Biomedical Signal Processing and Control 57 (2020) 101788

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Classification of heart sounds using fractional fourier transform based mel-frequency spectral coefficients and traditional classifiers

Zaid Abduh [a,*], Ebrahim Ameen Nehary [b], Manal Abdel Wahed [a], Yasser M. Kadah [a,c]

[a] Biomedical Engineering Department, Faculty of Engineering, Cairo University, Giza 12613, Egypt
[b] Electrical and Computer Engineering Department, Concordia University, Montreal, Canada
[c] Electrical and Computer Engineering Department, King Abdulaziz University, Jeddah 21589, Saudi Arabia

ARTICLE INFO

Article history:
Received 22 June 2019
Received in revised form 26 October 2019
Accepted 16 November 2019

Keywords:
Phonocardiogram
Heart sounds
Computer-aided auscultation
Fractional fourier transform
Mel-frequency spectral coefficients
Spectral subtraction

ABSTRACT

Heart sounds are a rich source of information for early diagnosis of cardiac pathologies. Distinguishing normal from abnormal heart sounds requires a specially trained clinician. Our goal is to develop a machine learning application that tackle the problem of heart sound classification. So we present a new processing and classification system for heart sounds. The automated diagnostic system is described in terms of its preprocessing, cardiac cycle segmentation, feature extraction, features reduction and classification stages. Conventional architectures will be used to identify abnormal heart sounds then the performances of the proposed systems will be compared. The conventional architectures include the following traditional classifiers: SVM, KNN and ensemble classifier (bagged Trees, subspace KNN and RUSBoosted tree). The proposed system is verified on the publicly available dataset of the heart sounds. The cross-validation and local hold out train-test methods are used to perform the experiments and obtain and compare the results. The proposed system showed potential for achieving excellent performance compared to previous methods on the same dataset with a score of 0.9200 at a sensitivity of 0.8735 and specificity of 0.9666 using a support vector machine classifier with cubic kernel. The details of the methodology and the results are presented and discussed.

1. Introduction

Cardiovascular disease (CVD) still the first leading cause of death with an estimated 17.9 million people worldwide died from CVD-related conditions in 2016, representing 31 % of all global deaths [1]. Heart diseases can be diagnosed by several techniques involve medical imaging procedures, which are rather costly and cumbersome and hence of limited availability to many people. On the other hand, the simplest CVD diagnostic technique is heart sound auscultation, which is an old yet very effective diagnostic tool to check the condition of the heart. Early detection of abnormal heart sounds provides the needed time to prevent cardiovascular system disruption.

The phonocardiogram (PCG) is a graphical description of heart sound. The PCG signal contains helpful information that aids diagnose heart disease and assess the quality of cardiovascular system function [2]. In general, heart sounds are often difficult to interpret due to their low intensity and dominating frequencies near the lower limit of the human hearing range. So, auscultation needs a lot of training and experience.

Computer-aided auscultation is an automated diagnostic tool utilize the computer to help the physician in the initial inspection of abnormal heart sounds.

In the past few decades, many automated analysis algorithms were developed to assess patients based on the PCG alone without electrocardiogram (ECG) synchronization. Deep learning and conventional architectures classification models were utilized. Also, different types of features were implemented. This area has received a lot of research work and still.

Potes et al. [3] proposed an ensemble of a feature-based classifier and a deep learning based classifier to boost the classification performance of heart sounds. A total of 124 time-frequency features were extracted from PCG signal and input to AdaBoost ensemble classifier. PCG signals were decomposed into four frequency bands to train convolutional neural network (CNN). The final decision was based on combining the outputs of AdaBoost and the CNN. Rubin et al. [4] proposed an algorithm that transformed 1-D PCG signal into 2-D time-frequency heat map representations using Mel-frequency cepstral coefficients (MFCC). Convolutional neural network was used to automatically classify normal versus

* Corresponding author.
E-mail address: zaabduh@gmail.com (Z. Abduh).

https://doi.org/10.1016/j.bspc.2019.101788
1746-8094/

abnormal heart sound recordings. The PhysioNet/Computing in Cardiology Challenge 2016 presented a dataset to develop, test and compare several algorithms with conventional architectures and deep learning to classify heart sounds. Many research groups contributed new methods in response to this challenge [5–7]. Gokhale [8] introduced an algorithm that utilizes Hilbert envelope and wavelet features with boosted trees ensemble classifier. Goda et al. [9] used time-frequency domain features and support vector machine (SVM) classifier. An algorithm was proposed by Grzegorczyk et al. [10] to classify heart sound recording based on deep neural networks. Conventional neural network and auto-encoder deep neural network were used with forty-eight time and frequency domain features. Tschannen et al. [11] introduced another technique where deep features (generated by a wavelet-based convolutional neural network) and time-frequency domain features were both used with L_2-SVM classifier. Homsi et al. [12] presented an approach that use time, frequency, wavelet and statistical domain features with a nested set ensemble classifier that included random forest, LogitBoost and cost-sensitive classifiers. Langley et al. [13] utilized wavelet entropy to classify unsegmented and short duration PCG signals where a wavelet entropy threshold was determined from the training set then PCG signals with entropy below the threshold were classified as abnormal. In Singh-Miller et al. [14], spectral features with discriminative model based on random forest regressor were applied for classification of heart sound recordings. Vernekar et al. [15] demonstrated Markov features with a weighted ensemble classifier that include four AdaBoost ensemble classifiers and four artificial neural networks (ANN) to classify PCG signals. Plesinger et al. [16] implemented a fuzzy logic like approach with logical rules and probability assessment based on histograms to classify heart sounds. Nabhan and Warriek [17] attempted to improve the work in [12] by detection the outlier signal and separated it from standard range signal by using an interquartile range threshold. Abdollahpur et al. [18] proposed an algorithm that assessed the signal quality of the segmented cardiac cycle then a total of ninety features including time domain, time-frequency, perceptual and Mel-frequency cepstral coefficient (MFCC) were extracted from the correctly segmented cycles only. The classification was performed using three feed-forward neural networks followed by a voting system. Langley and Murray [19] tried to improve the work in [13] where short and unsegmented heart sounds recordings were classified using feature threshold based classifier. Spectral amplitude and wavelet entropy features were calculated using FFT and wavelet analysis. A decision tree was then used to combine the spectral amplitude and wavelet entropy. Maknickas and Maknickas [20] applied CNN to classify heart sound records with Mel-frequency spectral coefficients (MFSC), difference and second-order difference of the MFSC calculated and fed to CNN as three dimensions for each frame. Whitaker et al. [21] proposed combining sparse coding and time domain features for heart sounds classification. Six SVM classifiers used. Abduh et al. [39] applied fractional Fourier transform based Mel-frequency spectral coefficients (FrFT-MFSC) and stacked auto-encoder deep neural network to classify heart sound records. In spite of the excellent performance of the new method, there is always the concern of overfitting problem in such methods. Deep learning methods require large amounts of training data that are difficult to meet by biomedical data sets, and the current problem is no exception. Therefore, the estimation of the many parameters associated with the deep learning neural network architecture using the much fewer training samples available from our data set raises a major concern that the system may not generalize well. Consequently, studies in this field have attempted to propose augmentation or simulation methods to generate more data samples from the limited available ones. This raises concerns about how realistic such synthetic data samples are and whether this approach should be acceptable for medical diagnostic systems. A number of regulatory organizations as AAMI (Association for the Advancement of Medical Instrumentation), Medicines & Healthcare products Regulatory Agency (MHRA), and British Standards Institute (BSI) held a joint workshop in 2019 to discuss governance and regulation of emerging artificial intelligence and machine learning technologies in healthcare and provided their conclusions in an important position paper [40]. Among their recommendation was a particularly relevant one about the critical importance of quantity and quality of data input to intelligent systems. Even though such recommendations are not yet incorporated into active regulations, it seems that systems based on such approaches using synthetic data will be much harder to be cleared by regulatory agencies given the burden of proof of their validity. Therefore, it is desirable to develop a system that is based on traditional classification with much fewer parameters to estimate and would allow the limited available data set to be sufficient to train, validate and test the system in a robust manner.

In this work, we develop a computer-aided auscultation system based on the fractional Fourier transform based Mel-frequency spectral coefficients features set developed by our group [39] combined with traditional classifiers. The new system is described in terms of its preprocessing, cardiac cycle segmentation, feature extraction, feature reduction and classification stages. The proposed system with several variants of its implementation are verified on the dataset of the PhysioNet/Computing in Cardiology Challenge 2016 and compared to previous work on the same dataset using the evaluation metric of this challenge.

2. Methodology

The components of the proposed computer-aided auscultation system are shown in Fig. 1 and are detailed as follows.

2.1. Preprocessing

Ambient sounds, lung sound, internal body noise, cough and stethoscope movement are the main interferences in heart sounds recording and analysis. Therefore, effective filtration is of critical important to enhance the heart sounds signal by reducing the influence of background noise and removing spike noise. In this work, a two-stage preprocessing system is utilized. In the first stage, a 3rd-order Butterworth band-pass filter with corner frequencies of 15 and 800 Hz is used to select the useful bandwidth of the heart sounds. In the second stage, spectral subtraction denoising method was applied [22]. This method was reported to be highly effective for background noise reduction in such challenging applications as EEG signal denoising for brain-computer interface [23]. The advantage of this technique is its adaptive noise estimation. The noise power is estimated from the frequencies outside of the range of frequencies of heart sounds then a weighted version of it is subsequently subtracted from the power spectrum of raw heart sounds and used to reconstruct the denoised signal [22]. In this work, spectral subtraction filtration was used with a weighting factor of 0.5.

2.2. Cardiac cycle segmentation

In this stage, each PCG signal is split into cardiac cycles. It is essential for the recognition of the systolic or diastolic states, permitting succeeding categorizing of abnormal states in this areas. Many algorithms were implemented. Some was carried out by the usage of a reference signal such as the ECG, the segmentation algorithms require to record the ECG in parallel. It will assist to recognize heart sounds. Other algorithms do not use ECG as a reference. In this work Springer's improved version of Schmidt's segmentation algorithm [24] utilized to split PCG signal into cardiac

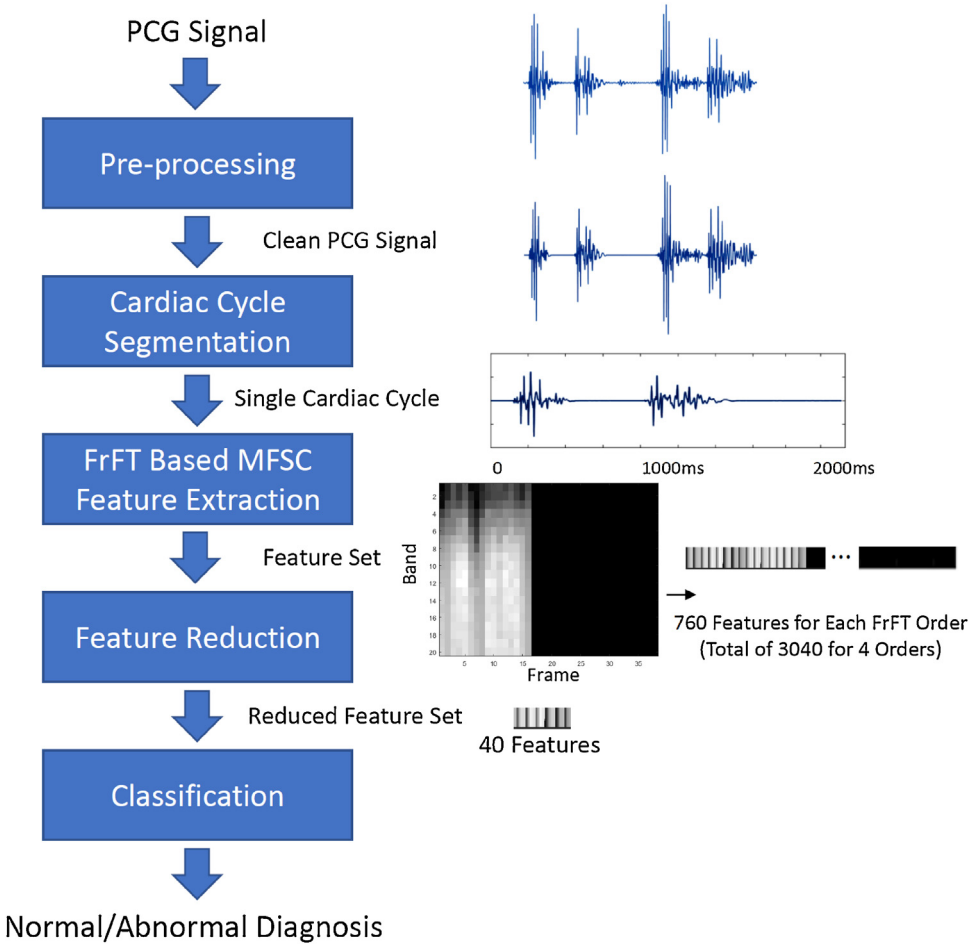

Fig. 1. Block diagram of the proposed approach for classification of heart sounds with an illustration of the signals obtained at each step (PCG: Phonocardiogram, FrFT: Fractional Fourier transform. MFSC: Mel-frequency spectral coefficients).

cycles. Then, each complete cardiac cycle is used for processing. This algorithm does not require ECG synchronization and uses a logistic regression hidden semi-Markov model (HSMM) to estimate the most likely sequence of states by incorporating information about expected heart sound state durations. To overcome the problem of variable time length of cardiac cycles (and hence size of their digital signals) in subsequent processing steps, the size of all signals was set to be the longest cardiac cycle found across all PCG recordings (here, it was around 2 s). For cardiac cycles with shorter length, they were zero-padded to that length. This ensured uniform frequency resolution for all signals.

2.3. Feature extraction

In this stage, the segmented cardiac cycles are processed to extract a set of features that best represents their salient characteristics for optimal classification accuracy. Generally speaking, since phonocardiogram signal is essentially similar to speech signal, popular feature representations currently used for speech signals such as spectral and cepstral features can be potentially useful in this application as well [25]. Among the most successful of these are the Mel-frequency cepstral coefficients (MFCC) transform the original PCG signal into a time-frequency representation of the distribution of signal energy [4]. The MFCC features correspond to the cepstrum of the log filter-bank energies. The Mel-frequency scale is approximately linear for frequencies below 1 kHz and logarithmic for frequencies above 1 kHz [27]. This is motivated by the fact that the human auditory system becomes less frequency-selective as frequency increases above 1 kHz.

In this work, we use the fractional Fourier transform based Mel-frequency spectral coefficients features (FrFT-MFSC) developed by our group [39]. These features are based on a modified version of MFCC whereby Fractional Fourier transform is used instead of discrete Fourier or the discrete cosine transforms to allow for a flexible time-frequency expansion. The log-energy was computed directly from the Mel-frequency spectral coefficients.

The Fractional Fourier Transform (FrFT) is the generalization form of Fourier transform and indicates a rotation of a signal in time-frequency plane. While the discrete Fourier or cosine transforms can obtain the frequency components of a signal, they are essentially decomposing the signal in terms of harmonic components that are not localized in time. Therefore, they cannot describe local variations in the signal as compared to a time-frequency analysis method. The fractional Fourier transform offers a generalization of the Fourier transform with the Fourier spectrum and the time domain signal are both special cases of this transform. Hence, it allows a more flexible time-frequency representation than other methods such as the spectrogram, Wigner distribution or ambiguity function in that it can transform signals to any intermediate domain between time and frequency. Hence, FrFT is suitable for non-stationary signal processing and has wide applications in basic signal analysis and speech recognition.

The continuous form of FrFT with ath-order of a signal $s(t)$ can be defined within $0 \leq |a| \leq 2$ through the linear operator as [28–31],

$$\left(F^a s\right)(w_a) = \int_{-\infty}^{\infty} K_a(w_a, t)\, s(t)\, dt. \tag{1}$$

Here, the kernel $K_a(w_a, t)$ is given by:

$$K_a(w_a, t) = \begin{cases} k_\emptyset\, exp(j\pi \left(w_a^2 \cot(\emptyset) - 2w_a t \csc(\emptyset) + t^2 \cot(\emptyset)\right)), \ if\ a \neq 0,\ \pm 2, \\ \delta(w_a - t),\ if\ a = 0, \\ \delta(w_a + t),\ if\ a = \pm 2 \end{cases} \tag{2}$$

where, $\emptyset = \frac{a\pi}{2}$ and $k_\emptyset = \frac{\exp(-j(\frac{\pi\, sgn(\emptyset)}{4} - \frac{\emptyset}{2}))}{\sqrt{|\sin(\emptyset)|}} = \sqrt{1 - j\cot(\emptyset)}$, and w_a means the variables in ath-order fractional Fourier transform. Similar to the discrete Fourier transform (DFT), the discrete fractional Fourier transform (DFrFT) matrix $F^a(m, n)$ is obtained as the discrete version of Eq. (2) as,

$$F^a(m, n) = \sum_{k=0}^{N-1} u_k(m) \exp(\frac{-j\pi k a}{2})\, u_k(n)\,. \tag{3}$$

Here, u is discrete Hermite-Gaussian function and a is the fractional order. The discrete fractional Fourier transform of a signal is just the matrix vector multiplication of this transform matrix in Eq. (3) with the signal vector.

Fig. 2 shows the steps of FrFT-MFSC extraction, which are derived as follows:

1 The signal is pre-emphasized by a filter H(z)= (1 – 0.9 z^{-1}) that boosts the higher frequencies to balance the spectrum of sounds steep roll-off in the high frequency region [38]. Then the signal is framed and windowed into the selected frames length.
2 The fractional Fourier transform of the windowed signal is calculated.
3 The powers of the obtained spectrum are mapped into the Mel-scale using triangular overlapping windows.
4 The logs of the powers are calculated at each Mel-frequency to obtain the FrFT-MFSC values.

We perform normalization to make all features in the range of [0,1] [26]. The Mel-scale is defined as follows where f is the frequency in Hz:

$$Mel(f) = 2595\, log_{10}\left(1 + \frac{f}{700}\right). \tag{4}$$

For FrFT-MFSC feature extraction, the processing parameters were as follows: four fractional orders (a = 0.90,0.95,1.0,1.1). The frame length parameter was chosen to cover the duration of each of the fundamental heart sounds S1 and S2 (nominal values of S1 = 122 ms and S2 = 114 ms). So, the chosen frame length is taken as 125 ms. The frame shift was taken close to 50 % of the frame length as is commonly used in the literature and chosen to be 50 ms. Fig. 3 illustrates the frames for a single segmented cardiac cycle after zero-padding to 2000 ms. The number of bands was taken as 20 based on experimentation. For each fractional order, a 20 (bands) × 38 (frames) spectral coefficients matrix was computed. That is, for each cardiac cycle, a total of 3040 features were obtained.

2.4. Feature dimensionality reduction

Principal component analysis (PCA) is used to convert possibly correlated features into linearly uncorrelated ones with an orthogonal transformation. It also identifies components that do not contribute much to the representation (or variance or the data) and hence can be removed. In this work, the percentage of variance accounted for in PCA was set as 95 %.

2.5. Classification

Many factors affect the performance of traditional classifiers and therefore should be considered during the mathematical model building and as well as in the generalization for real data analysis. The most important factors include the distribution of data in the feature space, balance of class data samples, heterogeneity, diversity of training data and presence of noise. Furthermore, each classifier has relative advantages and disadvantage compared to others. For example, k-nearest neighbor (KNN) is a memory-based learning technique [32] and therefore training and testing data should both be available all the time. Also, decision tree classifiers are preferred for noisy datasets [33]. On the other hand, boosted ensemble classifier is reported to perform best for imbalanced data [34].

In this work, we check the performance of different traditional classifiers on a set of features derived from dataset by using cross-validation method. We carry out training to search for the best classification model, including support vector machines (using linear, quadratic, cubic, and Gaussian kernels), k-nearest neighbor (using linear, cosine, cubic, and weighted distance metrics), and ensemble classification (bagged Trees, subspace KNN and RUSBoosted tree) [35,36]. Table 1 summarizes the classification parameters considered in this study.

Cross-validation and local holdout methods were used to train and test the different classifiers. In the cross-validation experiments, the feature vector data were divided into 5 folds. Each fold was held out in turn for testing and the model was trained for that fold using all data outside the fold. Then, each model performance was assessed using the data inside the fold, and the overall results are computed as the average over all folds. In the local holdout method experiments, 80 % of feature vector data was used as training set and the remaining 20 % was used for testing and both were chosen randomly. This 5-fold cross-validation and 80-20 % holdout experiments were performed 5 times by randomly selecting/dividing the given data in order to estimate the mean performance variability of the classifier.

2.6. Database description

In any pattern recognition problem, database selection plays important role in model building and generalization. Database should be large, diverse and understandable. So analysis and comparison of the different implemented algorithms will be easy.

The performance of the proposed system was verified using the database of the PhysioNet/Computing in Cardiology Challenge 2016 [7]. The used dataset includes 3153 recordings. The sure labeled data of this dataset includes 2868 recordings collected from six databases. Normal patient records are 2249 whereas abnormal patient records are 619, lasting from 5 s to just over 120 s. All recordings have been resampled to 2000 Hz and have been provided as ".wav" format. They were recorded in different real-world clinical and nonclinical environments and include recordings of varying amounts of noise. The data were recorded from both normal subjects and pathological patients, and from both children and adults. The same patient could have between 1 and 6 recordings in the database. The data were recorded from different locations on the body (including aortic area, pulmonic area, tricuspid area and mitral area, among others). The data are clearly imbalanced since the number of normal recordings are much larger than that of abnormal recordings. The dataset was divided randomly into two independent sets with 80 % for training and 20 % for testing.

It should be noted that this dataset incorporates realistic challenges of noisy, imbalanced data in addition to different data

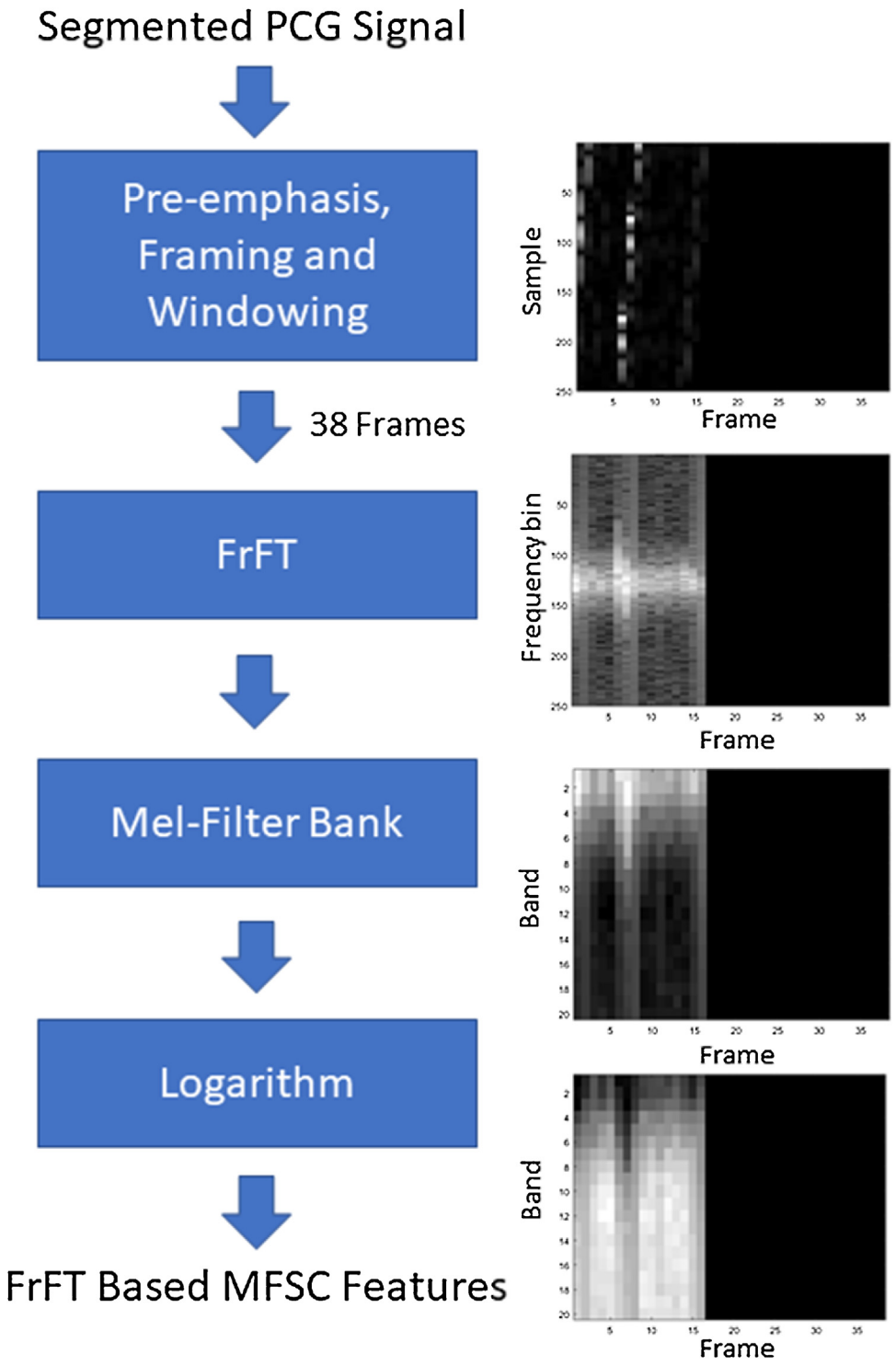

Fig. 2. Feature extraction using fractional Fourier transform Mel-frequency spectral coefficients with an illustration of the processed data at each step (PCG: Phonocardiogram, FrFT: Fractional Fourier transform. MFSC: Mel-frequency spectral coefficients).

collection environments (that is, collection was done by different research groups). Heterogeneity in the collection of the recordings introduces differences with potential to make classifier training more difficult. In particular, a classifier trained on one population might perform poorly when applied to another.

2.7. Performance evaluation

To evaluate the performance of classification process, the confusion matrix is computed with the abnormal cases as the positive

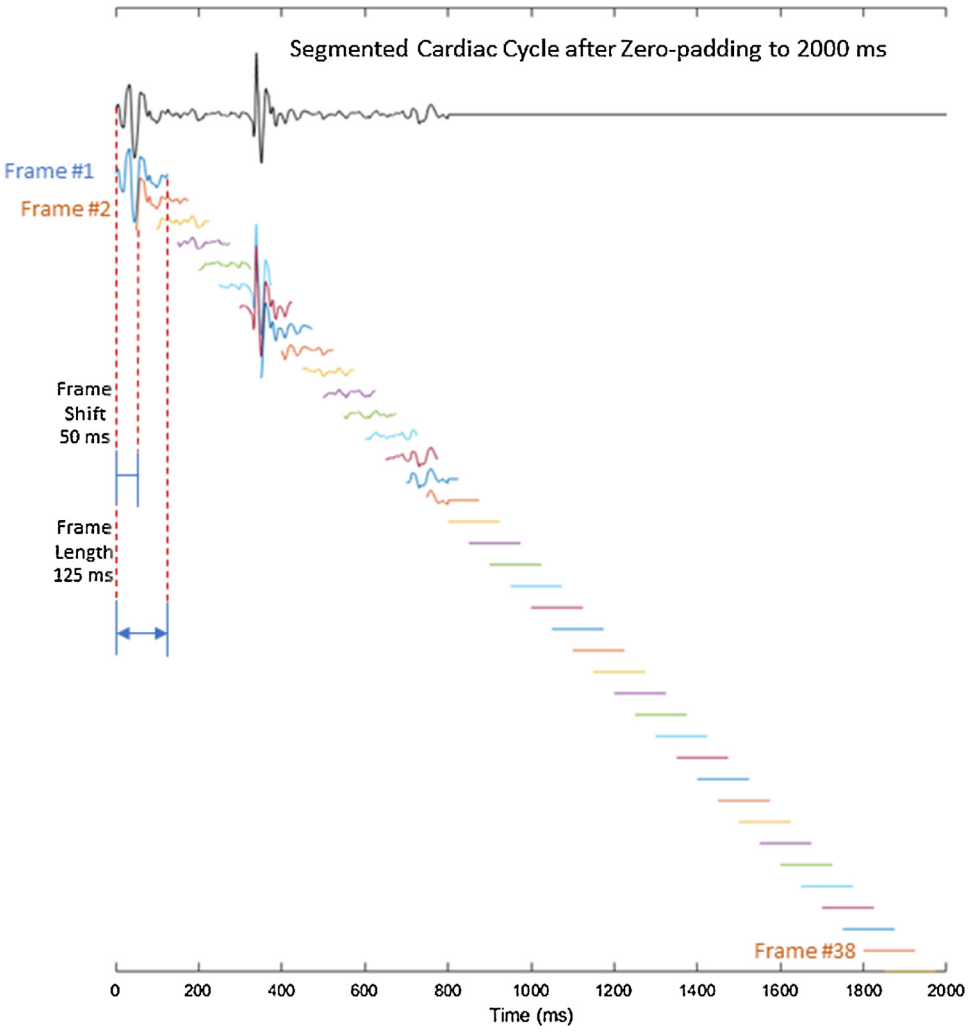

Fig. 3. The frames used for a single segmented cardiac cycle after zero-padding.

Table 1
Parameters of the used classifiers.

Classifier		Parameters
SVM	Linear	Linear kernel, Sequential Minimal Optimization solver
	Quadratic	Quadratic kernel, polynomial, 2nd order, Sequential Minimal Optimization solver
	Cubic	Cubic kernel, polynomial, 3rd order, Sequential Minimal Optimization solver
	Gaussian	Gaussian kernel, polynomial, Sequential Minimal Optimization solver
Ensemble Classifier	Bagged Trees	Combine Weights: Weighted Average, training cycles = 200
	Subspace KNN	Combine Weights: Weighted Average, training cycles = 200
	RUSBoosted Tree	Combine Weights: Weighted Average, training cycles = 200
KNN	Linear	Euclidean distance, Distance weight = Equal, k = 10
	Cosine	Cosine distance, k = 10
	Cubic	Cubic distance, k = 10
	Weighted	Distance weight, k = 10

class and used to calculate the values for sensitivity, specificity and accuracy as,

$$Sensitivity\,(Se) = \frac{TP}{TP + FN}, \tag{5}$$

$$Specificity\,(Sp) = \frac{\mathrm{TN}}{TN + FP}, \tag{6}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{7}$$

where TP, TN, FP, and FN are the confusion matrix entries representing true positive, true negative, false positive and false negative cases, respectively.

For imbalanced data, the error rate and the accuracy are not appropriate to measure the classification performance because they do not consider misclassification costs. As a result, they tend to be strongly biased to favor the majority class and are sensitive to class skews [34,39]. Therefore, an alternative evaluation score based on the average between sensitivity and specificity was chosen as the official evaluation metric for the PhysioNet/Computing in Cardiology Challenge 2016 and defined as,

$$Score = \frac{Se + Sp}{2}. \tag{8}$$

Table 2
FrFT- MFSC features classification results using 5-fold cross-validation train-test method showing the mean and standard deviation (SD) of 5 experiments.

Classifier			Sensitivity	Specificity	Accuracy	Score
SVM	Linear	Mean	0.6352	0.9429	0.8815	0.7890
		SD	0.0009	0.0002	0.0002	0.0004
	Quadratic	Mean	0.8284	0.9568	0.9312	0.8926
		SD	0.0074	0.0014	0.0007	0.0032
	Cubic	Mean	0.8657	0.9634	0.9439	0.9145
		SD	0.0051	0.0012	0.0014	0.0027
	Gaussian	Mean	0.7989	0.9569	0.9254	0.8779
		SD	0.0110	0.0024	0.0019	0.0049
Ensemble	Bagged Trees	Mean	0.7387	0.9718	0.9254	0.8553
		SD	0.0115	0.0015	0.0028	0.0059
	Subspace KNN	Mean	0.8365	0.9703	0.9437	0.9034
		SD	0.0094	0.0019	0.0022	0.0047
	RUSBoosted Tree	Mean	0.7845	0.8502	0.8371	0.8173
		SD	0.0131	0.0049	0.0048	0.0070
KNN	Linear	Mean	0.8277	0.9656	0.9381	0.8967
		SD	0.0088	0.0011	0.0020	0.0045
	Cosine	Mean	0.8012	0.9718	0.9378	0.8865
		SD	0.0277	0.0084	0.0015	0.0097
	Cubic	Mean	0.7305	0.9728	0.9245	0.8517
		SD	0.0416	0.0064	0.0034	0.0177
	Weighted	Mean	0.7901	0.9712	0.9351	0.8807
		SD	0.0082	0.0010	0.0013	0.0038

Table 3
FrFT- MFSC features classification results using 80 % as training and 20 % as testing showing the mean and standard deviation (SD) of 5 experiments.

Classifier			Sensitivity	Specificity	Accuracy	Score
SVM	Linear	Mean	0.6292	0.9425	0.8807	0.7861
		SD	0.0006	0.0014	0.0002	0.0003
	Quadratic	Mean	0.8208	0.9305	0.9082	0.8757
		SD	0.0024	0.0605	0.04802	0.0298
	Cubic	Mean	0.8735	0.9666	0.9469	0.9200
		SD	0.0018	0.0035	0.0006	0.0013
	Gaussian	Mean	0.7879	0.9583	0.9239	0.8731
		SD	0.0020	0.0011	0.0007	0.0012
Ensemble	Bagged Trees	Mean	0.7372	0.9709	0.9245	0.8540
		SD	0.0033	0.0006	0.0008	0.0017
	Subspace KNN	Mean	0.8453	0.9698	0.9447	0.9075
		SD	0.0054	0.0008	0.0004	0.0024
	RUSBoosted Tree	Mean	0.7907	0.8495	0.8386	0.8201
		SD	0.0149	0.0097	0.0055	0.0043
KNN	Linear	Mean	0.8286	0.9659	0.9377	0.8972
		SD	0.0013	0.0023	0.0002	0.0016
	Cosine	Mean	0.7845	0.9770	0.9382	0.8808
		SD	0.0036	0.0017	0.0003	0.0013
	Cubic	Mean	0.7173	0.9775	0.9252	0.8474
		SD	0.0074	0.0015	0.0020	0.0042
	Weighted	Mean	0.7887	0.9716	0.9356	0.8802
		SD	0.0012	0.0009	0.0003	0.0005

The results in this study present the values of all the above measures.

3. Results and discussion

3.1. Experimental verification

The performance of the proposed system was verified using the dataset of the PhysioNet/Computing in Cardiology Challenge 2016 described above [7]. Each of the PCG records was preprocessed using the Butterworth band-pass filter of order 3 with corner frequencies 15 and 800 Hz then enhanced further using the spectral subtraction denoising with weight 0.5. Each record was segmented into cardiac cycles resulting in 79,492 cardiac cycles. The longest cardiac cycle found across all PCG recordings has a length of around 2 s. So, if a cardiac cycle had a length less than 2 s, the time series was zero-padded to that length.

For FrFT-MFSC feature extraction, a total of 3040 features were computed for each cardiac cycle. To improve the classification process, all features vector values were normalized to interval [0,1]. Then, the feature reduction process was performed using PCA such that the percentage of variance represented by the results is 95 %. This process reduced the dimensionality of the feature space down to 40 that are computed as weighted linear combination from all 3040 features.

Performance comparison was done for several traditional classifiers that included support vector machine (SVM), k-nearest neighbor (KNN), bagged Trees ensemble classifier, subspace KNN ensemble classifier, and RUSBoosted tree ensemble classifier with parameters listed in Table 1.

3.2. Results

Table 2 summarizes the results of classification using 5-folds cross validation train-test method. The cross-validation train-test method appears to reduce the effect of over-fitting and helps in model evaluation more effectively when using imbalanced data. It should be noted that most classifiers in our experi-

Table 4
Comparison of the performance metrics of different variants of the proposed method.

Classifier	Sensitivity	Specificity	Score
SVM - Cubic Kernel	0.8735	0.9666	0.9200
Subspace KNN Ensemble classifier	0.8453	0.9698	0.9075
KNN – Linear	0.8286	0.9659	0.8972
SVM -Quadratic Kernel	0.8284	0.9568	0.8926
KNN – Cosine	0.8012	0.9718	0.8865
KNN – Weighted	0.7901	0.9712	0.8807
SVM – Gaussian Kernel	0.7989	0.9569	0.8779
Ensemble classifier Bagged Trees	0.7387	0.9718	0.8553
KNN – Cubic Kernel	0.7305	0.9728	0.8517
Ensemble Classifier RUSBoosted Tree	0.7907	0.8495	0.8201
SVM - Linear Kernel	0.6352	0.9429	0.7890

ments achieved relatively balanced values for sensitivity and specificity. The SVM classifier with cubic kernel achieved the highest score value of 0.9145 with sensitivity and specificity of 0.8657 and 0.9634 respectively. The SVM classifier with cubic kernel achieved the highest sensitivity of 0.8657. On the other hand, SVM classifier with linear kernel achieved the lowest sensitivity of 0.6352. The KNN classifier with cubic distance metric classifier with achieved the highest specificity of 0.9728 while RUSBoosted tree ensemble classifier the lowest specificity of 0.8502.

Table 3 summarizes the results of classification using local holdout train-test method. Most classifiers here also achieved relatively balanced values for sensitivity and specificity. Also, the SVM classifier with cubic kernel achieved the highest score value of 0.9200 with sensitivity and specificity of 0.8735 and 0.9666 respectively. The SVM classifier with cubic kernel achieved the highest sensitivity of 0.8735 while the SVM classifier with linear kernel achieved the lowest sensitivity 0.6292. On the other hand, the KNN classifier with cubic distance metric achieved the highest specificity of 0.9775, while RUSBoosted tree ensemble classifier produced the lowest specificity of 0.8495.

The results of several variants of the proposed work are presented in Table 4. Also, the results reported in the previous work in the literature are presented in Table 5.

3.3. Discussion

The SVM classifier with cubic kernel achieved the highest score value of 0.9200 with sensitivity and specificity of 0.8735 and 0.9666 respectively. Hence, it provided the best estimate of the predictive score in both methods and relatively balanced values for sensitivity and specificity. Performance assessments of classifiers in both train-test methods were generally similar. The proposed system led to significant improvemeernel achieved the highest score value of 0.9200 with sensitivity and specificity of 0.8735 and 0.9666 respectively. Hence, it provided the best estimate of the predictive score in both methods and relatively balanced values for sensitivity and specificity. Performance assessments of classifiers in both train-test methods were generally similar. The proposed system led to significant improvements in classification performance using traditional classifiers, which indicates the robustness of the new system and its ability to handle diverse PCG recordings and signal quality conditions.

Table 4 shows several variants of the proposed method as they rank by the official challenge score. It also shows the classification methods used in each and also the sensitivity and specificity values. The comparison of such variants should address not only the score but also the balance between sensitivity and specificity values. As can be observed, two variants of the proposed method provide excellent performance metrics that outperform all previous methods applied on the same dataset in Table 5 except the method by our group in [39] that uses deep learning. Furthermore, other variants also appear to provide good performance metrics as compared to the pervious methods applied to the same dataset. Also, it should be noted that the sensitivity and specificity values obtained for all entries from the new system are relatively close. This indicate clear potential of the proposed system and indicate the robustness of the used features which appear to work with various traditional classification methods and parameters.

It is important to observe the effect of the class distribution of the training data, which is a very important factor that determine the quality of subsequent classification. In the challenge database, the number of samples seems to be large but unfortunately these samples are heavily imbalanced with much different numbers of normal and abnormal recordings. This causes conventional classifiers to often become biased toward the majority class and therefore increase misclassification rate for the minority class. Therefore, the performance of different classifiers was evaluated on the dataset by using two validation methods. The first is the cross-validation train-test method, which reduces the effect of over-fitting produced by noisy records. The second is the local holdout method that simulates real life analysis by building the model with 80 % of used data and use the rest to evaluate the model.

Furthermore, the question comes of that possible data overfitting occurs in this study. Overfitting in the context of machine learning refers to a system that models the details and noise in the training data in a way that adversely affects performance on new data. To investigate the extent of this problem in our study, the results were reported in Tables 2 and 3 as the mean and standard deviation of the outcome of 5-fold cross validation and 80-20 % local holdout several experiments respectively. This approach follows the well-known random resampling technique that is used to validate machine learning models. As can be observed, the standard deviations are less than 0.5 % of the mean value reported for all results, which is considered small. The overall analysis of all results, the mean of random variations in the cross-validation results was 0.786 % with median of 0.511 %, again both are fairly small. These results indicate that the problem of overfitting may not have a significant impact on the outcome of this study based on the small variability in the different resampling experiments. Nevertheless, it remains a concern to this study.

The results of the previous work are generally within a narrow score metric range from 0.73 to 0.93 even they use different types of features and classifiers as shown in Table 5. Also, it is noted that the best scores were achieved with the most complex models. For example, Potes et al. [11] used final decision rule based on combination of ensemble AdaBoost classifier and convolutional neural network. Also, Homsi et al. [12] employed a nested set of ensemble classifiers that includes cost-sensitive classifier (CSC), LogitBoost (LB) and random forest (RF). Abduh et al. [39] applied fractional Fourier transform based Mel-frequency spectral coefficients (FrFT-MFSC) and stacked auto-encoder deep neural network. The proposed features still gave good result with the much simpler classifier using SVM with cubic kernel. This indicates the potential of the used features and their impact in simplifying the development process of the classification system.

This study indicates the potential of FrFT-MFSC features for phonocardiogram analysis where they represent the distribution of PCG signal energy in a more effective manner even under different noise and recording environment conditions. Applying FrFT-MFSC features with different fractional orders provides a good tool to represent noisy PCG signal in different time-frequency planes where they also preserve locality in both time and frequency. The use of log-energy computed directly from the Mel-frequency spectral coefficients maintains such time-frequency localization. The alternative yet more common use of discrete cosine transform for this

Table 5
Comparison of the performance metrics of the previous methods reported in the literature.

Study	Classifier	Sensitivity	Specificity	Score
Abduh *et al* [39].	Stacked Autoencoder Deep Neural Network	0.8930	0.9700	0.9315
Plesinger et al. [16]	Fuzzy logic like approach	0.8690	0.9370	0.9030
Langley and Murray [19]	Amplitude spectrum and wavelet entropy threshold followed by decision tree	0.8690	0.9370	0.9030
Goda et al. [9]	SVM	0.9308	0.8470	0.8870
Nabhan and Warriek [17]	Nested ensemble of algorithm including random forest, LogitBoost and cost-sensitive classifier	0.8740	0.9140	0.8940
Abdollahpur et al. [18]	Three feed-forward NNs	0.8883	0.8851	0.8867
Whitaker et al. [21]	SVM	0.8867	0.8816	0.8841
Homsi et al. [12]	Nested ensemble of algorithms including random forest, LogitBoost and cost-sensitive classifier	0.9440	0.8690	0.8840
Tschannen et al. [11]	L2-SVM	0.9080	0.8320	0.8700
Potes et al. [3]	Final decision rule based on AdaBoost ensemble classifier and Convolutional neural network	0.8800	0.8200	0.8500
Singh-Miller *and* Singh-Miller [14]	Discriminative model based on random forest regressor	0.8100	0.8900	0.8500
Potes et al. [3]	Convolutional neural network (CNN)	0.7900	0.8600	0.8200
Vernekar et al. [15]	Weighted ensemble classifier including four AdaBoost ensemble and four ANN classifiers	0.7920	0.8430	0.8200
Potes et al. [3]	AdaBoost- ensemble classifier	0.7000	0.8800	0.7900
Langley *and Murray* [13]	Wavelet entropy threshold	0. 9500	0.6000	0.7800
Gokhale [8]	Boosted trees ensemble classifier			0.7600
Grzegorczyk et al. [10]	Conventional neural network and autoencoder	0.8300	0.6200	0.7300

computation projects the spectral energies into a new basis that would not maintain such localization.

Noting that the dataset is significantly imbalanced, it is interesting to observe from Table 4 that several variants of the proposed method that performed very well used conventional classifiers such as SVM and KNN that are not intended for imbalanced data. In fact, we can see also from Table 5 that SVM was also reported to perform well by several previous studies. This indicates that the dataset has sufficiently limited statistical variability that allows the characteristics of the underlying distribution of the abnormal class to be learned by such classifiers from the available number of samples. So, even though the use of ensemble classifiers is always recommended for imbalanced data in general [37], the dataset used in our problem presents an interesting special case as confirmed by our results and those from previous work.

Although it may seem that the computation of 3040 features is a lot to process, we aimed to analyze all this information to establish first that they are useful. Moreover, given the relatively slow sampling rate and the speed of today's computers, processors or FPGAs, it is practically possible to compute all such features in real-time and at reasonable cost as well.

Even though the PCA does not provide binary in/out reduction of features like the classical feature reduction approaches, one can still gain insight into which features are more relevant by observing their weights in the principal components that explain the 95 % of data variance. For example, in our results, we found that 466 features had significant weights above 90 % of the maximum weight in the principal component feature combination derived from PCA. Fig. 4 shows the distribution of these significant features across the different fractional orders. It indicates balanced importance for features from all orders with slight advantage to order 1.0 then 0.95. Fig. 5 shows the distribution of the same features over the different frames within the cardiac cycle. It clearly indicates that the initial 600 ms part of the cardiac cycle is the most relevant to the diagnosis. This range show more emphasis that coincides with both the S1 and S2 parts of the PCG and to a lesser extent with the systole part. So, the outcome from PCA gives further insight into the physiological correlation of the set of features emphasized by PCA based on weight that can be used to find out the most useful of them.

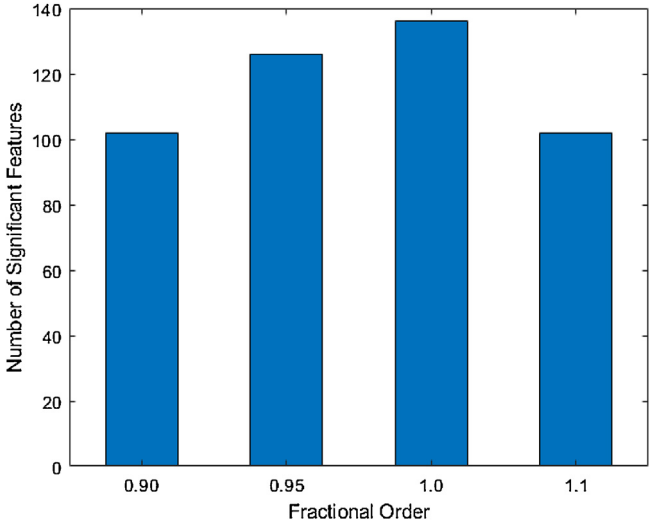

Fig. 4. The distribution of significant features in the principal component across different values of the fractional order.

3.4. Study limitations

Some difficulties were encountered in dealing with the data that are common to all similar studies. First, the variations in heart rate lead to temporal record length variations. This was addressed via zero-padding of all recordings to the length of the longest one. This maintains the same sampling rate while harmonizing frequency resolution among all records, which is critical to our method given its reliance on frequency domain features. Second, the inter-patient differences made it challenging to build a model that generalizes well across patients. This remains a challenge in biomedical signal measurement in general and there is not much to do to reduce

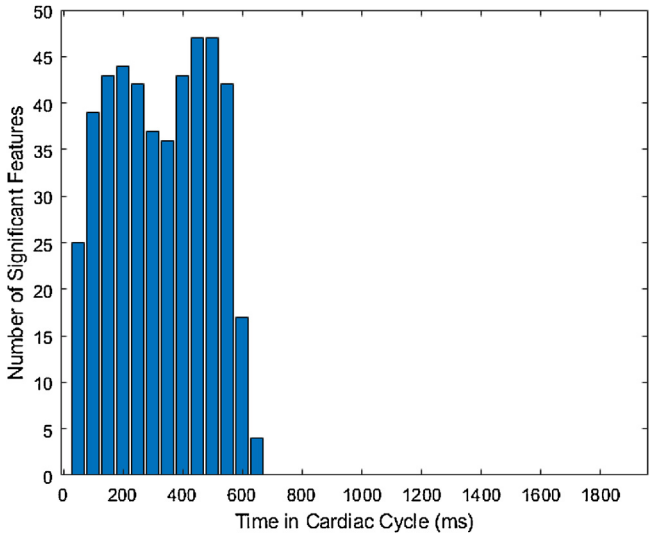

Fig. 5. The distribution of significant features in the principal component across different frames within the cardiac cycle.

such differences. Finally, differences introduced by heterogeneity in the collection of the recordings can render a classifier trained on one population much less effective when applied to another. Here, even though the features proposed seem to capture the salient features among all such populations as evident from the robust performance among populations, the problem of overfitting still cannot be ignored and therefore is considered among the potential limitations of the present study. Such limitations are open research points for future research.

4. Conclusions

A new approach is proposed to classify heart sounds. The main contribution of this approach is the combination of the features based on fractional Fourier transform based Mel-frequency spectral coefficients and traditional classifiers that offer simplicity and robustness against overfitting. The description of the proposed methodology and its implementation details are presented. The results of experimental verification using the database of the PhysioNet/Computing in Cardiology Challenge 2016 indicate that the presented approach and several variants of it performed well with an evaluation score reaching 0.92 in the best results obtained using a support vector machine classifier with cubic kernel. These results confirm the robust performance obtained by the new system compared to the previous work applied on the same data. This indicates the potential of the new system for clinical utility.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.bspc.2019.101788.

Declaration of Competing Interest

The authors have declared no conflict of interest.

References

[1] WHO, Cardiovascular Diseases World Statistics on WHO, 2017, last updated May. www.who.int/mediacentre/factsheets/fs317/en/ (Accessed October 6, 2019).
[2] C. Jian, G. Xingming, X. Shouzhong, Study on the signification and method of heart sound recognition, Foreign Med. Biomed. Eng. Fascicle 27 (2) (2004) 87–89.
[3] C. Potes, S. Parvaneh, A. Rahman, B. Conroy, Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds, Proc. 2016 Computing in Cardiology Conference (CinC) (2016) 621–624.
[4] J. Rubin, R. Abreu, A. Ganguli, S. Nelaturi, I. Matei, K. Sricharan, Classifying heart sound recordings using deep convolutional neural networks and mel-frequency cepstral coefficients, 2016 Computing in Cardiology Conference (CinC) (2016) 813–816.
[5] G.D. Clifford, C. Liu, B. Moody, D. Springer, I. Silva, Q. Li, et al., Classification of normal/abnormal heart sound recordings: the PhysioNet/Computing in cardiology challenge 2016, Proc. 2016 Computing in Cardiology Conference (CinC) (2016) 609–612.
[6] G.D. Clifford, C. Liu, B. Moody, J. Millet, S. Schmidt, Q. Li, et al., Recent advances in heart sound analysis, Physiol. Meas. 38 (2017) E10.
[7] C. Liu, D. Springer, Q. Li, B. Moody, R.A. Juan, F.J. Chorro, et al., An open access database for the evaluation of heart sound algorithms, Physiol. Meas. 37 (12) (2016) 2181–2213.
[8] T. Gokhale, Machine learning based identification of pathological heart sounds, 2016 Computing in Cardiology Conference (CinC) (2016) 553–556.
[9] M.A. Goda, P. Hajas, Morphological determination of pathological PCG signals by time and frequency domain analysis, Proc. 2016 Computing in Cardiology Conference (CinC) (2016) 1133–1136.
[10] I. Grzegorczyk, M. Soliński, A. Łepek Michałand Perka, J. Rosiński, J. Rymko, K. St\kepień, et al., PCG classification using a neural network approach, Proc. 2016 Computing in Cardiology Conference (CinC) (2016) 1129–1132.
[11] M. Tschannen, T. Kramer, G. Marti, M. Heinzmann, T. Wiatowski, Heart sound classification using deep structured features, Proc. 2016 Computing in Cardiology Conference (CinC) (2016) 565–568.
[12] M.N. Homsi, N. Medina, M. Hernandez, N. Quintero, G. Perpiñan, A. Quintana, et al., Automatic heart sound recording classification using a nested set of ensemble algorithms, Proc. 2016 Computing in Cardiology Conference (CinC) (2016) 817–820.
[13] P. Langley, A. Murray, Abnormal heart sounds detected from short duration unsegmented phonocardiograms by wavelet entropy, Proc. 2016 Computing in Cardiology Conference (CinC) (2016) 545–548.
[14] N.E. Singh-Miller, N. Singh-Miller, Using spectral acoustic features to identify abnormal heart sounds, Proc. 2016 Computing in Cardiology Conference (CinC) (2016) 557–560.
[15] S. Vernekar, S. Nair, D. Vijaysenan, R. Ranjan, A novel approach for classification of normal/abnormal phonocardiogram recordings using temporal signal analysis and machine learning, Proc. 2016 Computing in Cardiology Conference (CinC) (2016) 1141–1144.
[16] F. Plesinger, I. Viscor, J. Halamek, J. Jurco, P. Jurak, Heart sounds analysis using probability assessment, Physiol. Meas. 38 (8) (2017) 1685–1700.
[17] H.M. Nabhan, P. Warrick, Ensemble methods with outliers for phonocardiogram classification, Physiol. Meas. 38 (8) (2017) 1631–1644.
[18] M. Abdollahpur, A. Ghaffari, S. Ghiasi, M.J. Mollakazemi, Detection of pathological heart sounds, Physiol. Meas. 38 (8) (2017) 1616–1630.
[19] P. Langley, A. Murray, Heart sound classification from unsegmented phonocardiograms, Physiol. Meas. 38 (8) (2017) 1658–1670.
[20] V. Maknickas, A. Maknickas, Recognition of normal-abnormal phonocardiographic signals using deep convolutional neural networks and mel-frequency spectral coefficients, Physiol. Meas. 38 (8) (2017) 1671–1684.
[21] B.M. Whitaker, P.B. Suresha, C. Liu, G. Clifford, D. Anderson, Combining sparse coding and time-domain features for heart sound classification, Physiol. Meas. 38 (8) (2017) 1701–1729.
[22] S. Boll, Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. Acoust. 27 (2) (1979) 113–120.
[23] M.J. Alhaddad, M.I. Kamel, M.M. Makary, H. Hargas, Y.M. Kadah, Spectral subtraction denoising preprocessing block to improve P300-based brain-computer interfacing, Biomed. Eng. Online 13 (1) (2014) 13–36.
[24] D.B. Springer, L. Tarassenko, G.D. Clifford, Logistic Regression-HSMM-Based heart sound segmentation, IEEE Trans. Biomed. Eng. 63 (4) (2016) 822–832.
[25] D.P.W. Ellis, PLP, RASTA, MFCC And Inversion in Matlab, 2012, last updated. www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/ (Accessed October 16, 2017).
[26] Z. Abduh, M.A. Wahed, Y.M. Kadah, Robust computer-aided detection of pulmonary nodules from chest computed tomography, J. Med. Imaging Health Inform. 6 (3) (2016) 693–699.
[27] D. O'Shaughnessy, Speech Communications: Human and Machine, 2nd ed., Wiley-IEEE Press, 1999.
[28] A.C. McBride, F.H. Kerr, On Namias's fractional Fourier transforms, IMA J. Appl. Math. 39 (2) (1987) 159–175.
[29] L.B. Almeida, The fractional Fourier transform and time-frequency representations, Ieee Trans. Signal Process. 42 (11) (1994) 3084–3091.
[30] C. Candan, M.A. Kutay, H.M. Ozaktas, The discrete fractional Fourier transform, IEEE Trans. Signal Process. 48 (5) (2000) 1329–1337.
[31] S.-C. Pei, M.-H. Yeh, C.-C. Tseng, Discrete fractional Fourier transform based on orthogonal projections, IEEE Trans. Signal Process. 47 (5) (1999) 1335–1348.
[32] A. Clark, C. Fox, S. Lappin, Handbook of Computational Linguistics and Natural Language Processing, Wiley-Blackwell, 2010.
[33] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.
[34] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, A review on ensembles for the class imbalance problem: bagging-, boosting-, and

hybrid-based approaches, IEEE Trans. Syst. Man Cybern. Part C 42 (4) (2012) 463–484.
[35] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, 3rd edition, Prentice Hall, Englewood Cliffs, New Jersey, 2010.
[36] S. Sayad, An Introduction to Data Science, Last Updated, 2017 (Accessed October 16, 2017) http://www.saedsayad.com/further_readings.htm.
[37] V. López, A. Fernández, S. Garc'ia, V. Palade, F. Herrera, An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics, Inf. Sci. 250 (2013) 113–141.
[38] K.S. Rao, S.G. Koolagudi, Robust emotion recognition using spectral and prosodic features, Springer Briefs Speech Technol. (2013) 109.
[39] Z. Abduh, E.A. Nehary, M.A. Wahed, Y.M. Kadah, Classification of Heart Sounds Using Fractional Fourier Transform Based Mel-Frequency Spectral Coefficients and Stacked Autoencoder Deep Neural Network, J. Med. Imaging Health Inform. 9 (1) (2019) 1–8.
[40] The emergence of artificial intelligence and machine learning algorithms in healthcare: recommendations to support governance and regulation, in: Position Paper Prepared by BSI and AAMI. The AAMI/BSI Workshop on Artificial Intelligence and Machine Learning Algorithms in Health Technology, Arlington, Virginia, May, 2019.