Ensemble of Feature-based and Deep learning-based Classifiers for Detection of Abnormal Heart Sounds

Cristhian Potes[1], Saman Parvaneh[1], Asif Rahman[1], Bryan Conroy[1]

[1]Philips Research North America, Acute Care Solutions, Cambridge, MA, USA

Abstract

The goal of the 2016 PhysioNet/CinC Challenge is the development of an algorithm to classify normal/abnormal heart sounds. A total of 124 time-frequency features were extracted from the phonocardiogram (PCG) and input to a variant of the AdaBoost classifier. A second classifier using convolutional neural network (CNN) was trained using PCGs cardiac cycles decomposed into four frequency bands. The final decision rule to classify normal/abnormal heart sounds was based on an ensemble of classifiers combining the outputs of AdaBoost and the CNN. The algorithm was trained on a training dataset (normal= 2575, abnormal= 665) and evaluated on a blind test dataset. Our classifier ensemble approach obtained the highest score of the competition with a sensitivity, specificity, and overall score of 0.9424, 0.7781, and 0.8602, respectively.

1. Introduction

Heart auscultation is the primary tool for screening and diagnosis in primary health care [1]. Availability of digital stethoscopes and mobile devices provides clinicians an opportunity to record and analyze heart sounds (PCG) for diagnostic purposes. The goal of the 2016 PhysioNet/CinC Challenge is the development of algorithms to classify normal/abnormal heart sound recordings [2]. We proposed an ensemble of a feature-based classifier and a deep learning-based classifier to boost the classification performance of heart sounds.

2. Method and Material

A block diagram of the proposed approach to classify normal/abnormal PCG is shown in Fig. 1.

2.1. Challenge Database

The challenge database provided PCG recordings of healthy subjects and pathological patients collected at either a clinical or non-clinical environment. Details about the challenge dataset can be found in [2]. For algorithm development, in-house training and test sets were generated by randomly taking 80% and 20% of the records from each database, while keeping the same prevalence of abnormal classes. In-house training set was used for training and cross-validation of different models, and in-house test set was used for evaluation of the classification performance independently from the blind test dataset.

2.2. Pre-processing

Each PCG was resampled to 1000 Hz, band-pass filtered between 25 Hz and 400 Hz, and then pre-processed to remove any spikes in the PCG [3]. Furthermore, pre-processed PCGs were segmented into four heart sound states using a segmentation method proposed by Springer et al. [4]. Each PCG is comprised of more than one cardiac cycle (beat), and each beat is comprised of four heart sound states (i.e. S1, systole, S2, and diastole).

2.3. Feature-based Approach

In this approach, a variant of AdaBoost classifier [5] was trained for classification of normal/abnormal PCGs using time and frequency-domain features.

2.3.1. Time-domain Features

Mean and standard deviation (SD) of the following parameters were used as time-domain features (36 features):
1. PCG intervals: RR intervals, S1 intervals, S2 intervals, systolic intervals, diastolic intervals, ratio of systolic interval to RR interval of each heart beat, ratio of diastolic interval to RR interval of each heart beat, ratio of systolic to diastolic interval of each heart beat.
2. PCG amplitudes: ratio of the mean absolute amplitude during systole to that during the S1 period in each heart beat, ratio of the mean absolute amplitude during diastole to that during the S2 period in each heart beat, skewness of the amplitude during S1 period in each heart beat, skewness of the amplitude during S2 period in each heart beat, skewness of the amplitude during systole period in each

heart beat, skewness of the amplitude during diastole period in each heart beat, kurtosis of the amplitude during S1 period in each heart beat, kurtosis of the amplitude during S2 period in each heart beat, kurtosis of the amplitude during systole period in each heart beat, kurtosis of the amplitude during diastole period in each heart beat

2.3.2. Frequency-domain Features

The power spectrum of each heart sound state (i.e. S1, systole, S2, and diastole) was estimated using a Hamming window and the discrete-time Fourier transform. The median power across nine frequency bands (i.e. 25-45, 45-65, 65-85, 85-105, 105-125, 125-150, 150-200, 200-300, 300-400 Hz) corresponding to the S1, S2, systole, and diastole states of each cardiac cycle was calculated. Then, the mean of the median power of the nine frequency bands for all cycles were used as frequency-domain features (i.e. 9 frequency bands $\times$ 4 states $=$ 36 features). Additionally, 13 mel-frequency cepstral coefficient (MFCC) [6] were extracted from each state and each cardiac cycle. The mean of MFCCs across different cardiac cycles from the same heart sound recording was used as MFCC features (i.e. 13 MFCCs $\times$ 4 states $=$ 52 features).

2.3.3. AdaBoost-abstain Classifier

AdaBoost is an effective machine learning technique for building a powerful classifier from an ensemble of "weak learners". Specifically, the boosted classifier $H(\boldsymbol{x})$ is modeled as a generalized additive model of many base hypotheses:

$$H(\boldsymbol{x}) = b + \sum_t \alpha_t h(\boldsymbol{x}; \boldsymbol{\theta_t}) \quad (1)$$

where b is a constant bias that accounts for the prevalence of the categories, and each base classifier $h(\boldsymbol{x}; \boldsymbol{\theta_t})$ is a function of $\boldsymbol{x}$, with parameters given by the elements in the vector $\boldsymbol{\theta_t}$, and produces a classification output ($+1$ or -1). In our approach, each base classifier is a simple decision stump over one of the above features. We also allow each of the base classifiers to abstain from voting (output=0) using a modified version of AdaBoost, AdaBoost–abstain [5]. A final classification decision is assigned by taking the sign of $H(\boldsymbol{x})$, which results in a weighted majority vote over the base classifiers in the model.

2.4. Convolutional Neural Network-based Approach

Each PCG recording was decomposed into four frequency bands (i.e. 25-45, 45-80, 80-200, and 200-400 Hz) and segmented to different cardiac cycles using PCG segmentation. The decomposed cardiac cycle with S1, systole, S2, and diastole was the input to the CNN network shown in Fig.2. Each cardiac cycle had a 2.5 seconds duration corresponding to the longest cardiac cycle found across all PCG recordings. If a cardiac cycle had a shorter duration, then the time series was zero padded.

As shown in Fig. 2, four time series, one per each frequency band, are the inputs to the network. Each of the CNNs consist of three layers: the input layer followed by 2 convolution layers. The input layer corresponds to the cardiac cycle of a specific frequency band (i.e. length = 2500 samples). Each convolutional layer involves a convolution operation, a nonlinear transformation, and a maxpooling operation. The first convolutional layer has 8 filters of length 5, followed by ReLu, and a max-pooling of 2. The second convolutional layer has 4 filters of length 5, followed by ReLu, and a max-pooling of 2. The output of the 4-CNNs are flattened and input to a multilayer perceptron (MLP) network. The MLP network consists of the input layer (i.e. flattened output of the 4-CNNs), a hidden layer with 20 neurons, and the output layer (i.e. one node). The activation function in the hidden layer is a ReLu and the activation function in the output layer is a sigmoid. The output layer computes the class score (i.e. probability value, CNN_ABN) of abnormal heart sound. Dropout of 25% was applied after max-pooling of the second convolutional layer. Dropout of 50% and L2 regularization was applied at the hidden layer of the MLP network. Adam was used for stochastic optimization, and cross-entropy was chosen as the loss function to minimize.

2.5. Final Decision Rule

The output of the two classifiers, AdaBoost-abstain (AdaBoost_ABN) and CNN (CNN_ABN), were combined using a decision rule shown below (Algorithm 1) to produce the final classification result (normal/abnormal). The corresponding thresholds (thr_ABN and thr_CNN) were tuned to maximize the overall challenge score on the in-house training set [2].

Algorithm 1 Decision Rule

if (AdaBoost_ABN $>$ thr_ABN)OR(CNN_ABN $>$ thr_CNN) **then**
Abnormal PCG
else
Normal PCG
end if

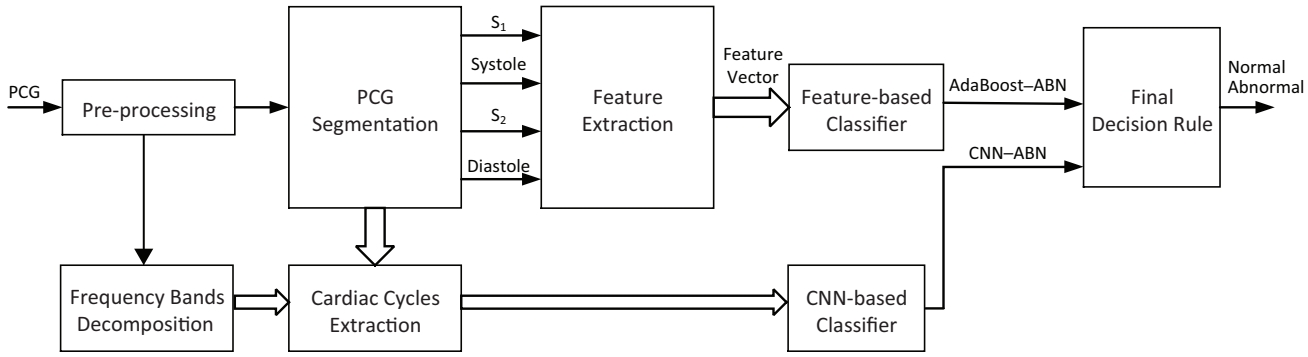

Figure 1. Block diagram of the proposed approach for classification of normal/abnormal heart sounds.

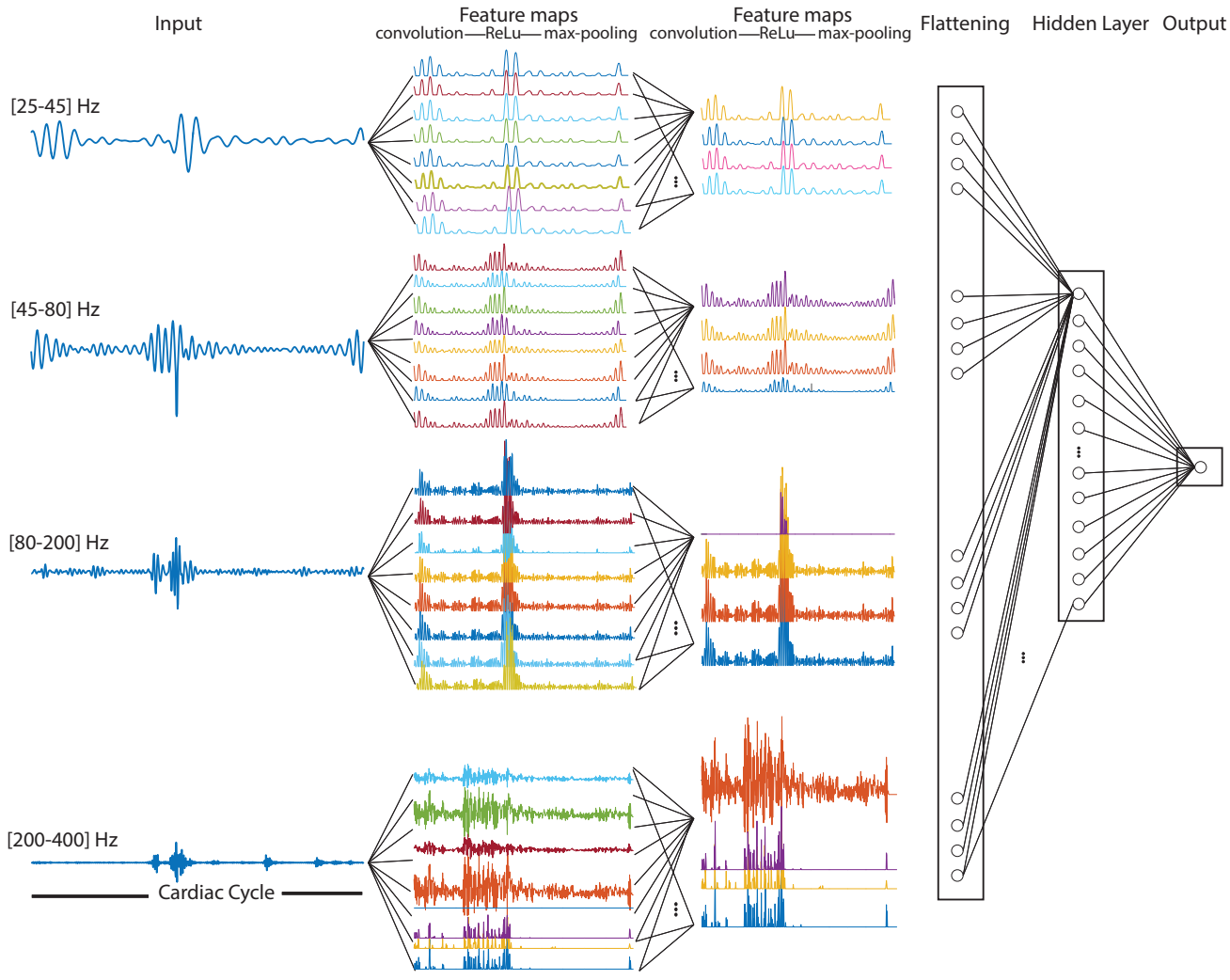

Figure 2. CNN architecture for classification of normal/abnormal heart sounds.

Table 1. Classification results on in-house test set and on a random subset of the blind test set (between parenthesis) when using AdaBoost-abstain, CNN, and their combination.

Classifier	Sensitivity	Specificity	Overall Score
AdaBoost-abstain	0.70 (0.88)	0.88 (0.82)	0.79 (0.85)
CNN	0.79 (0.88)	0.86 (0.80)	0.82 (0.84)
Classifier ensemble	0.88 (0.96)	0.82 (0.80)	0.85 (0.89)

3. Results

3.1. Classification Result using Feature-based Approach

One hundred twenty four features were fed into the AdaBoost-abstain classifier to classify normal/abnormal heart sounds; only 59 features were selected by the classifier after tuning parameters (e.g. number of iterations). Among the selected features, the top ten were the MFCC associated with S1, S2, and diastole states, SD of the kurtosis of the amplitudes during S1, and the mean and SD values of S1 and S2 intervals. AdaBoost–abstain provides an area under the receiver operating characteristic (AUC) of 0.91 on the in-house test set.

3.2. Classification Result using CNN-based Approach

Hyperparameters of the CNN network were tuned using the in-house training set, resulting in the following configuration: batch size of 1024, learning rate of 0.0007, and 200 epochs. Early stoppage was applied when the loss function stop decreasing. The CNN classifier provides a AUC equal to 0.92 on the in-house test set for classification of normal/abnormal heart sound.

3.3. Classification Result using Classifier Ensemble

The best results achieved at the official phase of the challenge using the complete blind test dataset was overall challenge score of 0.8602 (sensitivity and specificity equal to 0.9424 and 0.7781, respectively). The high sensitivity of our proposed algorithm is especially important for referring subjects for further screening. After the data challenge competition, the entry that led to the best classification results was downloaded and run on the in-house test set. The results achieved on the in-house test set using the AdaBoost-abstain, the CNN, and classifier ensemble are shown in Table 1. These results show that an ensembling of AdaBoost and CNN classifiers significantly increases the sensitivity (by 18% compared to AdaBoost alone, and 9% compared to CNN alone) but decreases the specificity (by 6% compared to AdaBoost alone and 4% compared to CNN alone). This conclusion is further confirmed with the results achieved from different entries at the official phase of the challenge (i.e. on a random subset of the blind test set) when using AdaBoost-abstain, CNN, classifier ensemble (see results between parenthesis in Table 1).

4. Conclusion

In this article, a novel approach for distinguishing normal/abnormal heart sounds is proposed that combines a classifier trained with time-frequency features and a deep-learning (CNN) classifier. Our results demonstrate the power of ensembling feature-based and representation learning classifiers for heart sound analysis.

Acknowledgments

The authors would like to thank Dr. Haibo Wang for his support setting up the GPU card for running the CNN models.

References

[1] Reed TR, Reed NE, Fritzson P. Heart sound analysis for symptom detection and computer-aided diagnosis. In Simulation Modelling Practice and Theory, volume 12. 2004; 129–146.

[2] Liu C, Springer D, Li Q, Moody B, Juan R, Chorro F, Castells F, Roig J, Silva I, Johnson A, Syed Z, Schmidt S, Papadaniil C, Hadjileontiadis L, Naseri H, Moukadem A, Dieterlen A, Brandt C, Tang H, Samieinasab M, Samieinasab M, Sameni R, Mark R, Clifford GD. An open access database for the evaluation of heart sound algorithms. Physiological Measurement 2016;37(9).

[3] 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;31(4):513–29.

[4] Springer DB, Tarassenko L, Clifford GD. Logistic regression-HSMM-based heart sound segmentation. IEEE Transactions on Biomedical Engineering 2016;63(4):822–832.

[5] Bryan C, Larry E, Cristhian P, Minnan XW. A dynamic ensemble approach to robust classification in the presence of missing data. Machine Learning 2015;.

[6] Kamarulafizam I, Salleh SH, Najeb JM, Ariff AK, Chowdhury A. Heart sound analysis using MFCC and time frequency distribution. In 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006, volume 14. ISBN 978-3-540-36839-7. ISSN 1680-0737, 2007; 946–949.

Address for correspondence:

Cristhian Potes
2 Canal Park, 3rd floor, Cambridge, MA 02141
cristhian.potes@philips.com