Classification of Normal/Abnormal Heart Sound Recordings: the PhysioNet/Computing in Cardiology Challenge 2016

Gari D Clifford[1,2], Chengyu Liu[1], Benjamin Moody[3], David Springer[4], Ikaro Silva[3], Qiao Li[1], and Roger G. Mark[3]

[1] Department of Biomedical Informatics, Emory University, Atlanta, USA

[2] Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, USA

[3] Institute for Medical Engineering & Science, Massachusetts Institute of Technology, USA

[4] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, UK

Abstract

In the past few decades heart sound signals (i.e., phonocardiograms or PCGs) have been widely studied. Automated heart sound segmentation and classification techniques have the potential to screen for pathologies in a variety of clinical applications. However, comparative analyses of algorithms in the literature have been hindered by the lack of a large and open database of heart sound recordings. The PhysioNet/Computing in Cardiology (CinC) Challenge 2016 addresses this issue by assembling the largest public heart sound database, aggregated from eight sources obtained by seven independent research groups around the world. The database includes 4,430 recordings taken from 1,072 subjects, totalling 233,512 heart sounds collected from both healthy subjects and patients with a variety of conditions such as heart valve disease and coronary artery disease. These recordings were collected using heterogeneous equipment in both clinical and nonclinical (such as in-home visits). The length of recording varied from several seconds to several minutes. Additional data provided include subject demographics (age and gender), recording information (number per patient, body location, and length of recording), synchronously recorded signals (such as ECG), sampling frequency and sensor type used. Participants were asked to classify recordings as normal, abnormal, or not possible to evaluate (noisy/uncertain). The overall score for an entry was based on a weighted sensitivity and specificity score with respect to manual expert annotations. A brief description of a baseline classification method is provided, including a description of open source code, which has been provided in association with the Challenge. The open source code provided a score of 0.71 (Se=0.65 Sp=0.76). During the official phase of the competition, a total of 48 teams submitted 348 open source entries, with a highest score of 0.86 (Se=0.94 Sp=0.78).

1. Introduction

Cardiovascular disease (CVD) continues to be the leading cause of morbidity and mortality worldwide with an estimated 17.5 million people having died from CVD-related conditions in 2012, representing 31% of all global deaths [1]. The burden on low to middle income countries (LMICs) is particularly worrisome, and yet high quality diagnostics can be often difficult to obtain in much of these resource constrained regions [2]. Although ultrasound and magnetic resonance imaging have displaced auscultation in the richer economies, heart sound auscultation remains a stalwart diagnostic of the ambulatory doctor. However, with patient to doctor ratios as high as 50,000:1 in some regions of the world, access to expert diagnosis is often impeded. A potential solution to this is to provide automated diagnosis on the mobile phone or in the cloud [2,3].

The 2016 PhysioNet/CinC Challenge seeks to create a large database to facilitate this, by drawing data from multiple research groups across the world, recorded in different real-world clinical and nonclinical environments (such as in-home visits). The data include not only clean heart sounds but also very noisy recordings, providing authenticity to the challenge. The data were also recorded from both normal subjects and pathological patients, providing a variety of signal sources. The data were also recorded from different locations, depending on the individual protocols used for each data set. However, they were generally recorded at the four common recording locations of aortic area, pulmonic area, tricuspid area and mitral area [4]. Until the current Challenge, no significant open-source heart sound database was available for researchers to train and evaluate the automated diagnostics algorithms upon.

The automated classification of pathology in heart sounds has been described in the literature for over 50 years, but accurate classification still remains a significant challenge. Typical methods for heart sound classification can be grouped into: artificial neural network-based classification, support vector machine-based classification, hidden Markov model-based classification and clustering-based classification. The current Challenge aims to encourage the development of algorithms to accurately classify heart sound recordings collected from a variety of clinical or nonclinical environments as normal or abnormal, and thus to further identify whether the subject of the recording should be referred on for an expert diagnosis. In addition, due to the uncontrolled environments, many recordings provided by the Challenge are corrupted by various noise sources. Classifications for the heart sound recordings were therefore three-level: normal (do not refer), abnormal (refer for further diagnostics) and unsure (too noisy to make a decision; retake the recording).

2. Challenge data

The data for this Challenge is describe in extensive detail in Liu *et al.* [4]. Briefly, the data consisted of eight heart sound databases collected independently by seven different research teams from seven countries, over a period of more than a decade. This resulted in a total of almost 30 hours of recordings containing 233,512 heart sounds from 116,865 heart beats, in 4,430 recordings taken from 1,072 subjects.

2.1. Expert labeling

All heart sound recordings were divided into two types based on expert labelling from the original data contributors: normal and abnormal. The normal recordings were from healthy subjects and the abnormal ones were from patients typically with heart valve defects and coronary artery disease (CAD). Heart valve defects include mitral valve prolapse, mitral regurgitation, aortic regurgitation, aortic stenosis and valvular surgery. All the recordings from the patients were generally labeled as abnormal. We do not provide more specific classification for these abnormal recordings.

2.2. Automated and hand corrected signal quality labels

To facilitate the challengers in training their algorithms to identify low signal quality recordings, we provided the signal quality labels. Signal quality was firstly evaluated using the method described by Springer *et al.* [5] and then manually hand-corrected by visual inspection. The recordings with poor signal quality (as judged by the researcher performing the hand correction) were labelled as 'unsure'.

2.3. Automated and hand correction of segmentation labels

Reference segmentation annotations of the four heart sound states (first sound S1, systole, second sound S2 and diastole) were provided for all recordings that were not labelled as 'unsure'. The reference annotations were obtained by applying the open source software (provided for the Challenge) developed by Springer *et al.* [6]. Subsequently, manual review was performed by a single individual to correct any obvious mistakes.

2.4. Training and test data

For all eight independent heart sound databases, four were divided into both training and test sets with a 70-30 training-test split. The other four databases were exclusively assigned to either training or test set. The training and test sets are two sets of mutually exclusive populations (i.e., no recordings from the same subject/patient are present in both training and test sets). The Challenge training set (a through f) includes a total of 3,153 heart sound recordings from 764 subjects/patients and the test set (b through e, plus g and i) included a total of 1,277 heart sound recordings from 308 subjects/patients. After the hand correction procedure for the segmentation annotations, there are a total of 84,425 beats in training set and 32,440 beats in test set. The recordings last from several seconds to up to more than one hundred seconds. All recordings were resampled to 2,000 Hz and have been provided in an uncompressed wav format. Table 1 briefly summarizes the Challenge data.

3. Example algorithms

The Challenge provided a very simple example benchmark classifier that relied on relatively obvious parameters extracted from the heart sound segmentation derived from the application of the provided open source code by Springer *et al.* [6]. First, a balanced heart sound database (472 abnormal and 472 normal recordings) from the training set was selected. Springers segmentation code was used to segment each selected heart sound recording to generate the time durations for the four states: S1, systole, S2 and diastole. Twenty features were extracted from the position information of the four states as detailed described in [4]. Then the twenty features were fed to a binary logistic regression classifier using forward selection to identify the most useful features. (The 'unsure' class was ignored and the output was therefore 'normal' or 'abnormal'.) Seven features were identified as the optimal number for classification.

In a 10 fold cross validation, a set of 5 features provided a sensitivity of 0.66, a specificity of 0.77 and a challenge

Database	# patients	# recordings	Proportion of recordings (%)			# beats
			Abnormal	Normal	Unsure	(after hand correction)
Training-*a*	121	409	67.5	28.4	4.2	14,559
Training-*b*	106	490	14.9	60.2	24.9	3,353
Training-*c*	31	31	64.5	22.6	12.9	1,808
Training-*d*	38	55	47.3	47.3	5.5	853
Training-*e*	356	2,054	7.1	86.7	6.2	59,593
Training-*f*	112	114	27.2	68.4	4.4	4,259
Total	764	3,153	18.1	73.0	8.8	84,425
Test-*b*	45	205	15.6	48.8	35.6	1,269
Test-*c*	14	14	64.3	28.6	7.1	853
Test-*d*	17	24	45.8	45.8	8.3	260
Test-*e*	153	883	6.7	86.4	6.9	26,724
Test-*g*	44	116	18.1	81.9	0	2,048
Test-*i*	35	35	60.0	34.3	5.7	1,286
Total	308	1,277	12.0	77.1	10.9	32,440

Table 1. Summary of the training and test sets used in 2016 PhysioNet/CinC Challenge. Note that there are approximately two heart sounds for each beat (S1 and S2).

score of 0.71 on the training data. It should be noted that this was not intended to be a good classifier, or properly trained, but merely an example set of code to enable a researcher to understand the mechanics of the submission process. The features were selected as trivially obvious and the classifier was not cross validated, and only a random subset of data were used.

We also implemented a simple unweighted voting algorithm by using the N best performing entries from the competition. We have shown in the past that this tends to give the highest score in any competition, assuming the weaker performers are removed.

4. Scoring

The overall score is computed based on the number of recordings classified as normal, abnormal or unsure. Table 2 details the determination rules.

The modified sensitivity *Se* and specificity *Sp* are defined as [4]:

$$Se = \frac{wa_1 \cdot Aa_1}{Aa_1 + Aq_1 + An_1} + \frac{wa_2 \cdot (Aa_2 + Aq_2)}{Aa_2 + Aq_2 + An_2},$$

$$Sp = \frac{wn_1 \cdot Nn_1}{Na_1 + Nq_1 + Nn_1} + \frac{wn_2 \cdot (Nn_2 + Nq_2)}{Na_2 + Nq_2 + Nn_2}.$$

where wa_1 and wa_2 are the percentages of good signal quality and poor signal quality recordings in all abnormal recordings respectively, and are used as weights for calculating *Se*, wn_1 and wn_2 are the percentages of good signal quality and poor signal quality recordings in all normal recordings respectively, and are used as weights for calculating *Sp*.

The overall Challenge *Score* is then given by:

$$MAcc = \frac{Se + Sp}{2}$$

Reference	Weights	Entry's output		
label		A (1)	U (0)	N (-1)
A, clean	wa_1	Aa_1	Aq_1	An_1
A, noisy	wa_2	Aa_2	Aq_2	An_2
N, clean	wn_1	Na_1	Nq_1	Nn_1
N, noisy	wn_2	Na_2	Nq_2	Nn_2

Table 2. Rules for determining the classification result. A: abnormal, U: unsure, N: normal.

5. Results, Discussions & Conclusions

A total of 348 open-source entries were submitted in the Challenge by 48 teams. Table 3 provides a breakdown of the top scoring entries ranked by $MAcc$. Although there is very little performance difference between the top three entries (in terms of the *MAcc*, we note that highest scoring entry by Potes *et al.* had a particularly high Se, and modest Sp. We also note that Potes *et al.* had the second highest overall Se. (The highest Se was 0.9633, but with a low Sp = 0.5589 and *MAcc*=0.7611, ranking 34th.) The third highest Se was as low as 0.8848, ranking 5th. Rubin *et al.* produced the highest Sp (0.9521), but with a relatively low Se of 0.7278 and an 8th place ranking. For an application which is forwarding subjects for further screening, as long as the resources can cope with the false positive rate, a higher sensitivity is perhaps best. However, the 2nd, 3rd and 4th contestants provide a good balance between Se

Rank	Entrant	*Se*	*Sp*	*MAcc*	Method note
1	Potes *et al.*	0.9424	0.7781	0.8602	AdaBoost & CNN
2	Zabihi *et al.*	0.8691	0.8490	0.8590	Ensemble of SVMs
3	Kay & Agarwal	0.8743	0.8297	0.8520	Regularized Neural Network
4	Bobillo	0.8639	0.8269	0.8454	MFCCs, Wavelets, Tensors & KNN
5	Homsi *et al.*	0.8848	0.8048	0.8448	Random Forest + LogitBoost
6†	Maknickas	0.8063	0.8766	0.8415	Unofficial entry - no publication
7	Plesinger *et al.*	0.7696	0.9125	0.8411	Probability-distribution based
8	Rubin *et al.*	0.7278	0.9521	0.8399	Convolutional NN with MFCs
17†	Voting of top N=38 algorithms	0.7120	0.9015	0.8068	Simple mode
43†	Sample entry	0.6545	0.7569	0.7057	See section 3

Table 3. Final scores for the top 8 of 48 entrants, the example algorithm provided and a simple voting approach. Best performances of competition entrants are in bold. † denotes an unofficial entry. MFCC = Mel Frequency Cepstral Coefficients. NN = Neural Network. SVM = Support Vector Machine. CNN = Convolutional NN. KNN = K Nearest Neighbors.

and Sp with only a 1.5% difference between them. A 2% spread exists between the top eight entrants.

Interestingly a voting algorithm was unable to beat 16 of the 48 contestants best entries. Using the N=38 best performing final independent entries ranked by the *MAcc* on the validation data results (not the hidden test data), a moderate score was achieved. In past competitions we have generally observed even a simple voting approach beats the best algorithm. In this particular case, the failure may be due to the fact that the training results were not representative of the algorithms performance on the test set. Moreover, a voting approach with features may lead to a far superior accuracy.

Finally, we note that the sample algorithm performed equally well on the training and test data. However, when stratifying by patient database, and performing a leave-one-database-out cross validation on the training data, the score varied between 0.47 on training set b and 0.86 on training set c, with a mean ± 1SD of 0.59 ± 0.15 across all training databases. This illustrates how hard it is to train an algorithm for new datasets with unseen recording conditions. The hidden test data contained two completely unseen databases (g and i).

In conclusion, the database provided for this Challenge represents the world's largest open access heart sounds database. We refer the reader to PhysioNet.org/challenge/2016 and our extensive documentation and online supplements from Liu *et al.* [4] for more details on the data and Challenge.

We also note that the database and algorithms are only a starting point. We hope to see the database grow and improve over time, particularly in response to the Challenge. A special issue in the journal *Physiological Measurement* will follow this competition and provide a forum for an extended editorial which will discuss the methods and results in more detail, provide the opportunity for entrants to revise their algorithms in light of other participants' methods, and address issues within the data.

Acknowledgments

This work was funded in part by the National Institutes of Health, grant R01-GM104987, the International Postdoctoral Exchange Programme of the National Postdoctoral Management Committee of China and Emory University. We are also grateful to Mathworks and Computing in Cardiology for sponsoring the competition prize money.

References

[1] WHO. 2015 world statistics on cardiovascular disease. URL `who.int/mediacentre/factsheets/fs317/en/`.

[2] Raghu A, Praveen D, Peiris D, Tarassenko L, Clifford GD. Engineering a mobile health tool for resource-poor settings to assess and manage cardiovascular disease risk: Smarthealth study. BMC Medical Informatics and Decision Making 2015;15(1):1–15.

[3] Springer D. Mobile phone-based rheumatic heart disease detection. Ph.D. thesis, University of Oxford, 2015.

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

[5] Springer D, Tarassenko L, Clifford GD. Automated signal quality assessment of mobile phone-recorded heart sound signals. Journal of Medical Engineering Technology 0; 0(0):1–14.

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

Address for correspondence:

Gari Clifford; gari@gatech.edu; http://gdclifford.info