Validating Features for Atrial Fibrillation Detection from Photoplethysmogram under Hospital and Free-living Conditions

Linda M. Eerikäinen[1,2], Lukas Dekker[1,3], Alberto G. Bonomi[2], Rik Vullings[1], Fons Schipper[2], Jenny Margarito[2], Helma M. de Morree[2], Ronald M. Aarts[1,2]

[1]Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands
[2]Philips Research, Eindhoven, The Netherlands
[3]Department of Cardiology, Catharina Hospital Eindhoven, Eindhoven, The Netherlands

Abstract

Atrial fibrillation (AF) is the most commonly experienced sustained arrhythmia, and it increases risks of stroke and congestive heart failure. Unobtrusive wearable solutions with photoplethysmography (PPG) have been proposed for AF detection and the performance has been mainly evaluated for short-term measurements in controlled measurement settings. In this study, we evaluate the predictive value of features from PPG for AF detection under both hospital and free-living conditions. PPG from the wrist was measured from 18 patients before and after cardioversion and from 16 patients (4 with 100% AF) for 24 hours. Single-lead ECG and 24-hour Holter were used respectively as gold standards. Six PPG-based inter-beat interval (IBI) variability and irregularity features were computed in three different sliding time windows. Thresholds for AF classification for every individual feature were determined with the data from the hospital conditions and tested with the measurements from free-living conditions. Overall, the best classification results were obtained by using a 120-s window, pNN40 resulting as the best feature. On average, the sensitivity was higher in the hospital conditions (92.3% vs. 71.6%) and the specificity higher in the free-living conditions (60.7% vs. 84.9%). In conclusion, testing the classification perfomance in free-living conditions is essential to properly evaluate AF detection models.

1. Introduction

Atrial fibrillation (AF) is the most commonly experienced sustained arrhythmia and its prevalence increases with age. The arrhythmia increases the risk of stroke to five-fold and the risk of congestive heart failure to three-fold. [1] Early diagnosis of AF has a great importance, especially for the prevention of stroke. However, AF can be asymptomatic and therefore can remain undiagnosed. For detecting paroxysmal events, long-term or frequent monitoring is needed.

A measurement technique suitable for unobtrusive long-term monitoring is photoplethysmography (PPG). PPG is an optical measurement, which records blood volume changes in the vascular bed of the tissue, enabling extraction of cardiovascular parameters, such as heart rate.

PPG-based solutions intended eventually for long-term monitoring purposes have been proposed for AF detection based on features determining the irregularity or variability of the inter-beat intervals (IBIs) [2–4]. However, the results in these studies have been reported only on short-term recordings up to 10 minutes. We previously showed that a Markov model could predict AF in free-living conditions using PPG data [5]. During continuous long-term monitoring the measurements are more prone to noise and the feature values might be less accurate compared to the short-term setting. In addition, different solutions are using different time windows for the feature computation. In this paper, we evaluate the most common irregularity and variability features for IBIs with three different window lengths, in both a hospital condition before and after an electrical cardioversion procedure, and in a free-living condition during 24-hour measurements.

2. Data

The data for the study were recorded in two different settings in Eindhoven, The Netherlands: before and after an electrical cardioversion (CV) procedure in the hospital and in 24-hour measurements in free-living conditions. The study was approved by the local medical ethical committee and every patient provided written informed consent before participating. An overview of the datasets is presented in Table 1. During 24-hour measurements, patients either had 100% AF (4 patients) or no AF.

 ISSN: 2325-887X DOI:10.22489/CinC.2017.050-177

Table 1. Datasets

	Number of patients	Males (%)	Age (y) (m ± sd)	Total rec. length, non-AF (hh:mm)	Total rec. length, AF (hh:mm)
CV	18	56	75 ± 11	13:41	16:26
24h	16	63	65 ± 14	298:32	89:57

2.1. Measurements in hospital conditions

The measurements in hospital conditions were performed in the department where patients are treated with electrical cardioversion. 20 patients assigned for AF treatment were included in this part of the study. The patients were measured approximately one hour before and one hour after the procedure with PPG and accelerometer sensors at the wrist with a data logging device equipped with the Philips Cardio and Motion Monitoring Module (CM3 Generation-3, Wearable Sensing Technologies, Philips, Eindhoven). As a reference, a single-lead electrocardiogram (ECG) was measured from the chest with Actiwave Cardio (CamNtech Ltd., Cambridge, United Kingdom). At the beginning and at the end of the recording a synchronization protocol was performed by shaking both devices simultaneously.

The rhythm before and after the procedure was evaluated by a clinical expert by looking at the ECG. Figure 1 shows an example of 30 s of PPG signal and corresponding IBIs before and after the cardioversion. Two patients with unsuccessful cardioversion were excluded from further analysis to include only patients with both AF and regular rhythm. Baseline characteristics and medication information of the patients were collected afterwards from the patient record.

2.2. Measurements in free-living conditions

The measurements in free-living conditions were performed in 16 patients assigned for a 24-hour Holter examination. PPG and accelerometer data were measured at the non-dominant wrist with the same wrist-wearable device as in the hospital conditions. The ECG was recorded with a 12-lead Holter monitor (H12+, Mortara, Milwaukee, WI, USA). The Holter monitor was attached first to the patient by following normal hospital procedures and a synchronization protocol was performed by tapping the wrist-wearable device and pressing the event button on the Holter simultaneously. The same procedure was performed when the patient arrived at the hospital the following day to return the devices. During the measurement period, patients were keeping a diary of their activities, complaints, and medication. The diary was handed in at the time when the measurement ended.

The ECG recordings were analyzed by trained analysts,

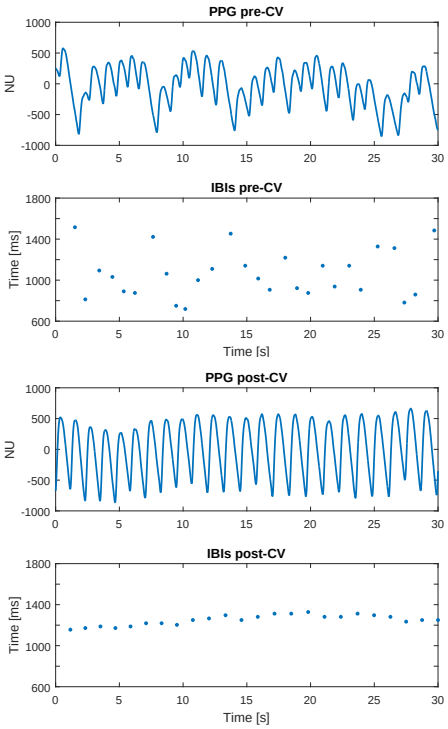

Figure 1. 30-s segments of PPG and corresponding IBIs before the cardioversion (above) and after the cardioversion (below) for a representative patient.

supported by software (Veritas, Mortara, Milwaukee, WI, USA) that automatically detects the time and type of the beat. Every heart beat in the ECG was labelled either as sinus rhythm, AF, premature supraventricular or ventricular contraction, artifact, or unknown. The output of the software was verified or corrected by the analysts. In addition, baseline characteristics and medication intake information were collected.

3. Methods

The goal of the analysis was to compare features describing irregularity or variability of IBIs and the ability of the features to classify the rhythm as AF and non-AF in the measurement settings described in Section 2. Before starting the feature computation, the PPG was filtered with a 0.3 Hz high-pass filter and a 5 Hz low-pass filter. Heart beats were detected from the PPG pulses and after the pulse extraction, the beat information in PPG and ECG were aligned. The IBIs from PPG were computed as time differences between two consecutive pulses.

3.1. Features

For feature computation, outlier removal was made based on IBI length and IBIs $<$ 200 ms and $>$ 2200 ms

were discarded. The features expressing variability or irregularity of the IBI sequence studied were Root Mean Square of Successive Differences (RMSSD), Shannon Entropy (ShE), the percentage of interval differences of successive intervals greater than 40 ms (pNN40) and greater than 70 ms (pNN70), and Sample Entropy (SampEn). 40 ms and 70 ms were selected based on Corino et al. [4] where the combination of the two features was found to be the most discriminative feature combination for AF.

ShE is a measure that has been successfully used to quantify the irregularity of IBI sequences during AF [2, 6]. For calculating the entropy, first the probability distribution of the IBIs is computed assigning the intervals to fixed number of bins with equal size. The probablity of the IBI to fall in the bin i is

$$p(i) = \frac{n_{(i)}}{l - n_{outliers}}, \quad (1)$$

where $n_{(i)}$ is the number of IBIs that fall in the bin i, l the total number of IBIs in the window, and $n_{outliers}$ the number of IBIs considered as outliers. When the probabilities for every bin are known, ShE is

$$\text{ShE} = -\sum_{i=1}^{N} p(i) \frac{log(p(i))}{log(N)}. \quad (2)$$

N is the number of bins and was selected to be $N = 16$, which is the minimum number of bins to obtain a reasonable accuracy [6].

SampEn evaluates similar patterns in the time series and a lower value indicates more self-similarity in the time series. In detail, SampEn is the negative natural logarithm of the conditional probability that two sequences similar to each other at m points are similar also at $m + 1$ points. SampEn was computed according to [7]

$$\text{SampEn} = -\ln(A/B) = -\ln(A) + \ln(B), \quad (3)$$

where A is the number of similar sequences of length $m+1$ and B the number of similar sequences of length m within tolerance r. Two SampEn features were generated by setting m equal to 1 and 2 (SampEn1 and SampEn2), and r equal to 0.25 times the standard deviation of the series as in [4].

The features were computed in three different window lengths: 30 s, 60 s, and 120 s, by sliding with 30 s.

3.2. Performance metrics

The statistical measures used to assess the predictive value of the features were sensitivity (Sens = TP/(TP+FN)), specificity (Spec = TN/(TN+FP)), and accuracy (Acc = (TP+TN)/(TP+FP+TN+FN)), where TP is the number of true positives, TN true negatives, FP false positives, and FN false negatives.

3.3. Cross-validation

The measurements in the hospital conditions were considered to be more controlled because of their shorter duration and the patients were in supine position during the entire measurement period. Therefore, the hospital dataset was used as a training set for defining the thresholds for every individual feature for every window length. This was done with a stratified leave-one-out cross-validation. One patient was held as a test data and the set of remaining patients was used to define the threshold which would give an optimum cut-off point on the receiver operating characteristic curve according to Youden's index. The procedure was repeated 18 times leaving each patient for testing one time.

The classification to AF and non-AF in the free-living conditions was based on the thresholds defined with the dataset in the hospital conditions. The mean of the thresholds of the cross-validation were selected as the final ones for every feature and window length.

4. Results

Table 2. Sensitivity and specificity in the hospital

Feature	30 s		60 s		120 s	
	Sens	Spec	Sens	Spec	Sens	Spec
ShE	89.1	54.3	89.1	53.9	88.9	54.5
RMSSD	88.2	44.7	87.9	44.9	88.5	44.2
pNN40	93.5	64.6	94.7	63.9	97.1	66.0
pNN70	91.1	61.1	95.7	59.4	93.9	60.2
SampEn1	86.1	63.6	85.4	68.3	93.2	67.4
SampEn2	88.0	57.5	89.2	66.4	92.3	71.7
Mean	89.3	57.7	90.3	59.5	92.3	60.7

The sensitivity and specificity of the AF classification in the hospital conditions for every window length are presented as mean values over all patients in Table 2 and accuracy is presented in Table 3. The standard deviation for sensitivity varied between 5.5–25.7%, for specificity between 17.5–41.5%, and for accuracy between 11.4–19.7%. The highest sensitivity (97.1%) and the highest accuracy (83.8%) were obtained with pNN40 and the highest specificity was with SampEn2 (71.7%) with a 120-s window.

Table 3. Accuracy in the hospital conditions

Feature	30 s	60 s	120 s
ShE	73.4	73.2	73.4
RMSSD	68.4	68.4	68.4
pNN40	83.8	81.5	83.8
pNN70	80.7	79.9	79.1
SampEn1	75.3	76.4	80.9
SampEn2	74.7	77.8	82.3
Mean	76.1	76.2	78.0

The results for the free-living conditions were calculated only in terms of sensitivity and specificity, and are listed

in Table 4. The standard deviations for sensitivity ranged between 9.5–13.8% and for specificity 2.8–13.3%. The highest sensitivities were obtained with a 120-s window and were similar for all the features ranging from 69.0% to 72.7%. pNN40 was the feature with the highest specificity (94.3%). Accuracy was not computed due to having only four patients with AF and 12 without AF in the dataset. The accuracies would not be comparable to the accuracies in the hospital dataset which is more balanced.

Table 4. Sensitivity and specificity in free-living

Feature	30 s		60 s		120 s	
	Sens	Spec	Sens	Spec	Sens	Spec
ShE	40.0	96.1	60.5	92.5	72.3	89.3
RMSSD	40.0	87.8	60.3	82.8	72.4	78.7
pNN40	40.1	96.5	60.3	95.1	72.3	94.3
pNN70	40.3	96.3	60.7	93.8	72.7	93.1
SampEn1	37.7	78.0	57.1	78.9	70.7	76.5
SampEn2	35.9	73.7	55.7	76.1	69.0	77.6
Mean	39.0	88.1	59.1	86.5	71.6	84.9

5. Discussion

This is the first study evaluating features for variability and irregularity of IBIs both in hospital and free-living measurement conditions. When comparing the sensitivity and specificity, the results showed a difference between the two conditions. On average, e.g. with 120-s window length, the sensitivity was higher in the hospital conditions (92.3%) compared to the free-living conditions (71.6%). The specificity, on the contrary, was higher in free-living (84.9%) than in the hospital (60.7%).

In addition to different measurement conditions, the different patient profiles might cause differences in the performance. In hospital conditions, after the electrical cardioversion there might still be irregularities, such as premature contractions, present in the rhythm. Five patients experienced a large number of irregulaties after the procedure which explains the low mean specificity and the high standard deviation of specificity (up to 41.5%). In free-living conditions, the density of premature contractions was lower on average which might explain the higher specificity.

The thresholds for the features were trained with the data recorded in the hospital. A large amount of irregularities in the non-AF group in the data set might have caused the thresholds to be higher than optimal thresholds for the free-living conditions causing the sensitivity to drop.

In free-living conditions, sensitivity improved significantly when increasing the window length. Specificity did decrease, but to a smaller extent. This indicates that a longer window length gives better classification results. This was expected, because the type of features used in this study become more reliable with more data. Interestingly, in the hospital conditions the window length did not seem to influence significantly the classification performance.

6. Conclusion

The classification performance of the PPG-derived features changed between the hospital and free-living conditions. Thus, testing the classification performance in free-living conditions is essential to properly evaluate AF detection models.

Acknowledgements

This research was performed within the framework of the strategic joint research program on Data Science between TU/e and Philips Electronics Nederland B.V. The authors would like to thank N. Sturkenboom (MD), L. Verborg (MD), R. Eerdekens (MD), L. van den Heuvel, the personnel of the cardioversion department and the Holter department, and the Holter analysts of the Catharina Hospital for their help and contributions in the data collection.

References

[1] Camm AJ, Lip GYH, De Caterina R, Savelieva I, Atar D, Hohnloser SH, Hindricks G, Kirchhof P. 2012 focused update of the ESC Guidelines for the management of atrial fibrillation. Europace 2012;14:1385–413. ISSN 1532-2092.

[2] Chong JW, McManus DD, Chon KH. Arrhythmia discrimination using a smart phone. IEEE Journal of Biomedical and Health Informatics 2015;19(3):1–4. ISSN 21682194.

[3] Lemay M, Fallet S, Renevey P, Sol J, Pruvot E, Vesin JM. Wrist-Located Optical Device for Atrial Fibrillation Screening : A Clinical Study on Twenty Patients. Computing in Cardiology 2016;43.

[4] Corino VDA, Laureanti R, Ferranti L, Scarpini G, Lombardi F, Mainardi LT. Detection of atrial fibrillation episodes using a wristband device. Physiol Meas 2017;38:787–799.

[5] Bonomi AG, Schipper F, Eerikäinen LM, Margarito J, Aarts RM, Babaeizadeh S, Morree HMD, Dekker L. Atrial Fibrillation Detection using Photo-plethysmography and Acceleration Data at the Wrist. Computing in Cardiology 2016; 43.

[6] Dash S, Chon KH, Lu S, Raeder Ea. Automatic real time detection of atrial fibrillation. Annals of Biomedical Engineering 2009;37(9):1701–1709. ISSN 00906964.

[7] Richman JS, R. MJ. Physiological time-series analysis using approximate entropy and sample entropy. Am J Physiol Heart Circ Physiol 2000;278:H2039–H2049.

Address for correspondence:

Linda M. Eerikäinen
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
L.M.Eerikainen@tue.nl