Intelligent Environments 2019
A. Muñoz et al. (Eds.)

doi:10.3233/AISE190061

Toward Early Detection and Monitoring of Chronic Heart Failure Using Heart Sounds

Martin GJORESKI[a,1], Anton GRADIŠEK[a], Borut BUDNA[a], Matjaž GAMS[a] and Gregor POGLAJEN[b]
[a] *Department of Intelligent Systems, Jožef Stefan Institute, Ljubljana, Slovenia*
[b] *Advanced Heart Failure and Transplantation Programme, Department of Cardiology, UMC Ljubljana, Slovenia*

Abstract. Chronic heart failure (CHF) affects over 26 million of people worldwide and represents a significant societal, logistic and financial burden both for the patients and for the healthcare system, necessitating novel management approaches of this patient population. In this paper, we explore the possibilities of detecting heart failure worsening based on heart sounds using machine-learning methods. First, we developed a method that distinguishes between healthy individuals and those with a decompensated CHF episode. Our method includes filtering, segmentation, feature extraction, and machine learning, and was tested with a leave-one-subject-out evaluation technique on the data from 193 individuals. The method achieved 82% accuracy, outperforming the baseline classifier for 14 percentage points. In the next stage, we explored the differences between decompensated and recompensated states of CHF patients. We identified ten features for which there is statistically significant difference ($p<0.001$) in the features distributions, when calculated between decomensated and recompensated state of CHF. These features may be the key for developing algorithms for continuous personalized remote monitoring of the CHF patients.

Keywords. Chronic heart failure, wearable device, heart sound

1. Introduction

Chronic heart failure (CHF) is a chronic progressive condition where the heart is unable to pump enough blood to meet the metabolic needs of tissues and organs at the physiological filling pressures [1]. The incidence of CHF is increasing by 2%. In developed world, CHF affects 1-2 % of total population and 6-10 % of people older than 65 years. Despite the progress in medical- and device-based treatment approaches in the last decades, the overall prognosis of CHF is still dismal as 5-year survival of this population only reaches 50%. In the typical clinical course of CHF we observe alternating episodes of compensated (when the patient feels well) and decompesated phases when symptoms and signs of chronic heart failure (such as dyspnea, orthopnea, pulmonary edema, lived congestion, pulmonary edema etc.) can readily be established. During the latter episodes patients often require hospital admission for treatment with intravenous medications (diuretics, inotropes) to achieve successful recompensation. Early HF worsening detection would allow a treating physician to timely adjust patient's medical management and thus avoid hospital admission. Currently an experienced physician can detect worsening of HF by examining the patient and through characteristic changes in the changes in heart failure biomarkers (determined from the patient's blood). Additionally, in some patients, characteristic changes in heart sounds

[1] martin.gjoreski@ijs.si

can accompany heart failure worsening and can be heard using phonocardiography. One of these "typical" CHF-associated sounds is the third sound (S3) that appears 0.1-0.2 s after the second sound S2. Unfortunately, clinical worsening of the CHF patient likely means that we are already dealing with a fully developed CHF episode. Recently, it has been demonstrated that some physiological parameters (such as appearance of additional heart sounds or increased blood pressure in the pulmonary circulation) already start to show a several weeks before CHF patient develops a clinically evident heart failure deterioration. This is also a window where outpatient-based treatment interventions can reverse CHF deterioration and return the patient to the compensated state without the need for hospital admission.

In this paper, we focus on the detection of state of CHF (compensated vs. decompensated) based on the analysis of heart sound recordings. Our work builds upon the initial studies where we demonstrated that it is possible to distinguish between healthy individuals and patients in the decompensated CHF episode using a stack of machine-learning classifiers, showing promising results on a limited dataset [1]. We expand upon this approach using a considerably larger patient dataset. Furthermore, we investigate the difference in heart sounds during the transition, between decompensated and recompensated state of CHF, with the aim of developing personalized monitoring models. Early detection of worsening of CHF has the potential to lead towards reduction of hospitalizations due to worsening CHF, with both improving the quality of life of patients and decreasing the financial and logistic burden on the patient and the health system.

2. Dataset

Heart sound recordings were obtained using a professional digital stethoscope 3MTM Littmann Electronic Stethoscope Model 3200. Our dataset (Table 1) contains recordings of 110 "healthy" people (meaning that they had no medical condition that would manifest itself in abnormal heart sound) and 51 people diagnosed with CHF. For 22 CHF patients, recordings were obtained both during the decompensation episode (when hospitalized) and during the compensated phase (when discharged from the hospital). The recordings were always obtained at Erb's point and each recording was up to 30 s long (stethoscope's limit). For some healthy people, more than one recording was obtained to increase the amount of data in the study (recordings of patients were obtained in clinical settings which limited the available time). The study was approved by the medical ethics committee beforehand.

Table 1. Overview of the experimental data recorded on healthy individuals and on patients in decompensated and recompensated CHF episodes

	Decompensated	Recompensated	Healthy	Overall
# Subjects	51	22	110	183
# Recordings	52	22	159	233
# Segments	2017	865	6272	9154
Duration (min)	17	7	52	76

3. Methodology

The machine-learning stacking scheme used in this study is presented in Fig. 1. The following subsections describe in detail the key steps of the pipeline.

3.1. Filtering and segmentation

According to Choi et al [2], the majority of cardiovascular sounds are most likely to occur in the frequency range below 1 kHz, thus, we filtered the raw audio files using low-pass Butterworth filter with a threshold of 1 kHz. For segmentation of the filtered audio signal, we used a sliding window of 1 s with an overlap of 0.5 s. This provides audio segments with a duration of 1 s. The size of the window was chosen with the reasoning that for an average heart rate of 60 beats per minute, the 1 s window should contain at least one heartbeat, thus containing all the relevant information about the CHF-related sounds.

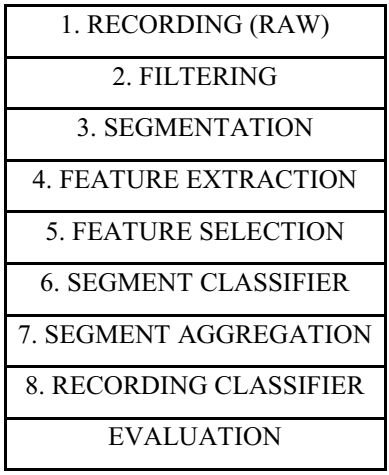

Figure 1. Machine-learning stacking scheme used for the analysis performed in this study.

3.2. Feature extraction

3.2.1. Audio features

We used the OpenSMILE tool [3] for large feature-space extraction. The tool was originally created for acoustic emotion recognition in 2009 but later expanded to more general uses. For example, in addition to affect recognition, it is widely used for music information retrieval (e.g., chord labelling, beat tracking, and onset detection). We extracted 1941 audio features, including statistical features (e.g., variance, skewness, kurtosis), energy-based features (e.g., energy in bands from 250 to 1 kHz), frequency-based features (25 %, 50 %, 75 %, and 90 % spectral roll-off points) and voicing-related features (e.g., jitter, shimmer, Harmonics-to-Noise Ratio). The complete list of features is described by Schuller et al. [4].

3.2.2. Heart rate variability features

Heart rate variability (HRV) describes the variation in the time interval between the heartbeats (R-R intervals). Fig. 2 presents an example of filtered audio data with marked R-R intervals that were detected automatically using a peak detection algorithm that finds the maximum of the first heart tone, S1. Once the peaks were detected, the R-R intervals simply represent the distance between these peaks.

We calculated the following HRV features: mean of the R-R intervals, the standard deviation of the R-R intervals, the square root of the mean of the squares of differences between the adjacent R-R intervals, the percentage of differences between adjacent R-R intervals that are greater than 50 ms, and the Poincaré plot indexes of HRV (SD1 and SD2). Each of the segments within the same recording was then attributed the HRV-based feature value of the entire recording [5].

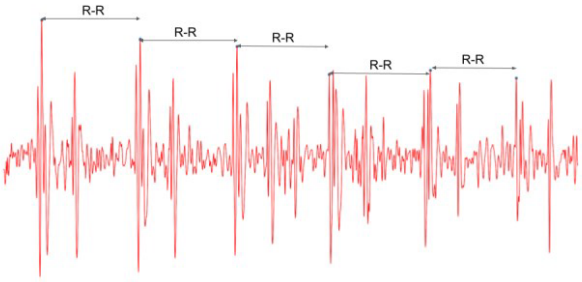

Figure 2. Representative filtered segment with detected R-R intervals

3.3. Feature selection

Since the number of available features (6 HRV and 1941 audio features) is several times larger than the number of available recordings, we selected only the best features to avoid overfitting. We ranked the features using Mutual information values between the features and the class values on the training data. Mutual information is a measure that estimates the dependency between two random variables. In the training phase, we used only the top-ranked *n* features, where *n* was set to be equal to the number of training samples.

3.4. Machine learning classifiers

The ML pipeline contains two ML classifiers, a segment-based and a recording-based classifier. The *segment-based* classifier takes as input 1 s segments (represented via the extracted features), and outputs the estimated class probabilities for each segment. The *recording-based* classifier takes as input a recording, and outputs estimated class probabilities for the recording. The motivation here is the fact that all the segments in a chosen recording belong to the same class, therefore the aggregation of segment-based forecasts should improve the overall classification.

The last classifier in the stacking scheme is the recording-based classifier. The features for the recording-based classifier are calculated using mean aggregation over the features of all segments in the recording. In addition, the average class probabilities for each segment, estimated by the segment-based classifier, are appended as features. Since the recording-based classifier is a meta-learner that utilizes the output of the segment-based classifier, a holdout set is required for its training. For that reason, we used a leave-one-subject-out (LOSO) technique on the training data, to generate the meta-training dataset. All models were trained using the Random Forest (RF) algorithm. We used the RF algorithm because of its robustness to noise in the input features.

4. Experimental results

We evaluated the proposed approach in two experimental setups: classifying CHF patients (both decompensated and recompensated) vs. healthy, and classifying decompensated vs. recompensated recordings. LOSO cross-validation was used for the two experimental setups. The first experimental setup simulates a situation where the algorithm detects whether a new user shows symptoms of CHF. The second experimental setup is aimed at detecting worsening of the condition in patients that have already been diagnosed with CHF, and represents a first step in building a method that would be able to notify the user to seek medical help before a hospitalization is required. The results are presented in the following subsections.

4.1. CHF patients vs. healthy analysis

The results of the LOSO evaluation are presented in Table 2. The table contains the results achieved by the proposed *Stacking classifier* and by the *Segment classifier*. The recording predictions for the Segment classifier are calculated by taking the majority prediction for all segments in one recording.

The first two rows in Table 2 present the confusion matrices for the Stacking classifier and for the Segment classifier. The entries in the confusion-matrices represent recordings. The rest of the rows in Table 2 present the performance metrics. The majority class is 68 % i.e., 68 % of the recordings belong to healthy individuals. The Stacking approach has achieved the accuracy of 82 %, which is 14 percentage points higher that the majority class and 6 percentage points higher than the Segment-based approach. In addition, compared to the Segment-based approach, the Stacking approach has significantly higher precision, recall, and F1-score.

Table 2. Confusion matrices and performance measures (recall, precision, F1-score and accuracy)

	Stacking		**Segment-based**	
	Patient	**Healthy**	**Patient**	**Healthy**
Patient	49	29	30	47
Healthy	14	151	17	148
Recall	0.63	0.92	0.39	0.90
Precision	0.78	0.84	0.64	0.76
F1-score	0.70	0.88	0.48	0.82
Accuracy	0.82		0.74	
Majority	0.68			

4.2. Decompensated vs. recompensated analysis

In this scenario, we aimed at some level of personalization. For 22 out of 51 patients, there is one recording in the decompensated phase (i.e., at the beginning of hospitalization) and one recording in the recompensated phase (i.e., at hospital discharge). Since the number of paired recordings is small and not suitable for ML experiments, we preformed statistical tests to check whether there is a statistically significant difference in the feature values when calculated from the recompensated recordings compared to the decompensated recordings (i.e., before and after a medical intervention). We used the Wilcoxon signed-rank test, which is a non-parametric statistical hypothesis test that tests whether two related paired samples come from the same distribution. In particular, it tests whether the distribution of the differences x - y is symmetric about zero [6]. In this case, x are the values of the features extracted from the decomposition recordings and y are the values of the features extracted from the decomposition. The Wilcoxon test is an alternative to the paired Student's t-test, when the population cannot be assumed to be normally distributed, such as in our case.

The statistical tests showed that there are 10 features for which there is a statistically significant change in the distribution. Those features are:

- 99th percentile of the first derivative of the 4th MFCC coefficient;
- Quartile deviation of the first derivative of the 4th MFCC coefficient;
- 1st percentile of the first derivative of the spectral roll-off;
- 1st percentile of the standard deviation of the spectral roll-off;
- Percentile range (0-1) of the Spectral roll-off;
- Flatness of the psychoacoustic sharpness;
- Percentile range (0-1) of the psychoacoustic sharpness;
- Percentile range (0-1) of the first derivative of the psychoacoustic sharpness;
- Standard deviation of the psychoacoustic sharpness;
- 99th percentile of the spectral entropy.

The Mel-frequency Cepstrum Coeffiecients (MFCC) are the coefficients of the MFC representation of the sounds. The MFC is a representation of the short-term power spectrum of the sound [7]. The spectral roll-off is a measure of the amount of the right-skewedness of the power spectrum of the signal. Similarly, the psychoacoustic sharpness and the spectral entropy are two features which quantify the spectral characteristics of the sound.

The boxplots of the 10 features are presented in Fig. 3. The boxplots clearly show that there is difference in the values of the features between the recordings of patients in recompensated and decompensated episodes.

5. Discussion

In line with our previous research based on the heart sound analysis [1], the stacking-based approach is promising in determining patient's state of CHF and detecting subclinical signs of threatening CHF deterioration. The accuracy is somewhat lower than what we obtained when comparing healthy controls to CHF patients. However, before, we were dealing only with CHF patients in the fully decompensated state, where the

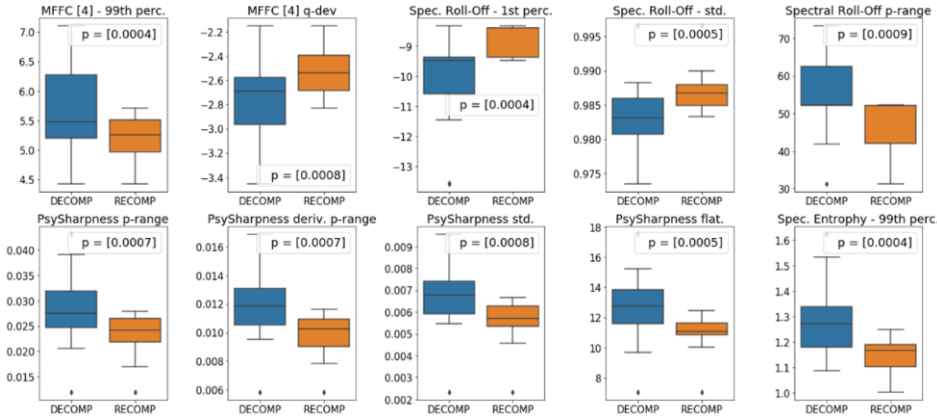

Figure 3. Boxplot and p-values that show a statistically significant difference in the values when calculated from the recompensated recordings compared to the decompensated recordings (i.e., before and after the CHF therapy).

effects of increased ventricular filling pressures on heart sounds are much more pronounced. In addition, the dataset for CHF patients was considerably larger and more diverse in the current study, which makes the results more representative. Furthermore, the inclusion of the recording-based features (derived from HRV) expands the feature room upon which we operate.

The results of personalized approach in detecting preclinical worsening of CHF condition are encouraging. We have identified at least 10 features that show statistically significant difference for CHF patients before and after the medical intervention. All of the 10 features quantify the spectral representation of the sounds. Thus, the spectral features might be more informative compared to the temporal features for the specific task. To further check whether these features can be used to build personalized models, a larger dataset will be required. Ideally, a pilot study would include CHF patients regularly recording their heart sounds, perhaps using a modified mobile/wearable device instead of a professional stethoscope. However, a labelled dataset for the specific modified mobile/wearable device would be required in order to tune the ML models.

In future work, we plan to analyze the effects that different hyperparameters have on the methods accuracy. This analysis includes sampling rates, window size, more advanced feature selection methods, e.g., wrapper method instead ranking methods. Alternatively, dimensionality reduction methods instead of feature selection, e.g., deep autoencoders, can turn out to be efficient. An intrinsic problem in building personalized models lies in the fact that only a small number of recordings will always be available for each individual. Here, the data from healthy patients can be utilized by using multi-task learning techniques, where one task would be healthy vs. non-healthy and the other task would be decompensated vs recompensated. Finally, we plan to test our method on other related datasets, e.g., the dataset used in the PASCAL Classifying Heart Sounds Challenge [8] or the dataset used in the PhysioNet/Computing in Cardiology Challenge [9].

References

[1] M. Gjoreski, M. Simjanoska, A. Gradišek, A. Peterlin, M. Gams and G. Poglajen, "Chronic Heart Failure Detection from Heart Sounds Using a Stack of Machine-Learning Classifiers," 2017 International Conference on Intelligent Environments (IE), Seoul, 2017, pp. 14-19.

[2] S. Choi, Z. Jiang, "Comparison of envelope extraction algorithms for cardiac sound signal segmentation", Expert Systems with Applications, vol. 34, no. 2, pp. 1056-1069, 2008.

[3] F. Eyben, F. Weninger, F. Gross, B Schuller: "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor", In Proc. ACM Multimedia (MM), Barcelona, Spain, ACM, ISBN 978-1-4503-2404-5, pp. 835-838, October 2013. doi:10.1145/2502081.2502224

[4] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, M. Pantic, "AVEC 2011–The First International Audio/Visual Emotion Challenge", Affective Computing and Intelligent Interaction pp 415-424, 2011.

[5] F. Shaffer, JP. Ginsberg. An Overview of Heart Rate Variability Metrics and Norms. Front Public Health. 2017;5:258. Published 2017 Sep 28. doi:10.3389/fpubh.2017.00258

[6] F. Wilcoxon, Frank, "Individual comparisons by ranking methods", 1948, Biometrics Bulletin. 1 (6): 80–83. doi:10.2307/3001968.

[7] M. Sahidullah, G. Saha. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. Speech Communication, 54(4), pp.543-565, 2012.

[8] P.J. Bentley, G. Nordehn, M. Coimbra, S. Mannor, R. Getz, "The PASCAL Classifying Heart Sounds Challenge 2011," www.peterjbentley.com/heartchallenge/.

[9] D. Gari. Clifford et al. "Classification of normal/abnormal heart sound recordings: The PhysioNet/Computing in Cardiology Challenge 2016", Computing in Cardiology Conference (CinC), 2016.