m.monteiro@tecnico.ulisboa.pt