HHS Public Access

Author manuscript

Physiol Meas. Author manuscript; available in PMC 2024 October 08.

Published in final edited form as:

Physiol Meas. ; 38(8): E10–E25. doi:10.1088/1361-6579/aa7ec8.

Editorial: Recent advances in heart sound analysis

Gari D. Clifford[1,2], **Chengyu Liu**[1], **Benjamin Moody**[3], **Jose Millet**[4], **Samuel Schmidt**[5], **Qiao Li**[1], **Ikaro Silva**[3], **Roger G. Mark**[3]

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

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

[3]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA

[4]ITACA Institute, Universitat Politecnica de Valencia, Valencia, Spain

[5]Department of Health Science and Technology, Aalborg University, Aalborg, Denmark

Abstract

Heart sounds have been widely studied and have been demonstrated to have value for detecting pathologies in clinical applications. Over the last few decades, the use of heart sound signals has become increasingly uncommon and its practice in modern medicine somewhat diminished, although research into automated analysis has continued. Unfortunately, a comparative analyses of algorithms in the literature have been hindered by the lack of high-quality, rigorously validated, and standardized open databases of heart sound recordings. The 2016 PhysioNet/Computing in Cardiology (CinC) Challenge addressed 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 comprises a total of 4,430 recordings collected from 1,072 healthy subjects and patients with a variety of conditions, including heart valve disease and coronary artery disease.

This editorial reviews the background issues for this Challenge, the design of the Challenge itself, the key achievements, and the follow-up research generated as a result of the Challenge, published in the concurrent special issue of *Physiological Measurement*. Additionally we make some recommendations for future changes in this the field of heart sound signal processing as a result of the Challenge.

In the Challenge, participants were asked to classify recordings as normal, abnormal, or unsure. The overall score for an entry was based on a weighted sensitivity and specificity score with respect to manual expert annotations. To aid researchers, we provided a simple baseline classification method and a complex open source code base for segmenting the heart sounds, based on a hidden semi-Markov model.

During the official phase of the Challenge, a total of 48 teams submitted 348 open source entries, with a highest score of 0.860 (Se=0.942, Sp=0.778). Subsequently, for this special issue, researchers reported the new highest score of 0.855 (Se=0.890, Sp=0.816) in the follow-up phase

gari@gatech.edu .

of the Challenge, indicating that the Challenge entrants achieved exceptional results which were extremely dicult to improve (even when there is a trade-off between Sp and Se) upon in the 4 months available post-Challenge. We expect that future researchers will be able to use the extensive database generated for the Challenge to significantly improve on the approaches detailed here.

1. Introduction

Auscultation of heart sound recordings or the phonocardiogram (PCG) has been shown to be valuable for the detection of disease and pathologies (Leatham (1975); Raghu et al. (2015)). The automated classification of pathology in heart sounds has been studied for over 50 years. Typical methods can be grouped into: artificial neural network-based approaches (Uguz (2012)), support vector machines (Ari et al. (2010)), hidden Markov model-based approaches (Saracoglu (2012)) and clustering-based approaches (Quiceno-Manrique et al. (2010)). However, accurate automated classification still remains a significant challenge due to the lack of high-quality, rigorously validated, and standardized open databases of heart sound recordings.

The 2016 PhysioNet/Computing in Cardiology (CinC) Challenge sought to create a large database to facilitate this, by assembling recordings from multiple research groups across the world, acquired in different real-world clinical and nonclinical environments (such as in-home visits), to encourage the development of algorithms to accurately identify, from a single short recording (10-60s), as normal, abnormal or poor signal quality, and thus to further identify whether the subject of the recording should be referred on for an expert diagnosis (Liu et al. (2016)). Until this Challenge, no significant open-access heart sound database was available for researchers to train and evaluate the automated diagnostics algorithms upon (Clifford et al. (2016)). Moreover, no open source heart sound segmentation and classification algorithms were available. The Challenge changed this situation significantly.

This editorial reviews the follow-up research generated as a result of the Challenge, published in the concurrent special issue of *Physiological Measurement*. Additionally we make some recommendations for promising research avenues in the field of heart sound signal processing and classification as a result of the Challenge.

2. Challenge data

Data for the Challenge consisted of heart sound recordings from eight independent databases (labelled alphabetically, a to i, excluding h, which was a fetal PCG database) sourced from seven contributing research groups. We refer the reader to Liu et al. (2016) for a detailed description of the data collection, as well as the division of training and test data sets. We should note that both training and test sets are unbalanced, i.e., the number of normal recordings does not equal that of abnormal ones. Challengers therefore had to consider this when they trained and test their algorithms. Figure 1 details the exact distribution of data across all the constituent databases.

3. Example algorithms and scoring

3.1. Benchmark classifier algorithm

We provided a benchmark classifier that relied on relatively obvious parameters extracted from the heart sound segmentation code. For the detailed description of this benchmark classifier, challengers can refer to Liu et al. (2016); Clifford et al. (2016). Here we briefly describe how the benchmark classifier is constructed and how it works. First, a balanced database from training set was selected. Then, Springers segmentation code (Springer, Tarassenko and Clifford (2016)) was used to segment heart sound recording. Twenty features were extracted according to the position and waveform amplitude information of the segmented signals. A forward likelihood ratio selection was used to train the binary logistic regression (BLR) model. Finally, seven features were identified as the predictable features, and a derived BLR prediction formula was constructed for normal/abnormal heart sound recordings classification. In a 10 fold cross validation, the constructed BLR model provided a sensitivity of 0.66, a specificity of 0.77 and a Challenge 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, and to provide a simple baseline for Challenge entrants to beat in the early stages of the Challenge.

3.2. Voting algorithm

We also implemented a voting approach to combine together varying numbers of the submitted algorithms (Clifford et al. (2016)). A simple unweighted voting of using the N best performing final entries from the Challenge, ranked by their score on the training data (to prevent over-fitting on the test scores), was implemented. N was varied from 1 to 48 with tied, absent or no vote was treated as ‘normal’ type.

3.3. Scoring

A modified accuracy ((MAcc)) with the combination of sensitivity (Se) and specificity (Sp) for scoring as:

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

The score on the complete test set determines the ranking of the entries. For details on the scoring mechanism please see Liu et al. (2016); Clifford et al. (2016).

4. Results of the Challenge

A total of 348 open-source entries were submitted in the Challenge by 48 teams. Table 1 provides a detailed summary for the top official scoring entries published in the CinC conference proceedings, ranked by the $MAcc$ index. Please note that we did not include the unofficial entries here. We reported the best Challenge scores (Se, Sp and $MAcc$) for each team from the complete hidden test data. We also summarized the methods the challengers used, mainly focusing on the following:

A total of 348 open-source entries were submitted in the Challenge by 48 teams. Table 1 provides a detailed summary for the top official scoring entries published in the CinC conference proceedings, ranked by the *MAcc* index. We reported the best Challenge scores (*Se, Sp* and *MAcc*) for each team from the complete hidden test data. We also summarized the methods the challengers used, mainly focusing on the following:

1. The type of segmentation procedure, if any, employed.
2. Types of features used.
3. Number of features used.
4. How features selection was performed, if at all.
5. What and how many features remained after feature selection, if applicable.
6. What classifier was used.
7. For training the classifier, how the training data were split.
8. How the researchers adjusted for class imbalances during training.

From Table 1, it can be seen that there was very little performance difference between the top three entries. The highest scoring entry by Potes et al. had a *MAcc* of 0.8602, with a highest *Se* (0.9424) and a modest *Sp* in the list. The second highest *Se* was as low as 0.8848, ranking 5th in the Challenge. Rubin et al. produced the highest *Sp* (0.9521), but with a relatively low *Se* of 0.7278 and ranked a 7th place. 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, 4th and 5th contestants provide a good balance between *Se* and *Sp*. A 2% spread exists between the top six entrants.

The sample entry generated a *Se* of 0.6545 and a *Sp* of 0.7569, resulting in a *MAcc* of 0.7051. To test if the results could be improved by combining multiple approaches, we designed a “voting” algorithm as follows. We calculated the performance of each of the 348 official entries, using a set of 600 records that were selected randomly from the public training data, but disjoint from the validation subset that competitors used for self-scoring. We then ranked entries according to their modified accuracy on this subset, and discarded all but the top entry from each participating team. The “voting” algorithm V_N (for $N = 2 \ldots 48$), is then defined as the output given by a plurality of the top N entries from that list (or 0, “uncertain”, if no plurality exists.) The voting algorithm did not show any improvement over the best individual submissions; the best result was $N = 3$, with $Se = 0.7173$, $Sp = 0.9309$, and $MAcc = 0.8241$.

5. Review of Articles in the Special Issue

A total of 8 articles were reviewed and revised in time to be accepted for this special issue. Most authors had originally entered the Challenge, and submitted updated versions of their algorithms, which should be made available by the authors through open source licenses. Each algorithm published in this issue is reviewed below according to the eight aspects summarized in section Results of the Challenge. The purpose of this summary is to allow the reader to quickly identify both the commonalities and the originality of all the approaches.

Finally, the last article in this special issue and review (Liu et al. (2017)) involves the systematic evaluation for the open source code for heart sound segmentation proposed in Springer, Tarassenko and Clifford (2016), which was also the heart sound segmentation method made available for the Challenge.

5.1. Abdollahpur et al. (2017)

The algorithm proposed by Abdollahpur et al. (2017) used a novel cycle quality assessment (CQA) method for assessing the signal quality of the segmented cardiac cycle. Features were extracted only on the cycles which higher signal quality and superior segmentation. The method achieved a *MAcc* of 0.8263 in the last phase of the Challenge (Abdollahpur et al. (2016)).

The authors note that the recordings were down sampled to 1 kHz and filtered by the fourth order Butterworth high pass (25 Hz) and low pass (600 Hz) filters. Spikes were removed using the algorithm proposed by Schmidt et al. (2010). Then, after the heart sound segmentation with Springer's HSMM model (Springer, Tarassenko and Clifford (2016)), correctly segmented heart cycles without excessive noise or spikes were selected for further feature extraction process using a novel CQA method detailed in Abdollahpur et al. (2016). Frequency and amplitude criteria were applied for detecting correctly segmented heart sound cycles. A total of 90 features were calculated from the time domain, time-frequency, perceptual and mel-frequency cepstral coefficient (MFCC) analysis. Before starting the main classification process, the derived 90 dimensional feature vector was mapped to a new feature space by applying a Fishers discriminant analysis. The main classification procedure was then performed using three feed-forward NNs and a voting system among classifiers. A final *MAcc* score of 0.826 was achieved on the hidden test data.

5.2. Homsi and Warrick (2017)

The algorithm proposed by Homsi and Warrick (2017) used an ensemble based classification with a special consideration for outliers and achieved a *MAcc* score of 0.801 for the hidden test data in the Challenge.

In this paper, a total of 131 features in time, frequency, wavelet and statistical domains were extracted from the heart sound signals. Outlier signals were detected and separated from those with a standard range using an interquartile range threshold. Then, feature extreme values were given special consideration, and finally features were reduced to the most significant ones using a feature reduction technique. In the classification stage, the selected features either for standard or outlier signals were fed separately into an ensemble of 20 two-step classifiers. The first step of the classifier included a nested set of ensemble algorithms which was cross validated on the training data, while the second step used a voting rule of the class label. The results showed that the proposed method achieved an overall score of 0.9630 for standard signals and 0.9018 for outlier signals on a cross-validated experiment using the training data. This method achieved an overall score of 0.801 on the hidden test set (0.796 sensitivity and 0.806 specificity).

5.3. Kay and Agarwal (2017)

Kay and Agarwal (2017) proposed an algorithm that employed DropConnected neural networks trained on time-frequency and inter-beat features for heart sound classification. This algorithm achieved a $MAcc$ of 0.8520 on the test data, and ranked third in the Challenge (Kay and Agarwa (2016)). This paper provides an extensive analysis concerning the profile differences of the open training data, including the recording numbers, recording sensors, unbalanced data and the specific pathology of the recordings.

In this paper, first, the heart sounds were segmented using Springer's the open-source segmentation algorithm based on a hidden semi-Markov model (HSMM) (Springer, Tarassenko and Clifford (2016)). Then, a total of 675 features were extracted from the analysis of continuous wavelet transform (220), MFCC (400), inter-beat behaviour (20 and complexity measures (35). Then, the extracted features were normalized and the dimensionality was reduced to 50 using principal component analysis (PCA). Subsequently, the features were used as the input to a fully-connected, two-hidden-layer neural network, trained by error backpropagation, and regularized with DropConnect. When the algorithm was submitted to be evaluated on the test data, a number of different networks were trained with a range of hyper-parameters and different training sets. The networks are then ensembled based on their scores. The best result obtained by the ensemble of networks, on the test data, was 0.8520, which is the third best performance in the Challenge. The authors also updated their algorithm by excluding the training-e set for training since the recording sensor type for training-e set is different from others. However, a significantly worse score of 0.580 was obtained because 69% of recordings in the test set are from dataset-e indicating that the algorithm is sensitive to the recording type and struggles to generalize from one dataset to another.

5.4. Langley and Murray (2017)

Most algorithms for automated analysis of heart sound require segmentation of the signal into the characteristic heart sounds. Langley and Murray (2017) aimed to assess the feasibility for accurate classification of heart sounds on short, unsegmented recordings.

At the first step, initially the 5 second segment (seg 1) at the start of each heart sound recording was analyzed. For some recordings with considerable noise at the start of the recordings, so a repeated 5 s segments (seg 2) with lowest noise was extracted for each recording. Segments were zero-mean but otherwise had no prepossessing or segmentation. Then normalized spectral amplitude was determined by FFT and wavelet entropy was calculated by wavelet analysis ('Gaus4' mother wavelet). For each of these a simple single feature threshold based classifier was implemented and the frequency/scale and thresholds for optimum classification accuracy determined. The analysis was then repeated using relatively noise free 5 s segments (seg 2) of each recording by applying a Wavelet entropy measure for signal noise assessment. Spectral amplitude and wavelet entropy features were then combined in a classification tree (Langley and Murray (2016)). Detailed results were reported as follows. There were significant differences between normal and abnormal recordings for both wavelet entropy and spectral amplitude across scales and frequency. In the wavelet domain the differences between groups were greatest at

highest frequencies whereas in the frequency domain the differences were greatest at low frequencies (12 Hz). Abnormal recordings had significantly reduced high frequency wavelet entropy, suggesting the presence of discrete high frequency components in these recordings. Abnormal recordings exhibited significantly greater low frequency (12 Hz) spectral. Classification accuracy was greatest for wavelet entropy and was further improved by selecting the lowest noise segment (seg 2). Classification tree with the combined features gave an accuracy (not $MAcc$) of 0.79 (Sp = 0.80, Se = 0.77). The study demonstrated the feasibility of accurate classification without segmentation of the characteristic heart sounds.

5.5. Maknickas and Maknickas (2017)

Maknickas and Maknickas (2017) describe the use of mel-frequency spectral coefficients (MFSC) fed to a CNN, and which achieved a $MAcc$ of 0.8415 in the last phase of the Challenge, ranked sixth overall with an unofficial entry. There are existing studies which leverage MFCC analysis for heart sound classification Chauhan et al. (2008). However, the authors claimed that MFSC analysis could outperform MFCC since during the calculation of the MFCC, the discrete cosine transform (DCT) projects the spectral energies into a new basis that may not maintain locality. However, MFSC uses the log-energy computed directly and can avoid this situation.

In this paper the authors describe a process which first splits the training heart sound files into equal numbers of normal and abnormal data files. Then MFSC (i.e., MFCC with no DCT) was calculated for each file, and was cut into frames with width and height of both 128 samples. The difference and second-order difference of the MFSC were also calculated as second and third dimensions of the frame. All frames were normalised. Then CNN was trained to predict the normal/abnormal label for each frame in the file, and used the average of all predicted frame labels as the final label of the file. Finally, the model with best performance was selected during the training phase. Testing on the separate validation set achieved the highest score when using 256 hidden layers for the deep CNN, although the score slightly improved on the selected training data when increasing the number of hidden layers from 128 to 2048. Therefore, the Challenge results were achieved by weights and topology of 256 hidden layers and the final score was 0.842, just 0.018 below the highest score of 0.860. This impressive result indicates the potential of CNNs for future use, but also illustrates how enormous volumes of data are likely to be required to out-perform well chosen features and standard classification approaches.

5.6. Plesinger et al. (2017)

Plesinger et al. (2017) proposed an algorithm based on fuzzy logic which they termed ‘probability assessment’ for normal/abnormal heart sound classification, which achieved a $MAcc$ of 0.8411 in the last phase of the challenge, and was ranked 7^{th} highest (Plesinger et al. (2016)). The presented solution produced different results in specific databases. For database-c, it gave 100% sensitivity and specificity in both training and testing. Database-e also provided an extremely high score. However, the method failed to accurately classify database-g and database-i (not present in the training set), where it reported nearly all records as normal. This poor performance with these completely hidden databases indicates the method also struggles to generalize to unseen data.

In their methods, they first derived amplitude envelopes in five frequency bands low frequency (LF, 15-90 Hz), middle frequency (MF, 15-90 Hz), high frequency (HF, 100-250 Hz), super frequency (SF, 200-450 Hz) and ultra frequency (UF, 400-800 Hz) were computed using an FFT band-pass filter and Hilbert transformation. Then invalid time segments were checked for each 1 s window. Then heart sounds S1 and S2 were detected using amplitude envelopes in the LF band. The averaged shapes of the S1/S2 pair were computed from amplitude envelopes in all five bands (15-90 Hz; 55-150 Hz; 100-250 Hz; 200-450 Hz; 400-800 Hz). A total of 228 features were extracted from the statistical properties and the symmetry of the averaged shapes, and the independent of S1 and S2 detection. Then the features are processed using logical rules and probability assessment based on histograms, and a fuzzy logic like approach, which they termed ‘PROBAfind’. This software contains a function suggesting a feature with the best impact on the sum of final sensitivity and specificity, and can be used as a semi-automatic feature selection method. The authors found 53 features were selected as the normal/abnormal/unsure classification. A final score $MAcc$ of 0.8411 achieved on the hidden test data (7^th place in the Challenge), indicating that the performance of probability assessment is comparable to other machine-learning approaches. However, it the human oversight required and long training time required for this approach is a significant limitation and may have led to the lack of generalization.

5.7. Whitaker et al. (2017)

Whitaker et al. (2017) proposed an algorithm combining sparse coding and time domain features for normal/abnormal heart sound classification, which achieved a $MAcc$ of 0.807 in the Challenge (Whitaker and Anderson (2016)). This study introduced sparse coding as a tool for unsupervised feature extraction in heart sound classification, and was also the first to use matrix norm sparse coding in a practical classification setting for Heart Sounds. Previous work by Da Poian et al. (2017) has demonstrated the utility of this technique, using on compressed sensing for Atrial Fibrillation detection in the ECG. As the first step, Whitaker et al. used Springer’s HSMM segmentation code (Springer, Tarassenko and Clifford (2016)) to separate each audio file into five arrays of smaller audio segments. The first four arrays contained a list of all S1, systole, S2 and diastole sounds respectively. The fifth array contained copies of the full heart cycles, starting at the start of the S1 state and ending at the last sample in diastole. Each state or sound segment was converted to the frequency domain with an N-point FFT and sparse coding was applied on the aforementioned five data matrices as a form of unsupervised feature extraction. In sparse coding, frequency-domain data is decomposed into a dictionary matrix and a sparse coefficient matrix. The dictionary matrix represents statistically important features of the audio segments and becomes fixed after training. In effect it represents the basis functions. The sparse coefficient matrix is a mapping that represents which features are useful in each segment. Working in the sparse domain, the authors trained SVMs for each audio segment, as well as the full cardiac cycle. Then a sixth SVM was trained to combine the results from the preliminary SVMs into a single binary label for the entire heart sound recording. Compared with the CinC paper in Whitaker and Anderson (2016), this paper presented two novel modifications. The first modification involved a matrix norm in the dictionary update step of sparse coding to encourage the dictionary to learn discriminating features from the abnormal heart

recordings. The second combined the sparse coding features with twenty time domain features described in Liu et al. (2016) in the final SVM classification stage. The authors demonstrated an improved cross-validated *MAcc* of 0.893 (*Se* = 0.901 and *Sp* = 0.885). However, improved version did not generate a higher score on the hidden test data than their challenge's score. A new score *MAcc* of 0.803 (0.801 sensitivity and 0.806 specificity) in this follow-up phase was achieved.

This study showed that sparse coding is an effective way to define spectral features of the cardiac cycle and its sub-cycles for the purpose of classification. In addition, it demonstrated that sparse coding can be combined with additional feature extraction methods to improve classification accuracy. Further work may incorporate additional features to improve the classification accuracy or robustness to novel data and noise.

5.8. Liu et al. (2017)

A Hidden Markov model (HMM)-based approach has received increased interest for heart sound segmentation due to its robustness on processing noisy recordings, particularly when incorporating physiological models. The focus of this article was on evaluating the performance of the recently published logistic regression based HSMM heart sound segmentation method Springer, Tarassenko and Clifford (2016), which was open sourced for the Challenge. By using a wider variety of heart sound data in the PhysioNet/CinC Challenge 2016. The HSMM-based model was trained on the training-a dataset only (per the original work) and was tested on all other separate test datasets, which comprised 102,306 heart sounds. The results confirm the high accuracy of the HSMM-based algorithm with an average F_1 score of 98.5% for segmenting S1 and systole intervals and 97.2% for segmenting S2 and diastole intervals. The described evaluation framework, combined with the largest collection of open access heart sound data, provides essential resources for researchers who need to test their algorithms with realistic data and share reproducible results.

6. Discussion and Conclusions

In summary, the PhysioNet/Computing in Cardiology Challenge 2016 provided several key additions to the field of normal/abnormal heart sound classification.

First, the public release of the large, open assess and free heart sound database gives potential benefits to a wide range of users, especially for those who lack access to well-characterized real clinical signals.

Second, we note that even for the top performing entrants, the classification results differ significantly between each of the eight databases. The test sets g and i are two new databases and did not appear in the training data. For those two hidden databases, the challenger results are not as good as other databases, indicating that the algorithm generalization ability is sensitive to the recording source and requires improvement, or should always be retrained for specific recording scenarios and/or recording modalities/devices.

Third, there is very little performance difference between the top three entries, and only a 2% spread exists between the top six entrants, although these Challenge entrants used different classifier methods. This shows that there is not a ”best” classifier for this special normal/abnormal heart sound classification task. However, the ensemble method, i.e., combining two or more of the common classification methods, such as SVM, CNN, LR, RF and others, can create improved classification performances. We note however, that a naive approach of simple weighted voting between the top N algorithms ranked by training performance does not improve the modified accuracy and a more intelligent voting approach is needed - see below. Notably, the feature extraction stage in any classification related work can be the most crucial and important part. Although there are no widely accepted optimal features in heart sound classification, from this Challenge we can identify the MFCC, wavelet and time-frequency features as likely candidates.

Fourth, we note that voting method can produce superior results to even the best algorithm. Such an approach can also lead to a more robust implementation, although it may be significantly more computationally intensive. It is also important to note that too many naive voters can reduce the classification accuracy, as we have observed in earlier challenges, although not in this one. This may be due to the common use of a strong feature extractor provided for all entrants. In Zhu et al. (2014) and Zhu et al. (2015) a voting system for algorithms (and human) annotations of physiological data was described, which incorporates both the physiology and the individual annotator’s accuracy as a function of objective features (such as signal quality) to produce a weighted voting scheme to guarantee that all voters added extra information. Such approaches may become ever more important as computational power becomes increasingly less expensive. We also note that this means that all competitors in the Challenge added something to the final answer!

Fifth, the current approach in this Challenge classifies any input signal as normal or abnormal although “unsure” class was permitted. However, an efficient algorithm is needed for recognizing a good quality recording from a poor quality one. Due to the audio processing capabilities, mobile phones have the potential to facilitate the diagnosis of heart disease through automated auscultation. However, such a platform is likely to be used by non-experts, and hence, it is essential that such a device is able to automatically differentiate poor quality from diagnostically useful recordings since non-experts are more likely to make poor-quality recordings. In Springer, Brennan, Ntusi, Abdelrahman, Zuhlke, Mayosi, Tarassenko and Clifford (2016), an automated signal quality assessment of heart sound recordings was developed, which includes the first systematic evaluation of a heart sound signal quality classification algorithm (using a separate test dataset) and assessment of the quality of heart sound recordings captured by non-experts. This approach indicates a promising use case for low resource cardiac screening.

Sixth, we provided a state-of-the-art open source heart sound segmentation algorithm for this Challenge. This was utilized by the top entrants and indicates that it was fundamental to high performing classification algorithms. We note however that no researcher attempted to improve on the algorithm in either the Challenge or the subsequent special issue. The marginal increase in performance in this special issue indicates that improving the segmentation approach may be the best point of entry for any future researchers attempting

to improve classification performance. The inability of more complex classifiers (such as CNNs) to beat carefully chosen features and standard classifiers, indicates that it is more important to focus on the labelling and preprocessing than on the classifier. That is not to say that a superior classifier can provide an increase in performance, but that the feature extraction step provides more marginal improvement. We also note that despite our databases representing the largest public dataset of heart sound by many orders of magnitude, the databases may require a significant increase in size before deep learning is able to show any significant performance gains.

Finally we note some limitations of the Challenge. Although we have collated and provided all collected information from the data contributors, more detailed pathological information is needed for the heart sound recordings. Detection and properly identification of mitral stenosis, aortic stenosis and mitral insufficiency among others is still a challenge. We intend to work with industry and researchers alike to enhance the Challenge database in all these areas and would be grateful for continued contributions of data and source code, which we will post together with all the open source algorithms and annotated data from the 2016 PhysioNet/Computing in Cardiology Challenge. The latter can be found on PhysioNet's website at http://physionet.org/challenge/2016.

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 for providing free software licenses and sponsoring the Challenge prize money, and Computing in Cardiology for sponsoring the Challenge prize money and providing a forum to present the Challenge results. We would also like to thank the database contributors, and data annotators for their invaluable assistance. Finally, we would like to thank all the competitors and researchers themselves, without whom there would be no Challenge or special issue.

References

Abdollahpur M, Ghaffari A, Ghiasi S and Mollakazemi JM (2017). Detection of pathological heart sounds, Phys. Meas 38(8).

Abdollahpur M, Ghiasi S, Mollakazemi MJ and Ghaffari A (2016). Cycle selection and neuro-voting system for classifying heart sound recordings, Computing in Cardiology Conference (CinC), 2016, IEEE, pp. 1–4.

Ari S, Hembram K and Saha G (2010). Detection of cardiac abnormality from pcg signal using lms based least square svm classifier, Expert Syst. Appl 37: 80198026.

Banerjee R, Biswas S, Banerjee S, Choudhury AD, Chattopadhyay T, Pal1 A, Deshpande P. and Mandana KM (2016). Time-frequency analysis of phonocardiogram for classifying heart disease, Computing in Cardiology Conference (CinC), 2016, IEEE, pp. 573–576.

Bobillo IJD (2016). A tensor approach to heart sound classification, Computing in Cardiology Conference (CinC), 2016, IEEE, pp. 629–632.

Bouril A, Aleinikava D, Guillem MS and Mirsky GM (2016). Automated classification of normal and abnormal heart sounds using support vector machines, Computing in Cardiology Conference (CinC), 2016, IEEE, pp. 549–552.

Chauhan S, Wang P, Sing LC and Anantharaman V (2008). A computer-aided mfcc-based hmm system for automatic auscultation, Comput. Biol. Med 38(2): 221–233. [PubMed: 18045582]

Clifford GD, Liu C, Moody B, Springer D, Silva I andRoger G Mark QL. (2016). Classification of normal/abnormal heart sound recordings: the physionet/computing in cardiology challenge 2016, Computing in Cardiology Conference (CinC), 2016, IEEE, pp. 609–612.

Homsi MN, Medina N, Hernandez M, Quintero N, Perpian G, Quintana A and Warrick P (2016). Automatic heart sound recording classification using a nested set of ensemble algorithms, Computing in Cardiology Conference (CinC), 2016, IEEE, pp. 817–820.

Homsi MN and Warrick P (2017). Ensemble methods with qutliers for phonocardiogram classification, Phys. Meas 38(8).

Kay E and Agarwa A (2016). Dropconnected neural network trained with diverse features for classifying heart sounds, Computing in Cardiology Conference (CinC), 2016, IEEE, pp. 617–620.

Kay E and Agarwal A (2017). Dropconnected neural networks trained on time-frequency and inter-beat features for classifying heart sounds, Phys. Meas 38(8).

Langley P and Murray A (2016). Abnormal heart sounds detected from short duration unsegmented phonocardiograms by wavelet entropy, Computing in Cardiology Conference (CinC), 2016, IEEE, pp. 545–548.

Langley P and Murray A (2017). Heart sound classification from unsegmented phonocardiograms, Phys. Meas 38(8).

Leatham A (1975). Auscultation of the Heart and Phonocardiography, Churchill Livingstone, London.

Liu C, Springer D and Clifford GD (2017). Performance of an open-source heart sound segmentation algorithm on eight independent databases, Phys. Meas 38(8).

Liu C, Springer D, Li Q, Moody B, Juan RA, Chorro FJ, Castells F, Roig JM, Silva I, Johnson AEW, 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 and Clifford GD (2016). An open access database for the evaluation of heart sound algorithms, Phys. Meas 37(12): 21812213.

Maknickas V and Maknickas A (2017). Recognition of normal-abnormal phonocardiographic signals using deep convolutional neural networks and mel-frequency spectral coefficients, Phys. Meas 38(8).

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

Ortiz JJG, Phoo CP and Wiens J (2016). Heart sound classification based on temporal alignment techniques, Computing in Cardiology Conference (CinC), 2016, IEEE, pp. 589–592.

Plesinger F, Jurco J, Jurak P and Halamek J (2016). Discrimination of normal and abnormal heart sounds using probability assessment, Computing in Cardiology Conference (CinC), 2016, IEEE, pp. 801–804.

Plesinger F, Viscor I, Halamek J, Jurco J and Jurak P (2017). Heart sounds analysis using probability assessment, Phys. Meas 38(8).

Poian GD, Liu C, Bernardini R, Rinaldo R and Clifford GD (2017). Atrial fibrillation detection on compressed sensed ecg, Physiological Measurement 38(7): 1405–1425. **URL:** http://stacks.iop.org/0967-3334/38/i=7/a=1405 [PubMed: 28569241] **URL:**

Potes C, Parvaneh S, Rahman A and Conroy B (2016). Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds, Computing in Cardiology Conference (CinC), 2016, IEEE, pp. 621–624.

Quiceno-Manrique AF, Godino-Llorente JI, Blanco-Velasco M and Castellanos-Dominguez G (2010). Selection of dynamic features based on time-frequency representations for heart murmur detection from phonocardiographic signals, Ann. Biomed. Eng 38: 118–137. [PubMed: 19921435]

Raghu A, Praveen D, Peiris D, Tarassenko L and Clifford GD (2015). Engineering a mobile health tool for resource-poor settings to assess and manage cardiovascular disease risk: Smarthealth study, BMC Medical Informatics and Decision Making 15(1): 1–15. [PubMed: 25889846]

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

Ryu H, Park J and Shin H (2016). Classification of heart sound recordings using convolution neural network, Computing in Cardiology Conference (CinC), 2016, IEEE, pp. 1153–1156.

Saracoglu R. (2012). Hidden markov model-based classification of heart valve disease with pca for dimension reduction, Eng. Appl. Artif. Intell 25: 1523–1528.

Schmidt SE, Holst-Hansen C, Graff C, Toft E and Struijk JJ (2010). Segmentation of heart sound recordings by a duration-dependent hidden markov model, Physiol. Meas 31(4): 513–529. [PubMed: 20208091]

Singh-Miller NE and Singh-Miller N (2016). Using spectral acoustic features to identify abnormal heart sounds, Computing in Cardiology Conference (CinC), 2016, IEEE, pp. 557–560.

Springer DB, Brennan T, Ntusi N, Abdelrahman HY, Zuhlke LJ, Mayosi BM, Tarassenko L and Clifford GD (2016). Automated signal quality assessment of mobile phone-recorded heart sound signals, J. Med. Eng. Technol 40(7): 342–355. [PubMed: 27659352]

Springer DB, Tarassenko L and Clifford GD (2016). Logistic regression-hsmm-based heart sound segmentation, IEEE Trans. Biomed. Eng 63: 822–832. [PubMed: 26340769]

Tang H, Chen H, Li T and Zhong M (2016). Classification of normal/abnormal heart sound recordings based on multi-domain features and back propagation neural network, Computing in Cardiology Conference (CinC), 2016, IEEE, pp. 593–596.

Tschannen M, Kramer T, Marti G, Heinzmann M and Wiatowski T (2016). Heart sound classification using deep structured features, Computing in Cardiology Conference (CinC), 2016, IEEE, pp. 565–568.

Uguz H (2012). A biomedical system based on artificial neural network and principal component analysis for diagnosis of the heart valve diseases, J. Med. Syst 36(1): 61–72. [PubMed: 20703748]

Whitaker BM and Anderson DV (2016). Heart sound classification via sparse coding, Computing in Cardiology Conference (CinC), 2016, IEEE, pp. 805–808.

Whitaker BM, Suresha PB, Liu CY, Clifford GD and Anderson DV (2017). Combining sparse coding and time-domain features for heart sound classification, Phys. Meas 38(8).

Yang TI and Hsieh H (2016). Classification of acoustic physiological signals based on deep learning neural networks with augmented features, Computing in Cardiology Conference (CinC), 2016, IEEE, pp. 569–572.

Yang X, Yang F, Gobeawan L, Yeo SY, Leng S, Zhong L and Su Y (2016). A multi-modal classifier for heart sound recordings, Computing in Cardiology Conference (CinC), 2016, IEEE, pp. X1165–1168.

Yazdani S, Schlatter S, Atyabi SA and Vesin JM (2016). Identification of abnormal heart sounds, Computing in Cardiology Conference (CinC), 2016, IEEE, pp. 1157–1160.

Zabihi M and Rad AB (2016). Heart sound anomaly and quality detection using ensemble of neural networks without segmentation, Computing in Cardiology Conference (CinC), 2016, IEEE, pp. 613–616.

Zhu T, Dunkley N, Behar J, Clifton DA and Clifford GD (2015). Fusing continuous-valued medical labels using a bayesian model, Ann. Biomed. Eng 43(12): 2892–2902. [PubMed: 26036335]

Zhu T, Johnson AEW, Behar J and Clifford GD (2014). Crowd-sourced annotation of ecg signals using contextual information, Ann. Biomed. Eng 42(4): 871–884. [PubMed: 24368593]

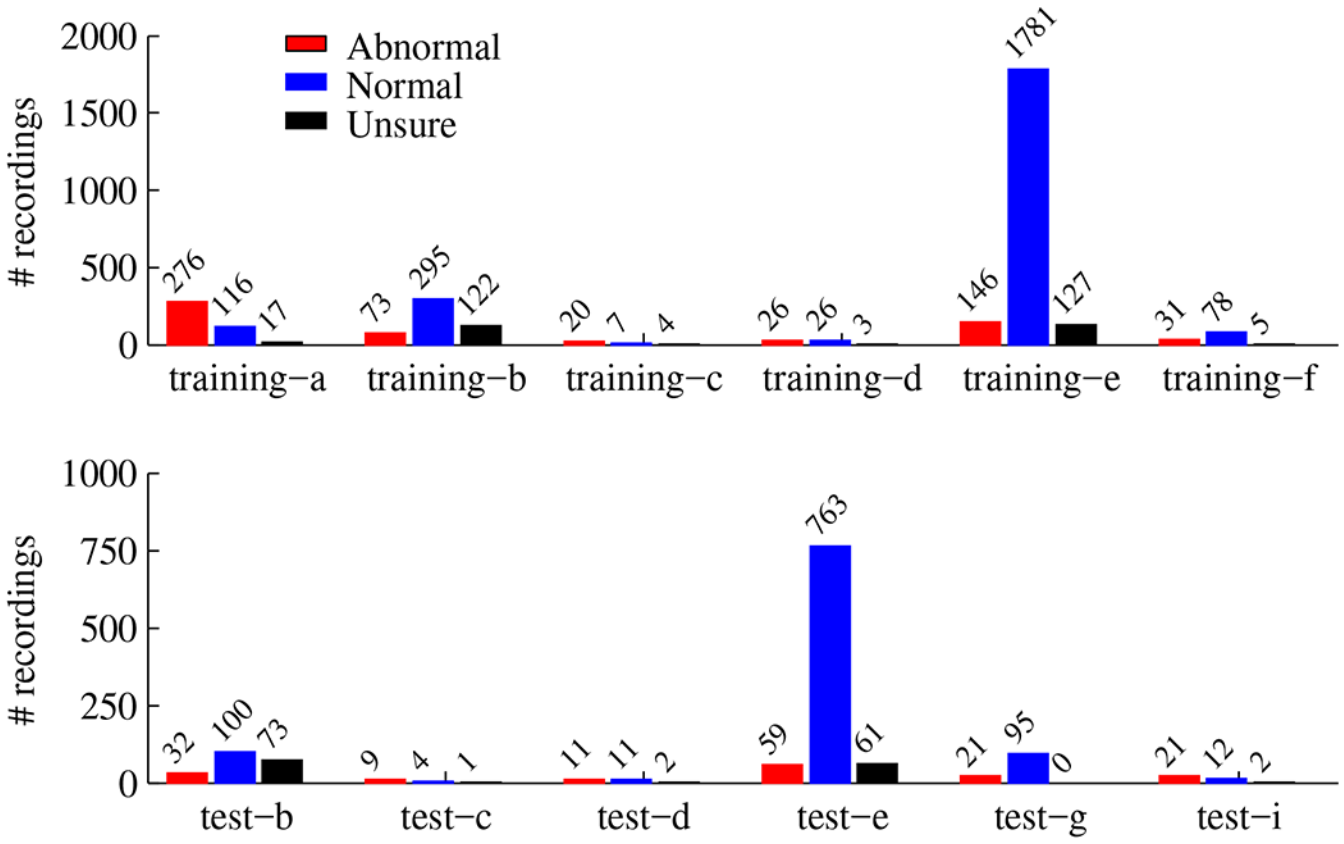

Figure 1:

Unbalanced data distribution for both training and test sets. Please note that the training and test databases with the same letter are related and are from the same data contributor, such as training-b and test-b.

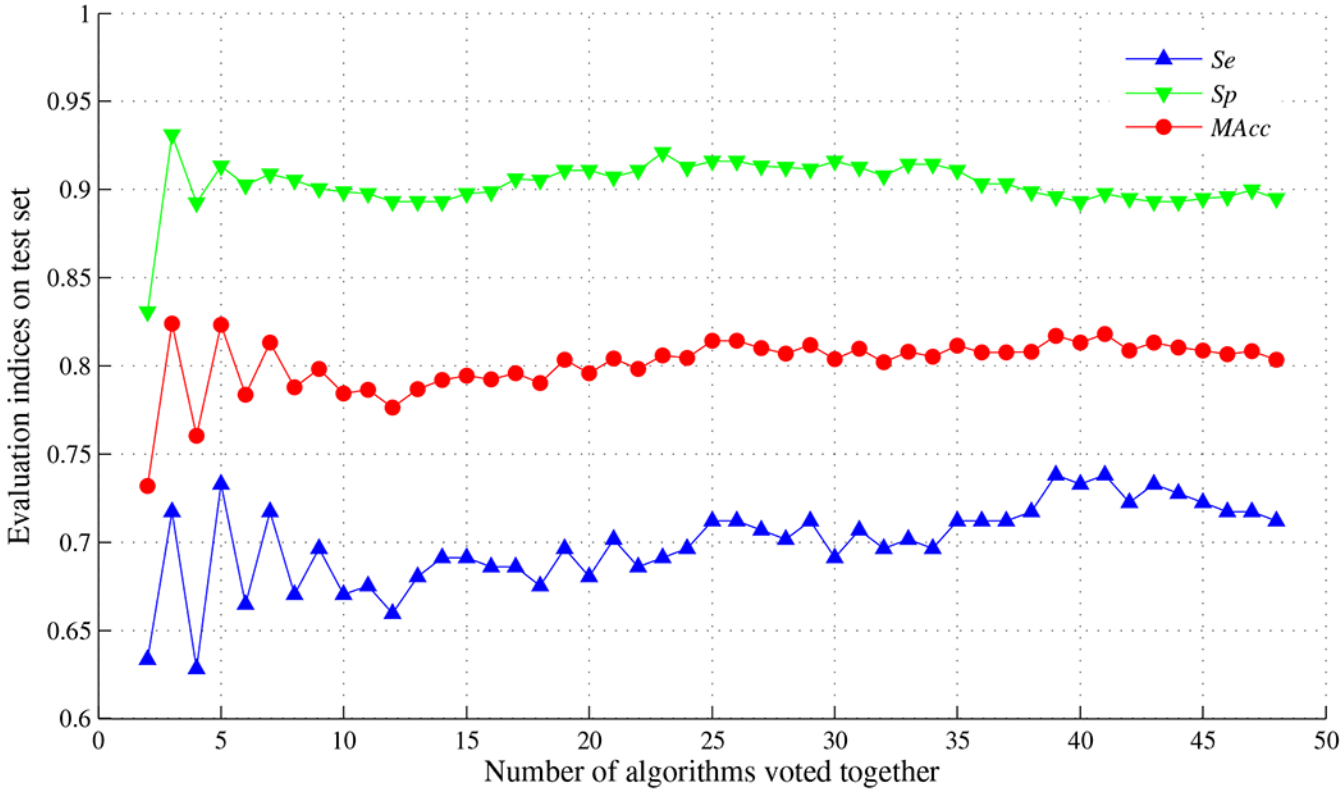

Figure 2:

Performance of voting algorithms as a function of number of algorithms. Algorithms were chosen by ranking them in descending order of score on the randomly selected 600 training recordings, and the test data score was reported (to prevent overestimation of the score).

Table 1:

Final scores for the top 20 of 48 official entrants, the example algorithm provided and a simple voting approach. Best performances of Challenge entrants are underlined. MFCC = mel-frequency cepstral coefficients. DTW = dynamic time warping. PCA = principal component analysis. FDA = fisher discriminant analysis. NN = neural network. LR = logistic regression. SVM = support vector machine. RF = random forest. ELM = extreme learning machine. CNN = convolutional NN. RNN = recurrent

Rank	Entrant	*Se*	*Sp*	*MAcc*	Segment	Feature method	# features	Feature selection	# selected features	Classifier	Training data division	Balancing data
1	Potes et al. (2016)	<u>0.9424</u>	0.7781	<u>0.8602</u>	Yes	Time-frequency	124	No	124	AdaBoost & CNN	80%/20% train/ test	No
2	Zabihi and Rad (2016)	0.8691	0.8490	0.8590	No	Time, frequency and time-frequency	40	Yes (wrapper)	18	Ensemble of NNs	20-fold CV	Yes
3	Kay and Agarwa (2016)	0.8743	0.8297	0.8520	Yes	Wavelet, MFCC and complexity	675	Yes (PCA)	70	DropConnected NN	10-fold CV	No
4	Bobillo (2016)	0.8639	0.8269	0.8454	Yes	Time-frequency, MFCCs and wavelets	142×4 × 172*tensor*	Yes (fisher score)	1000:1 reduction	LR, SVM & KNN	10-fold CV	No
5	Homsi et al. (2016)	0.8848	0.8048	0.8448	Yes	Time, frequency, wavelet, statistical	131	No	131	Ensemble of classifiers	10-fold CV	No
6	Plesinger et al. (2016)	0.7696	0.9125	0.8411	Yes	Frequency, statistical	315	Yes (PROBAfind)	51	Probability assessment	No	No
7	Rubin et al. (2016)	0.7278	<u>0.9521</u>	0.8399	Yes	MFCC	13	Yes (unknown)	6	CNN	80%/20% train/ test	No
8	Abdollahpur et al. (2016)	0.7696	0.8831	0.8263	Yes	Time, time-frequency, perceptual	89	Yes (FDA)	unknown	NNs voting	No	No
9	Tang et al. (2016)	0.8220	0.8149	0.8185	Yes	Multi-domain features	324	No	324	BPNN	Varied train/test division	No
10	Tschannen et al. (2016)	0.8482	0.7762	0.8122	Yes	Deep CNN-based features	12,160	Yes (PCA)	400	SVM	5-fold CV	No
11	Nilanon et al. (2016)	0.7696	0.8527	0.8111	Yes	Spectrogram, MFCC	unknown	No	unknown	LR, SVM, RF and CNN	5-fold CV	No
12	Whitaker and Anderson (2016)	0.8429	0.7716	0.8073	Yes	Frequency, sparse coding	unknown	No	unknown	SVM	1000/2153 train/test	No
13	Yang and Hsieh (2016)	0.7749	0.8287	0.8018	No	Augmented features	unknown	No	unknown	RNN	1/5 data for CV	No
14	Yazdani et al. (2016)	0.7487	0.8508	0.7998	Yes	Heartbeat, tape-long	unknown	No	unknown	Ensemble of classifiers	10-fold CV	Yes
15	Banerjee et al. (2016)	0.8010	0.7901	0.7956	Yes	Time-frequency	88	Yes (MIC)	31/88	RF	5-fold CV	Yes

Rank	Entrant	*Se*	*Sp*	*MAcc*	Segment	Feature method	# features	Feature selection	# selected features	Classifier	Training data division	Balancing data
16	Singh-Miller and Singh-Miller (2016)	0.7382	0.8499	0.7941	No	Spectral	unknown	Yes	25	RF	10-fold CV	No
17	Ryu et al. (2016)	0.6663	0.8775	0.7869	Yes	CNN-based features	unknown	No	unknown	CNN	3126/300 train/test	No
18	Yang et al. (2016)	0.6649	0.9088	0.7869	Yes	Audio signal analysis	unknown	Yes (RFE)	unknown	SVM & ELM	10-fold CV	No
19	Bouril et al. (2016)	0.7330	0.8398	0.7864	Yes	Time, frequency	74	Yes (unknown)	unknown	SVM	No	No
20	Ortiz et al. (2016)	0.7853	0.7855	0.7854	Yes	Time, MFCC, DTW	unknown	No	unknown	SVM	Varied train/test division	No
–	Sample entry	0.6545	0.7569	0.7051	Yes	Time, amplitude	20	Yes (likelihood ratio)	7	LR	10-fold CV	Yes
–	Voting results (best)	0.7173	0.9309	0.8241	–	–	–	–	–	–	–	–

Physiol Meas. Author manuscript; available in PMC 2024 October 08.

Table 2:

Summary of the papers included in this special issue.

Work in this special issue	*Se*	*Sp*	*MAcc*	Segment	Feature method	# features	Feature selection	# selected features	Classifier	Training data division	Balancing data
Abdollahpur et al. (2017)	0.7696	0.8831	0.8263 *	Yes	Time, time-frequency, perceptual	90	Yes (FDA)	unknown	NNs voting	train/test division	No
Homsi and Warrick (2017)	0.7960	0.8060	0.8010	Yes	Time, frequency, wavelet, statistical	131	Yes	19/17	Ensemble of classifiers	10-fold CV	No
Kay and Agarwal (2017)	–	–	0.5810	Yes	Wavelet, MFCC, interbeat and complexity	675	Yes (PCA)	50	DropConnected NN	10-fold CV	Yes
Langley and Murray (2017)	0.5589	0.9633	0.7611 *	No	Spectral amplitude and wavelet entropy	unknown	No	unknown	Decision tree	CV	No
Maknickas and Maknickas (2017)	0.8063	0.8766	0.8415#	No	MFSC	N/A	No	N/A	Deep CNN	train/test division	Yes
Plesinger et al. (2017)	0.8900	0.8160	0.8550	Yes	Frequency, statistical	228	Yes (PROBAfind)	53	Probability assessment	No	No
Whitaker et al. (2017)	0.8010	0.8060	0.8030	Yes	Time, frequency, sparse coding	unknown	No	unknown	SVM	1000/2153 train/ test	No

MFCC = mel-frequency cepstral coefficients. MFSC = mel-frequency spectral coefficients. PCA = principal component analysis. FDA = fisher discriminant analysis. NN = neural network. SVM = support vector machine. CNN = convolutional NN. CV = cross-validation.

*indicates the paper presents the same results from the Challenge official entries,

#indicates the paper presents the same results from the Challenge unofficial entries,

□indicates the paper presents new results in this follow-up phase.