A benchmark of heart sound classification systems based on sparse decompositions

Roilhi F Ibarra-Hernández, Nancy Bertin, Miguel A Alonso-Arévalo, Hugo A Guillén-Ramírez

A benchmark of heart sound classification systems based on sparse decompositions

Roilhi F. Ibarra-Hernández[a], Nancy Bertin[b], Miguel A. Alonso-Arévalo[a], and Hugo A. Guillén-Ramírez[a]

[a]Ensenada Center for Scientific Research and Higher Education (CICESE), Ensenada, B.C., Mexico
[b]Univ Rennes, Inria, CNRS, IRISA, Rennes, France

ABSTRACT

Background: Nowadays, cardiovascular diseases (CVD) remain the main cause of death worldwide. A heart sound signal or phonocardiogram (PCG) is the most simple, economical and non-invasive tool to detect CVDs. Advances in technology and signal processing allow the design of computer-aided systems for heart illnesses detection from PCG signals.
Purpose: The paper proposes a pipeline and benchmark for binary heart sounds classification. The features extraction architecture is focused on the use of Matching Pursuit time-frequency decomposition using Gabor dictionaries and the Linear Predictive Coding method of a residual. We compare seven classifiers with two different approaches: feature averaging and cycle averaging.
Methods: We test our proposal on the PhysioNet/CinC challenge 2016 database, which comprises a wide variety of heart sounds recorded from patients with normal and different pathological heart conditions. We conduct a 10-fold stratified cross-validation method to evaluate the performance of different classification algorithms. The feature sets were also tested when using an oversampling method for balancing.
Results: The benchmark identified systems showing a satisfying performance in terms of accuracy, sensitivity, and Matthews correlation coefficient. Results can be improved when using feature averaging and an oversampling strategy.

Keywords: Heart sounds, Phonocardiogram, Matching Pursuit, Linear Predictive Coding, Cross Validation test, Oversampling, SMOTE.

1. INTRODUCTION

Cardiovascular diseases (CVDs) are one of the main causes of death globally according to the World Health Organization.[1] A valuable method for CVDs detection is auscultation, *i.e.* listening to the mechanical valvular activity. When recorded, the resulting sound sequence is called a phonocardiogram (PCG). Auscultation is in fact a primary diagnosis method for initial detection of heart valves malfunctioning. It is the most economical and simple screening test, since the only device needed is a stethoscope. However, this technique has restrictions, since high quality recordings are required for reliable interpretations. Another constraint is the amount of experience and training to master auscultation skills for physicians, who are few in quantity and often located in urban areas. Nowadays, with the recent advances in signal processing and electronic stethoscope technologies, the design of automatic classification schemes from PCG recordings appears as a promising diagnosis method for CVDs detection. This paper aims at defining end-to-end pipelines for automatic classification of heart sounds as *healthy* or *pathological*, and at benchmarking instantiations of this pipeline with several signal representations and classifiers, in order to assess their potential in clinical practice.

Further author information: (Send correspondence to R.F.Ibarra)
R.F.: E-mail: frajo@cicese.edu.mx, Telephone: +52 (1)646 193 16 33

1.1 Heart sound signals

In a healthy or normal state, a PCG recording comprises two main components called *fundamental heart sounds* (FHS) and denoted S_1 and S_2 . Each FHS can be characterized by a common time length and energy concentration over low frequency regions. For instance the S_1 components are dominant in the region of 10 Hz to 140 Hz while S_2 components usually concentrate their energy around the 10 Hz to 200 Hz band.[2] *Murmurs* are sounds stemming from a turbulent blood flow due to valve malfunction, hence denoting a pathological or abnormal state. The energy distribution of murmurs in frequency vary widely and depending on its nature they can be as high as 600 Hz. There exists other sounds called S_3 and S_4 which can represent a pathological state in adults, but are physiological in children. The frequency content of these sounds may overlap with the S_1 and S_2 energy distribution. Figure 1 illustrates the waveform and the time-frequency content representative of a PCG cardiac cycle (union of S_1 and S_2 sounds) in both normal and pathological states.

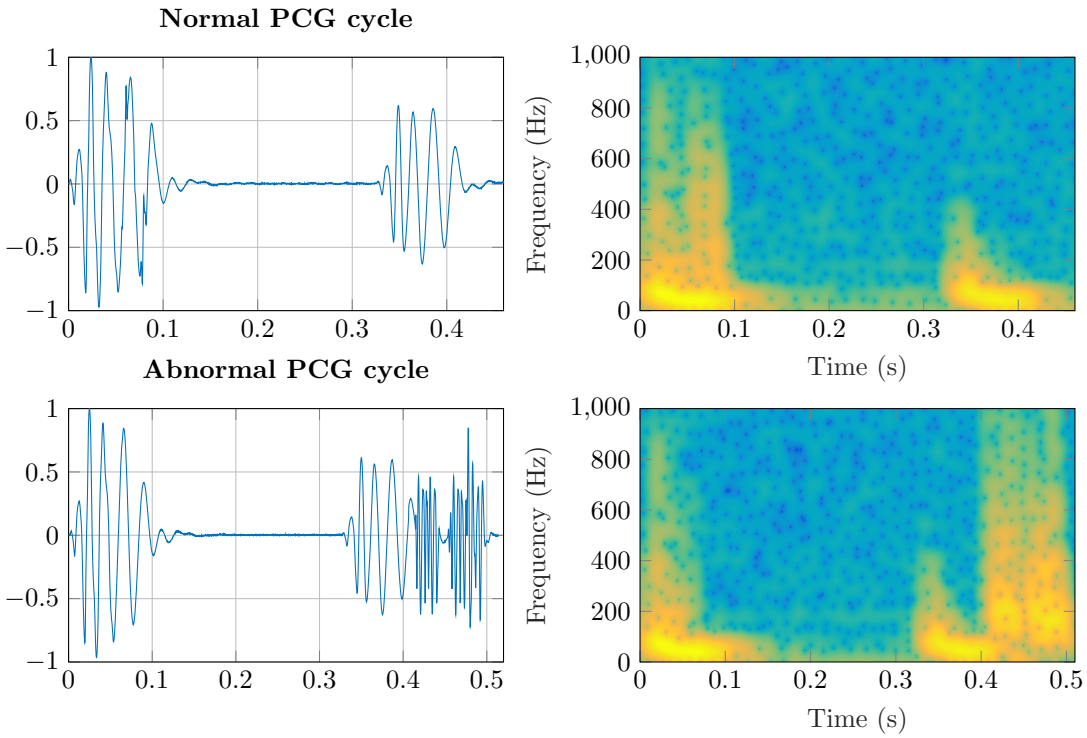

Figure 1. The time waveform and spectrogram representative of a PCG cardiac cycle in normal (top) and abnormal (bottom) conditions.

Murmur detection is a key task in the primary diagnosis of pathological states from a PCG recording. However, murmurs are highly nonstationary signals. In addition, PCGs are often corrupted by noises, such as those arising from speech or stethoscope movements. As a result, automatic heart sounds classification is a promising, but challenging task.

1.2 Prior art

The 2016 PhysioNet/Computing in Cardiology Challenge (CinC) was open in 2016 to encourage researchers to design and train classification methods for the binary labeling of PCG sounds as in *normal* or *abnormal* conditions.[3] It provides the largest public database of PCG recordings registered in a real clinical environment.[4] As for any classification task, each approach submitted to the challenge can be mainly characterized by two components: the chosen signal representation (features) and the type of classifier. During the CinC 2016 conference, most of the participants have used time-frequency features, in particular the wavelet decomposition coefficients

and Mel-Cepstral Frequency Coefficients (MFCCs).[5–7] Linear predictive coding coefficients have been also used.[7] In terms of classifiers, competitors mostly used Neural Networks,[5,8,9] Support Vector Machines[10,11] and Random Forest techniques.[12–14] In parallel, sparse representations of the heart sounds (Matching Pursuit decomposition and Linear Predictive Coding of the residual) have been shown to provide a compact and meaningful representation of PCG signals.[15] To our best knowledge, the usefulness of such a representation in a heart sounds classification task has not been assessed so far.

1.3 Contributions

This paper proposes a unified but modular pipeline for the development and the assessment of heart sounds classification systems based on sparse representations. A global overview of the pipeline and details on all its modules are given in Sec. 2. We provide a systematic evaluation of the derived systems as well, with a particular emphasis on the comparison between various design choices: cycle averaging vs. feature averaging, optional use of data oversampling for unbalanced classes compensation, and a wide range of classifiers. Experimental results, presented in Sec. 4 and discussed in Sec. 5, are obtained from a 10-fold Cross Validation test outputting accuracy, sensitivity and specificity scores, and Matthews Correlation Coefficients. This exhaustive evaluation allows us to conclude in Sec. 6 on the promising outperformance of a system combining cycle averaging, sparse representation, LPC residual coding and a Random Forest classifier, even in adverse conditions such as unbalanced classes and noise-corrupted recordings.

2. METHODS

Fig. 2 gives an overview of the proposed pipeline for heart sounds classification. Features extraction step gives two different sets of features: **A** and **B** (see Fig.3 for details). To address the probable case where the data would be unbalanced (more recordings labeled as *normal* than as *abnormal* in the training set), the Synthetic minority oversampling technique (SMOTE)[16] can be employed. After the SMOTE procedure, a 10-fold cross validation (CV) test is performed for different classifiers (summarized in Tab.1.) As described in Fig. 2, each fold operates on a different training/test data splitting (one tenth is used for test and the remaining for training).

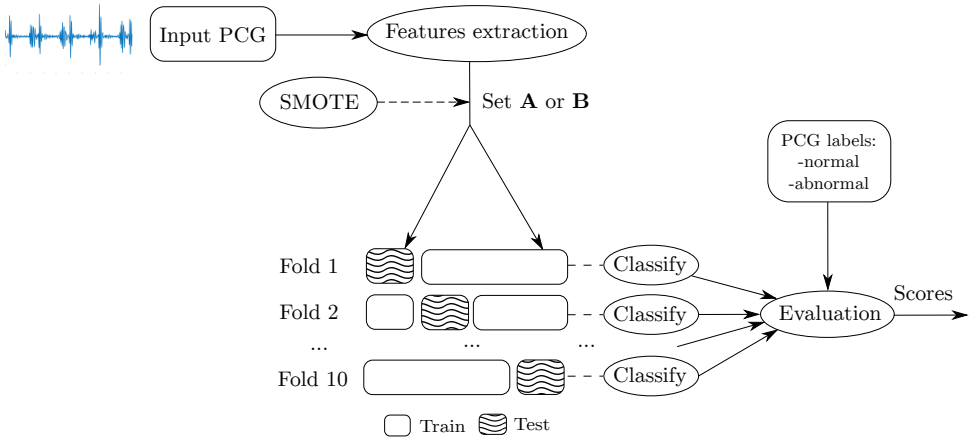

Figure 2. Overview of the proposed pipeline.

2.1 Features extraction

This paper compares two design choices for features extraction in order to represent each PCG: feature averaging and cycle averaging. Each approach gives as output two feature sets: **A** and **B**. Both methods share the following steps:

- **Cycle segmentation**: It consists in the time delineation of heart sound cycles. For this purpose we took the corrected annotations from Physionet/CinC challenge database.[3] Original annotations came as an output from Springer's method[17] and then corrected by the sponsors.

- **Matching pursuit (MP) decomposition**: It performs the time-frequency representation of a cycle or a FHS.[18] Features from MP are given by the parameters of the selected atoms from MP algorithm when using a Gabor dictionary (see Sec. 2.1.2).
- **Linear predictive coding (LPC)**: Due to its greedy nature, MP does not provide a perfect reconstruction of the signal. The difference between the MP reconstructed signal and the original PCG is called the residual. LPC models this residual in a compact form (see Sec. 2.1.3).

Fig. 3 describes the methods developed in this work to extract the feature sets **A** and **B**. The first method is shown on the left diagram and depicts the *feature averaging* approach. The output feature set **A** is actually obtained from the mean value of the MP and LPC parameters for the N segmented PCG cycles. The diagram on the right depicts a second method which uses a *cycle averaging* approach. MP and LPC parameters are extracted from an averaged cycle. By contrast with the first method (in which parameters are directly extracted from the whole cycle), in this second method, the averaged cycle is split into the two FHS. MP is then applied on each FHS separately, and two frequency parameters $F1$ and $F2$ are kept for each of them, as in a previous methodology.[19,20]

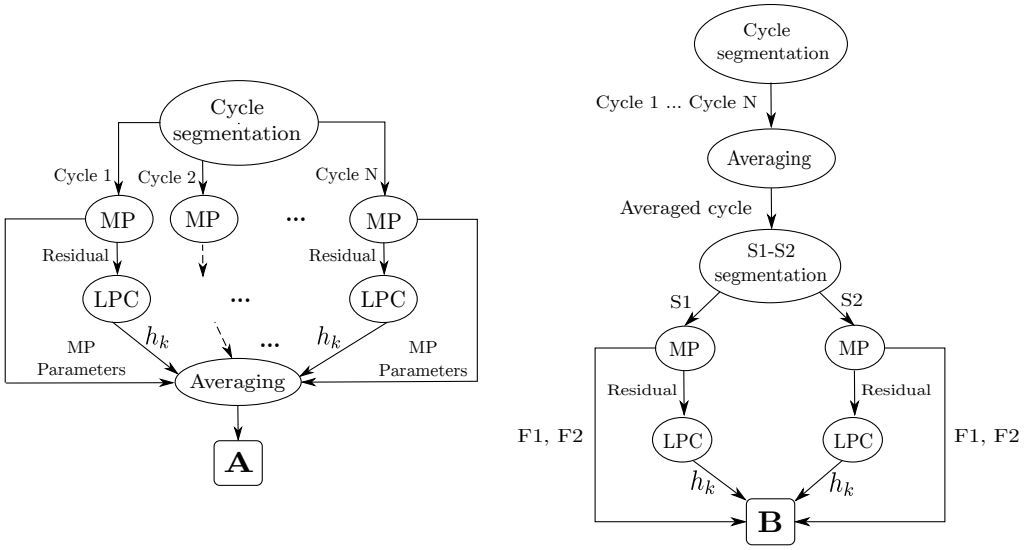

Figure 3. Feature extraction methods developed for this work. Left: the feature averaging approach (set **A** as output), right: the cycle averaging approach (set **B** as output).

2.1.1 Preprocessing

From observations of heart sounds pointed on 1.1, we introduce a band-pass filter between 25 Hz and 600 Hz to remove non informative frequencies in the input PCG signal. The original sampling rate of 4 kHz of recordings is preserved and the filter implemented is a sixth-order Butterworth filter.

2.1.2 Matching Pursuit (MP)

The Matching Pursuit method is a greedy and iterative algorithm which aims to find a linear combination of D elementary waveforms called *atoms* which approximates a signal $\mathbf{x}$ with minimal error. Each atom $\mathbf{g}_d$ is a elementary signal belonging to a redundant set of all possible predefined atoms called dictionary $\mathcal{D}$. MP iteratively selects atoms $\widehat{\mathbf{g}}_d$ to provide a decomposition of the form:

$$\mathbf{x} = \sum_{d=1}^{D} \alpha_d \cdot \widehat{\mathbf{g}}_d + \mathbf{r}, \tag{1}$$

where the coefficient α_d is a scalar weighting factor and $\mathbf{r}$ is a *residual* term. When using a dictionary of Gabor functions, the MP decomposition derives an adaptive time-frequency transform,[18] since this dictionary is composed by *well concentrated* waveforms in the time-frequency plane. In this work we use a pre-defined set of multiscale functions which is a collection $\mathcal{D} = \cup_{j=1}^{J} \mathcal{D}_j$ of *blocks* $\mathcal{D}_j$ of time-frequency atoms at different scales. The waveform of a Gabor atom in a multiscale dictionary is defined by the modulation, dilation, translation and sampling of a (continuous) window $w_j(t)$ as:

$$\mathbf{g}_{j,n,k}(m) = w_j(mT_s - nT_j) \exp\left(\frac{2i\pi k m T_s}{K_j}\right) \text{ for } 1 \leq m \leq M, \tag{2}$$

where the time location or window shift is defined as nT_j, the window length or scale L_j and is modulated at a frequency k/K_j, K_j is a predefined number of possible frequencies (FFT size), T_s is the sampling period and M the number of time samples. Fig. 4 illustrates the waveform of a Gabor atom, which is actually a cosine modulated Gaussian window. Other waveforms are shown at the right to see the effect of changing the modulation frequency.

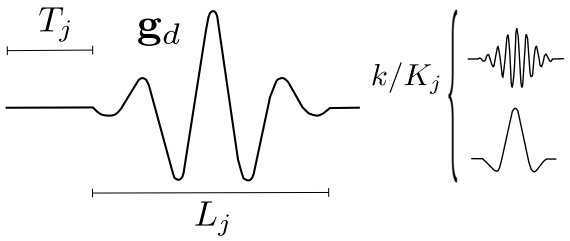

Figure 4. Time waveform of a Gabor atom and its defined parameters. The right panel illustrates the resulting waveforms when frequency parameter k/K_j varies.

2.1.3 Linear Predictive Coding (LPC)

As seen in Eq.1, after MP, the initial signal is decomposed into two main parts, a linear combination of Gabor atoms and a residual. The residual term $\mathbf{r}$ is expected to be rather uncorrelated with atoms, and thus has to be represented differently to be integrated in the feature sets. Instead of representing the temporal waveform, LPC tries to fit the spectrum of the signal. A number of works have used the LPC parameters[7,21] as features for PCG's classification. Let h_k be the set of LPC coefficients. They define a filter called *predictor*. The principle is that the residual $\mathbf{r}$ can be predicted as a linear combination of the previous samples:

$$r_n = -\sum_{i=1}^{p} h_i r_{n-i} + e_n, \tag{3}$$

where $n = 0, \cdots, N-1$ and e_n is the final residual. Filter coefficients h_i are added to the feature set.

3. EXPERIMENTAL SETTINGS

3.1 Heart sounds database

The set of PCG recordings from the Physionet/Cinc Challenge 2016[4] was employed to conduct the experiments in this work. This set is available online and contains 3,153 recordings from 764 subjects as training set. It has been assembled from six different subsets comprised by PCG sounds recorded under different conditions. Each recording has been labeled as normal or abnormal according to the presence or not of a murmur. Another given label is the quality of the recording as *good* or *bad*. However, this dataset is unbalanced since the class distribution of the signals is uneven: 665 recordings were marked as abnormal while 2,488 as normal.

3.2 Feature extraction settings

The MP decomposition of PCG signals was configured in the Matching Pursuit Toolkit library (MPTK) designed for MATLAB.[22] We used a Gabor dictionary of $J = 5$ blocks, each block corresponding to a common atom lengths L_j of 32, 64, 128, 256 and 512 samples. The selected number of atoms was $M = 15$ in order to reach almost a 99% of the energy to reconstruct a PCG cycle.[15] The atom parameters of frequency, amplitude, length, position (shift) and phase were extracted to be considered as features in features set **A** while in features set **B** just two atom frequencies were considered.

For the LPC representation of the residual signal output by MP, we used the MATLAB code from UCLA available on-line.[23] The selected order of the filter was $p = 15$.

Both feature sets **A** and **B** do not have the same number of samples. Combining the MP and LPC methods, features set **A** consists on $N_{features} = 90$ resulting from 75 parameters from MP (15 atoms with 5 parameters each one) and 15 from LPC. This choice is based on our previous work.[15] On the other hand, features set **B** contains $N_{features} = 19$, among which 4 are provided by MP (frequencies of the first two selected atoms for each FHS). This setting, in addition to the rest of set **B** extraction scheme (FHS segmentation and feature averaging), is in line with state-of-the-art.[19,20] Thus, set **B** acts as a comparison baseline for set **A**, whose performance in heart sound classification has not been assessed so far.

3.3 Classifiers settings

We tested the classification state of our proposed pipeline using seven different methods. Table 1 presents a brief description of each one by its name, acronym and the main parameters employed. The classifiers were configured according to the values depicted in Table using the sci-kit learn toolbox of Python.[24] In addition, feature sets were normalized to have zero mean and unit variance when using SVM. For the RF method we changed the number of estimators to 100 as some authors recommend in presence of unbalanced problems.[25,26]

3.4 Noisy recordings

Some of the recordings provided by Physionet database do not contain any annotation file to perform segmentation, due to their noise level*. To handle these files, we performed the processing of the signals in frames with time lengths of 900 ms (with respect to the approximate time duration of a cycle) and 200 ms (according to the typical length of a FHS.)

3.5 Classification performance

In order to measure the classifiers performance, the following metrics were calculated: accuracy (ACC), Sensitivity (SE), Specificity (SP) and Matthews Correlation Coefficient (MCC), defined as:

$$ACC = \frac{TP + TN}{TP + FN + TN + FP} \tag{4}$$

$$SE = \frac{TP}{TP + FN} \times 100 \tag{5}$$

$$SP = \frac{TN}{TN + FP} \times 100 \tag{6}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TN + FP)(TP + FN)(TN + FP)}}, \tag{7}$$

where the values TP, TN, FP and FN corresponds to the number of true positives, true negatives, false negatives and false positives respectively†.

*These were found as *low quality*, actually.

†In our case a true positive (TP) corresponds to a PCG entry with an abnormal condition correctly predicted.

Table 1. Brief description of the tested classifiers in this work. Parameters tunning and references for details.

Acronym	Full name	Main parameters	Value of parameters	Reference
CART	Classification and regression tree	Criterion to measure the quality of each split, spitter, strategy to choose the split at each node	Gini criterion, best split between trees.	Breiman et al[27]
KNN	K-nearest neighbors	Number of neighbors, weight function used in prediction, metric of distance to use for the tree	5 neighbors, uniform weight, Minkowski's distance.	Wenberger et al[28]
LDA	Linear discriminant analysis	Name of solver to reduce data dimensions, tolerance value threshold for rank estimation in the solver	Solver: SVD, tolerance of 0.0001	Friedman et al[29]
LR	Logistic regression	Penalty norm, number of iterations	L2 norm penalty, 100 iterations	Schmidt et al[30]
NB	Naive Bayes	Prior probabilities in the classes to adjust the data	No prior probabilities value	Chan et al[31]
RF	Random forest	Number of estimators, criterion to measure the quality of the split, use or not bootstrap to build trees.	100 estimators, Gini criterion, bootstrap: true	Breiman et al[32]
SVM	Support vector machines	Penalty parameter C of the error term, kernel function used in the algorithm, kernel coefficient gamma	$C = 2.37$ for unbalanced data and $C = 1$ for balanced, radial basis function as kernel with automatic gamma parameter	Cortes et al[33]

4. RESULTS

4.1 Classifiers performance test without class balancing

We conducted in sci-kit learn a 10-Fold CV approach. For the first test, data is plural. Table 2 shows the SE, SP, ACC and MCC average and standard deviation output metrics. In terms of SP, the SVM method reached the highest score of 92.29% when using features set **A**. However, the SE score reached by this combination is the lowest, even the MCC is not the highest presented. The RF method reached the highest values for the remaining metrics when using features set **A**: $SE = 75.80$, $MCC = 0.55$ and $ACC = 0.86$. Results from Table 2 present also greater standard deviation values for SE. We observe that for all feature sets and all classifiers, SP is considerably higher than SP.

4.2 Classifiers performance test with class balancing

For a second experiment, we perform a balancing applying the SMOTE technique when oversampling the minority class. The SMOTE library included in the imbalanced-learn toolbox of Python[34] was used for this purpose. Then, we conducted again the stratified 10-fold CV test to evaluate the classifiers performance. Table 3 shows the average metrics that outcomes from this experiment.

Table 2. Performance metrics resulting for our heart sounds cross validation test without balancing.

Model		Average performance			
Classifier	Features set	SE	SP	ACC	MCC
CART	A	50.19 ± 6.33	87.30 ± 1.91	0.79 ± 0.03	0.38 ± 0.08
KNN	A	48.26 ± 4.85	86.36 ± 1.21	0.78 ± 0.02	0.35 ± 0.06
LDA	A	58.32 ± 6.23	83.36 ± 0.86	0.81 ± 0.01	0.31 ± 0.05
LR	A	58.44 ± 6.97	83.36 ± 0.95	0.81 ± 0.02	0.32 ± 0.06
NB	A	49.56 ± 10.41	80.84 ± 0.68	0.79 ± 0.02	0.18 ± 0.06
RF	A	**75.80 ± 5.32**	88.03 ± 1.32	**0.86 ± 0.02**	**0.55 ± 0.06**
SVM	A	46.41 ± 1.99	**92.29 ± 0.76**	0.76 ± 0.02	0.45 ± 0.03
CART	B	42.70 ± 3.75	**84.76 ± 1.02**	0.76 ± 0.02	0.28 ± 0.05
KNN	B	60.63 ± 3.82	84.54 ± 0.85	**0.82 ± 0.01**	**0.36 ± 0.04**
LDA	B	48.11 ± 10.78	80.65 ± 0.76	0.79 ± 0.01	0.17 ± 0.07
LR	B	54.31 ± 12.73	80.43 ± 0.88	0.79 ± 0.01	0.17 ± 0.08
NB	B	58.59 ± 12.43	80.93 ± 0.68	0.80 ± 0.01	0.21 ± 0.07
RF	B	**72.74 ± 8.88**	83.17 ± 1.11	0.82 ± 0.02	0.36 ± 0.08
SVM	B	34.95 ± 4.69	81.91 ± 0.90	0.73 ± 0.02	0.16 ± 0.05

Table 3. Performance metrics resulting for our heart sounds cross validation test when using SMOTE balancing.

Model		Average performance			
Classifier	Features set	SE	SP	ACC	MCC
CART	A	82.61 ± 1.60	84.98 ± 1.85	0.84 ± 0.02	0.68 ± 0.03
KNN	A	70.74 ± 1.39	**97.10 ± 1.63**	0.79 ± 0.02	0.62 ± 0.03
LDA	A	76.95 ± 1.27	82.37 ± 2.14	0.79 ± 0.01	0.59 ± 0.03
LR	A	78.60 ± 1.60	82.38 ± 2.50	0.80 ± 0.02	0.61 ± 0.04
NB	A	84.54 ± 4.58	53.29 ± 0.63	0.56 ± 0.01	0.21 ± 0.03
RF	A	**91.60 ± 1.77**	92.10 ± 1.70	**0.92 ± 0.01**	**0.84 ± 0.03**
SVM	A	77.20 ± 1.09	78.77 ± 1.99	0.78 ± 0.01	0.56 ± 0.03
CART	B	80.71 ± 2.43	82.91 ± 1.82	0.82 ± 0.02	0.64 ± 0.03
KNN	B	79.24 ± 1.56	**95.87 ± 0.79**	0.86 ± 0.01	0.73 ± 0.02
LDA	B	70.55 ± 1.52	70.89 ± 2.44	0.71 ± 0.02	0.41 ± 0.04
LR	B	69.90 ± 1.79	67.97 ± 1.94	0.69 ± 0.02	0.38 ± 0.04
NB	B	82.80 ± 5.09	52.78 ± 0.66	0.55 ± 0.01	0.19 ± 0.03
RF	B	**89.10 ± 1.46**	91.55 ± 1.40	**0.90 ± 0.01**	**0.81 ± 0.02**
SVM	B	67.18 ± 4.03	53.41 ± 0.96	0.56 ± 0.02	0.15 ± 0.04

Compared to the metrics obtained without applying SMOTE, the SE values showed an increment for all tested classifiers and feature sets. The highest sensitivity $SE = 91.60\%$ value was reached for the combination of the features set **A** and the RF method. This approach also reached the highest accuracy $ACC = 92\%$ and Matthews Correlation Coefficient $MCC = 0.84$. Although KNN classifier with the features set **B** presented the highest SP, the remaining scores were not the highest values. Fig. 5 plots all the ACC scores obtained. An increase in the mean values is shown, except for the NB method and SVM when using as input the features set **B**.

5. DISCUSSION AND FUTURE WORK

Overall, the outcomes of the proposed classification schemes indicate that a combination of MP time-frequency features and LPC features, oversampling, and state-of-the-art classifiers result together in a satisfying performance for heart sound classification.

The obtained performances can be indicatively related to those reported for the 2016 PhysioNet/CinC entrants.[3] In the original challenge, performance was assessed by modified versions of SE, SP, and their arithmetic mean,

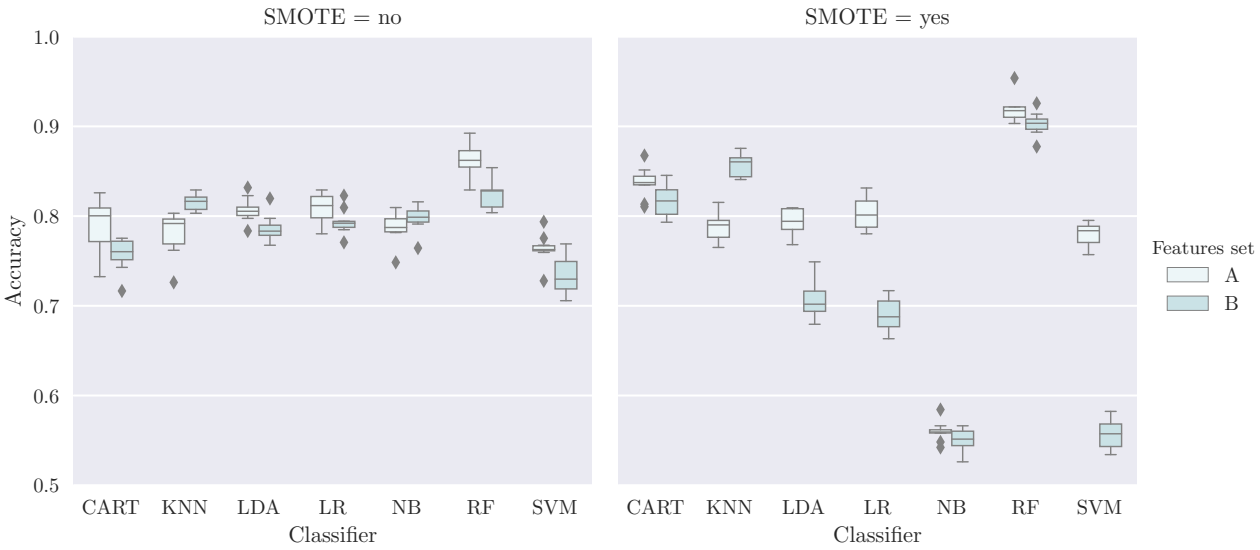

Figure 5. Accuracy scores for the experiments conducted. At the right chart data was oversampled with SMOTE while in the left is not.

denoted $Macc$[‡]. The highest ranked system,[9] based on a convolutional neural network, reached a mean accuracy $Macc = 86.02\%$ with $SE = 94.24\%$ and $SP = 77.81\%$. In comparison, in our experiments on similar data[§], the random forest classifier with feature set A and SMOTE oversampling reached a sensitivity $SE = 91.60\%$ and a specificity $SP = 92.10\%$, corresponding to a mean accuracy of $Macc = 91.85\%$. This suggests that some of the approaches presented here could be competitive with state-of-the-art systems submitted to the challenge.

Several novelties of the proposed approaches, emphasized below, contribute to their performance. Their analysis allows us to draw additional insights on heart sound classification strategies and to identify axes for future research.

5.1 Features

Several differences between feature sets **A** and **B**, and features used as input in state-of-the-art systems, can be stressed out:

- Compared with most of Physionet challenge entries (wavelets, time lengths of FHS and murmurs and MFCC coefficients mainly), the features extracted in both sets came from an MP-based approach using Gabor dictionaries. This proposal has been previously shown to provide with an efficient representation of PCG signals.[35,36]
- The LPC extracted features came from a representation of the output residual signal from the MP decomposition, instead of direct modeling of cardiac sounds waveform. In previous work we showed that this representation of the residual preserves some meaningful heart sound signal components.[37]
- We considered two main approaches to build and extract PCG signal feature sets: feature averaging and cycle averaging. Most of Physionet challenge submissions focused on the first one, while the second one had not been evaluated with these recordings so far.

[‡]It must be noted that metrics are slightly different from ours, calling for cautious interpretations. Contrary to the cited paper, we did not weight SP and SE according to the data labels as indicated for the Physionet/Cinc 2016 challenge. Little difference is expected from this. On the other hand, the average $Macc$ is different from the ACC criterion, which puts more importance on the SE/SP trade-off, and they cannot be directly compared.

[§]Our experiments were conducted exclusively on data included in the Physionet database. However, due to the unavailability of the two "hidden test sets", they are not strictly identical.

All else being equal, systems using the set **A** as input systematically outperformed their equivalent with the set **B**, although the number of features in **A** is greater. An additional test on cumulative feature importance for RF classification was performed (data not shown). This test revealed the importance of adding the LPC features: in set A, although they represent only about 17% of the input feature vector, 5 of them were ranked in the 10 most important features for the RF classifier. In addition, our results validate the new feature averaging approach against the previous cycle averaging approach.

However, as a 90-dimensional feature space remains demanding and possibly suboptimal (due to correlations between features), future work may include a feature selection step, to better deal with small training datasets and reduce the computational burden.

5.2 Oversampling

Unbalance between healthy and pathological classes is a typical situation in medical applications. It is known, in particular, to be very detrimental to sensitivity of classifiers, and to lead to uneven $SE - SP$ trade-offs.

Our results clearly confirm the dramatic impact of using an oversampling strategy to alleviate this problem. We observe an average improvement of 24.4 points in sensivity scores, and a global reduction of the variance on this indicator. In most cases, this is accompanied with a moderate variation of specificity (minor loss or improvement). Hence, the use of SMOTE brings the majority of our best systems to reach higher ACC and MCC scores, which denotes a more balanced trade-off between SE and SP, and a better global performance.

Exceptions are the NB classifiers (with both feature sets), SVM classifier (with set **A** only) and the LR and LDA classifiers (with set **B** only) which apparently don't benefit from oversampling, encountering an important drop in specificity. However, none of these classifiers ranked first in any of our experiments.

5.3 Classifiers and their tuning

As mentioned above, the RF classifier showed the higher metrics for both experiments, reaching a mean ACC of 92%, $SE = 91.60\%$, $SP = 92.10\%$ and $MCC = 0.84$ when using the feature set **A** and oversampling. Without oversampling, RF still ranks first with set **A** and first ex-aequo with set **B** (together with KNN.)

The achieved $SE - SP$ trade-off varies widely between classifiers. The highest specificity is achieved by KNN with oversampling (97.1% with set A and 95.87% with set B), at the price of a poor sensitivity. This may be linked to the relatively low number of considered neighbors (5) with respect to the feature set size (90). As sensitivity is concerned, RF ranks first again. This plays a large role in its global score, since sensitivity is obviously damaged by unbalanced classes for all classifiers, and the criterion that most of the tested classifiers struggle to fulfill.

It should be noted that all performance scores, but more particularly the $SE - SP$ trade-off, is known to be also largely ruled by the classifier's parameters. In our experiments, their values were chosen from default values and literature, and were not tuned nor optimized. Further work would be needed to explore their impact on performance, defining tuning strategies and identifying the classifiers' various operational regimes.

This is all the more relevant than in clinical practice, the smallest $SE - SP$ difference is not always seeked for. As noticed in the original challenge,[3] as false positives and false negatives have very different consequences on patients, physicians may favor one of the criteria over the other, depending on the clinical situation, other existing complementary tests and costs. As PCG classification is meant to act as a first stage screening test, one could expect that a better sensitivity would be preferred, even at the cost of a loss in specificity. This choice, which can be informed by benchmarks like the present one, however belongs to physicians and public health policy experts.

6. CONCLUSIONS

This paper proposes a methodical benchmark of several phonocardiogram binary classification systems, aiming at detecting pathological heart states. Two sets of time-frequency features were extracted from each heart sound, based on the MP decomposition and the LPC coding technique.[15] Seven common state-of-art classification methods were then tested in a 10 fold stratified cross validation technique. An oversampling technique, SMOTE,

was also added in order to compensate for unbalanced classes (less recordings with an abnormal condition, which is typical in biological data), and its impact on classifiers performance was assessed. Performance was measured through sensitivity, specificity, accuracy and Matthews correlation coefficient.

Among all conducted experiments, the Random Forest classifier combined with SMOTE technique outperformed all other configurations, reaching a competitive score for accuracy (92%), with a good sensitivity-specificity trade-off. Detailed insights from the experiments also validate the newest of the two used feature sets in the considered classification task. Its originality mainly relies on three novelties: an alternative strategy for feature extraction from a long recording (feature averaging) to the more common heart cycle averaging strategy; it does not require a segmentation of the fundamental heart sounds S1 and S2 within one heart cycle; the LPC coefficients that it includes are not computed on the whole signal, but only on the residual signal left after a sparse decomposition of the PCG in a Gabor dictionary.

As the identified scheme shows promising results for detecting abnormal conditions in PCG signals, future work will target performance and computational efficiency improvements through a refined preprocessing (denoising) and dimensionality reduction approaches (feature selection). Although excluded from our benchmark due to their high training data requirements, neural network based classifiers cannot be ignored nowadays and should also be tested against the proposed feature set.

ACKNOWLEDGMENTS

The first author would like to express his gratitude to the Rennes Métropole department, which funded its stay at Inria Rennes, France research center to support this work. M. Ibarra is a PhD student supported by the Mexican National Council for Science and Technology (CONACYT) through the Graduate Research Fellowship No. 477876. Authors also want to thank to the sponsors of the Physionet/Cinc 2016 Challenge for providing the largest open database of heart sounds to the scientific community.

REFERENCES

[1] WHO, "Cardiovascular diseases, fact sheet." World Health Organization: http://www.who.int/mediacentre/factsheets/fs317/en/ (2017). Accessed: 2018-05-02.

[2] Abbas, A. K. and Bassam, R., "Phonocardiography signal processing," *Synthesis Lectures on Biomedical Engineering* **4**(1), 1–194 (2009).

[3] Clifford, G. D., Liu, C., Moody, B., Springer, D., Silva, I., Li, Q., and Mark, R. G., "Classification of normal/abnormal heart sound recordings: The physionet/computing in cardiology challenge 2016," in [*Computing in Cardiology Conference (CinC), 2016*], 609–612, IEEE (2016).

[4] Liu, C., Springer, D., Li, Q., Moody, B., Juan, R. A., Chorro, F. J., Castells, F., Roig, J. M., Silva, I., Johnson, A. E., et al., "An open access database for the evaluation of heart sound algorithms," *Physiological Measurement* **37**(12), 2181 (2016).

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

[6] Munia, T. T., Tavakolian, K., Verma, A. K., Zakeri, V., Khosrow-Khavar, F., Fazel-Rezai, R., and Akhbardeh, A., "Heart sound classification from wavelet decomposed signal using morphological and statistical features," in [*Computing in Cardiology Conference (CinC), 2016*], 597–600, IEEE (2016).

[7] Zabihi, M., Rad, A. B., Kiranyaz, S., Gabbouj, M., and Katsaggelos, A. K., "Heart sound anomaly and quality detection using ensemble of neural networks without segmentation," in [*Computing in Cardiology Conference (CinC), 2016*], 613–616, IEEE (2016).

[8] Kay, E. and Agarwal, A., "Dropconnected neural network trained with diverse features for classifying heart sounds," in [*Computing in Cardiology Conference (CinC), 2016*], 617–620, IEEE (2016).

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

[10] Whitaker, B. M., Suresha, P. B., Liu, C., Clifford, G. D., and Anderson, D. V., "Combining sparse coding and time-domain features for heart sound classification," *Physiological measurement* **38**(8), 1701 (2017).

[11] Bobillo, I. J. D., "A tensor approach to heart sound classification," in [*Computing in Cardiology Conference (CinC), 2016*], 629–632, IEEE (2016).

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

[13] Antink, C. H., Becker, J., Leonhardt, S., and Walter, M., "Nonnegative matrix factorization and random forest for classification of heart sound recordings in the spectral domain," in [*Computing in Cardiology Conference (CinC), 2016*], 809–812, IEEE (2016).

[14] Singh-Miller, N. E. and Singh-Miller, N., "Using spectral acoustic features to identify abnormal heart sounds," in [*Computing in Cardiology Conference (CinC), 2016*], 557–560, IEEE (2016).

[15] Ibarra, R. F., Alonso, M. A., Villarreal, S., and Nieblas, C. I., "A parametric model for heart sounds," in [*Signals, Systems and Computers, 2015 49th Asilomar Conference on*], 765–769, IEEE (2015).

[16] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P., "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research* **16**, 321–357 (2002).

[17] Springer, D. B., Tarassenko, L., and Clifford, G. D., "Logistic regression-hsmm-based heart sound segmentation," *IEEE Transactions on Biomedical Engineering* **63**(4), 822–832 (2016).

[18] Mallat, S. G. and Zhang, Z., "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on signal processing* **41**(12), 3397–3415 (1993).

[19] Wang, W., Guo, Z., Yang, J., Zhang, Y., Durand, L.-G., and Loew, M., "Analysis of the first heart sound using the matching pursuit method," *Medical and Biological Engineering and Computing* **39**(6), 644–648 (2001).

[20] Wang, W., Pan, J., and Lian, H., "Decomposition and analysis of the second heart sound based on the matching pursuit method," in [*Signal Processing, 2004. Proceedings. ICSP'04. 2004 7th International Conference on*], **3**, 2229–2232, IEEE (2004).

[21] Redlarski, G., Gradolewski, D., and Palkowski, A., "A system for heart sounds classification," *PloS one* **9**(11), e112673 (2014).

[22] Krstulovic, S. and Gribonval, R., "MPTK: Matching pursuit made tractable," in [*Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*], **3**, III–III, IEEE (2006).

[23] Ozun, O., Steurer, P., and Thell, D., "Wideband speech coding with linear predictive coding (lpc)." UCLA Electrical Engineering Digital Speech Processing: http://www.seas.ucla.edu/spapl/projects/ee214aW2002/1/report.html (2017). Accessed: 2018-05-17.

[24] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[25] Vercio, L. L., Del Fresno, M., and Larrabide, I., "Detection of morphological structures for vessel wall segmentation in ivus using random forests," in [*12th International Symposium on Medical Information Processing and Analysis*], **10160**, 1016012, International Society for Optics and Photonics (2017).

[26] Mellor, A., Boukir, S., Haywood, A., and Jones, S., "Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin," *ISPRS Journal of Photogrammetry and Remote Sensing* **105**, 155–168 (2015).

[27] Breiman, L., Friedman, J., Olshen, R., and Stone, C., "Classification and decision trees," *Wadsworth, Belmont* **378** (1984).

[28] Weinberger, K. Q. and Saul, L. K., "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research* **10**(Feb), 207–244 (2009).

[29] Friedman, J., Hastie, T., and Tibshirani, R., [*The elements of statistical learning*], vol. 1, Springer series in statistics New York, NY, USA: (2001).

[30] Schmidt, M., Le Roux, N., and Bach, F., "Minimizing finite sums with the stochastic average gradient," *Mathematical Programming* **162**(1-2), 83–112 (2017).

[31] Chan, T. F., Golub, G. H., and LeVeque, R. J., "Updating formulae and a pairwise algorithm for computing sample variances," in [*COMPSTAT 1982 5th Symposium held at Toulouse 1982*], 30–41, Springer (1982).

[32] Breiman, L., "Random forests," *Machine learning* **45**(1), 5–32 (2001).

[33] Cortes, C. and Vapnik, V., "Support vector machine," *Machine learning* **20**(3), 273–297 (1995).

[34] Lemaître, G., Nogueira, F., and Aridas, C. K., "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research* **18**(17), 1–5 (2017).

[35] Zhang, X., Durand, L., Senhadji, L., Lee, H. C., and Coatrieux, J.-L., "Analysis-synthesis of the phonocardiogram based on the matching pursuit method," *IEEE Transactions on Biomedical Engineering* **45**(8), 962–971 (1998).

[36] Sava, H., Pibarot, P., and Durand, L.-G., "Application of the matching pursuit method for structural decomposition and averaging of phonocardiographic signals," *Medical and Biological Engineering and Computing* **36**(3), 302–308 (1998).

[37] Ibarra-Hernández, R. F., Alonso-Arévalo, M. A., Cruz-Gutiérrez, A., Licona-Chávez, A. L., and Villarreal-Reyes, S., "Design and evaluation of a parametric model for cardiac sounds," *Computers in biology and medicine* **89**, 170–180 (2017).

Answers to reviewers questions

We want to thank the reviewers for the comments made to our manuscript, SIP300-14, *A benchmark of heart sound classification systems based on sparse decompositions*. We completed the submission successfully taking into account these valuable observations. In this section, we would like to ask a couple of questions formulated during the review.

Question: The authors mention that murmurs are highly nonstationary signals, but aren't S1 and S2 sounds also nonstationary signals?

Our answer: We consider that murmurs are, by nature, highly nonstationary because of their characteristic broad and abrupt frequency changes. On the other hand, S1 and S2 sounds are quasi-stationary due to their waveform shape and limited characteristic bandwidth. However, time location and delineation in these sounds vary according to physiological conditions from each patient.

Question: It could be desirable a discussion about the advantages of the classifiers analyzed over the more recent techniques as deep learning or recurrent neural networks.

Our answer: Deep neural networks usually require much more data than traditional Machine Learning algorithms, at least datasets containing hundreds of thousands if not millions of labeled samples. Unfortunately, we do not have access to such amount of annotated cardiac sounds. Training of deep learning or neural networks is not a trivial task and due to the amounts of data required the implementation of the methods can be hard to parallelize [1].

[1] Chen, X. W., & Lin, X. (2014). Big data deep learning: challenges and perspectives. *IEEE access, 2*, 514-525.