How Accurately Can We Detect Atrial Fibrillation Using Photoplethysmography Data Measured in Daily Life?

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

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

Abstract

Photoplethysmography (PPG) is an unobtrusive measurement modality recently explored for the detection of atrial fibrillation (AF). When used in wrist-worn applications, PPG-monitoring can be used for long-term monitoring in daily life, which is beneficial when aiming to detect AF. The objective of this study was to investigate whether the performance of an AF detection model trained and tested on short measurements is generalizable to measurements in daily life. PPG, accelerometer, as well as reference ECG data were measured from 32 subjects (13 continuous AF, 19 no AF) in 24-hour monitoring in daily life. An AF detection model combining inter-pulse interval features was trained to classify AF or non-AF. Short measurements were obtained by selecting a 5-minute segment from each 24-hour recording and used for training the model. The accuracy was tested on both 5-minute segments and 24-hour data. Sensitivity, specificity, and accuracy of the model were 98.90%, 99.03%, and 98.98% with 5-minute data and 96.94%, 91.99%, and 93.91% with 24-hour data. False positive detections per patient worsened from being on average none during short recordings to (mean $\pm$ sd) 467 $\pm$ 328 in daily life. Thus, testing the AF detection models intended for long-term PPG-monitoring is essential with data from daily life in order to obtain a realistic estimate of the accuracy.

1. Introduction

Atrial fibrillation (AF) is a cardiac arrhythmia that affects approximately 3% of the adult population [1]. Early diagnosis of AF is essential because the arrhythmia increases the risk of stroke and heart failure. From a population health perspective, effective screening solutions are needed, but the challenges are in detecting asymptomatic and intermittent episodes of AF.

Studies with implantable devices have shown that increasing the monitoring period increases the percentage of subjects detected with AF in high risk populations [2, 3]. Currently, implantables are the only solution for monitoring on a long-term basis. These technologies are costly and require the implantation of the device. Therefore, the implantables cannot provide a solution for screening larger populations and other solutions that are unobtrusive and have a lower cost are needed.

Photoplethysmography (PPG) is an optical measurement modality that is used in wrist-worn wearables for heart rate monitoring. These devices can be worn for extended periods of time and could provide an unobtrusive long-term monitoring solution that would benefit AF screening. Detecting AF with wrist-worn PPG-based devices has been investigated in several studies with good detection performance [4–11].

In many studies about AF detection with wrist-worn devices, the devices are worn by study subjects for only a short period of time and the measurement setting is often in a controlled hospital environment [4–7]. Therefore, the studies give very little insight in the accuracy of AF detection during daily life. There are only few studies that show detection performances in 24-hour measurements [8–10] or longer [11]. In addition, a reduction in classification accuracy has been shown in daily life when features had been trained with data from patients having an electrical cardioversion [12] or when a model was tested in both settings [13]. In this work, our aim is to study how well a detection model trained with short measurements performs in daily life.

2. Methods

2.1. Data

The dataset used in the analysis consisted of 40 patients and was collected from patients that were scheduled for 24-hour Holter monitoring. Next to the 12-lead

 ISSN: 2325-887X DOI: 10.22489/CinC.2019.086

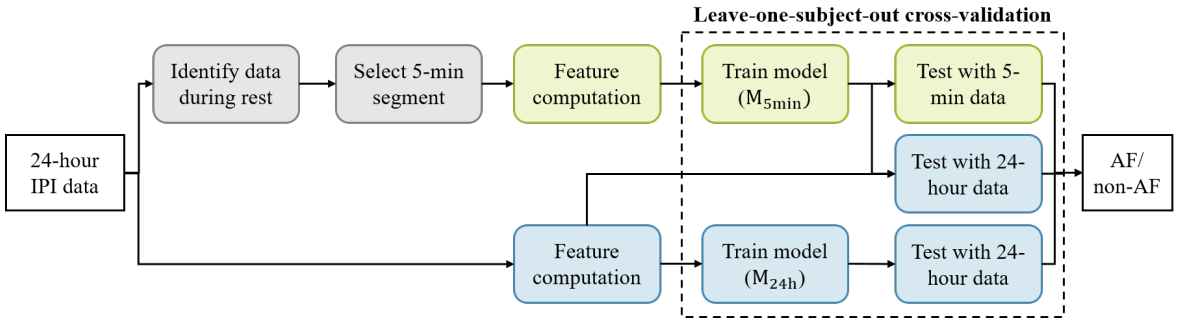

Figure 1. Flow chart of the inter-pulse interval (IPI) processing.

Holter monitor (H12+, Mortara, Milwaukee, WI, USA), the patients wore on their non-dominant arm a data logger, recording PPG in reflective mode and 3-axis accelerometer data with a Philips Cardio and Motion Monitoring Module (CM3 Generation-3, Wearable Sensing Technologies, Philips, Eindhoven). The sampling frequency of both PPG and accelerometry was 128 Hz and the dynamic range of the accelerometer was $\pm$ 8 g.

The ECG data was used as a reference and it was labeled beat-by-beat by an analyst using an automated rhythm detection software (Veritas, Mortara, Milwaukee, WI, USA). The labels were adjusted by the analyst when the automated labeling was not correct. The labels included normal, AF, premature atrial and ventricular contractions, paced, artifact, and unknown. During the 24-hour measurement period, patients were filling in a diary about their symptoms and activities and that was handed in when the monitoring period ended.

For the analysis, patients with atrial flutter, very noisy ECG reference, or very strong respiratory effect on heart rate variability during the night, were excluded. From the remaining 32 patients, 13 had continuous AF (age (years, m $\pm$ sd): 70 $\pm$ 9, males: 69%, BMI (kg/m^2, m $\pm$ sd): 28.3 $\pm$ 4.6) and 19 patients had no AF, but sinus rhythm and premature contractions (age, (years, m $\pm$ sd): 67 $\pm$ 13, males: 53%, BMI (kg/m^2, m $\pm$ sd): 27.9 $\pm$ 5.5).

2.2. Inter-pulse intervals

Inter-pulse intervals (IPIs) were extracted from the PPG signal for rhythm irregularity assessment to be used in the classification of AF and non-AF. The IPIs were calculated as the time difference between two consecutive pulses. For identifying individual pulses, the raw PPG data were first preprocessed by downsampling to 64 Hz and were then bandpass filtered to range from 0.3 to 5 Hz. Then the pulses were detected by identifying the fiducial points as described in [9]. The IPI time series were used to match the pulses to the labeled ECG beat times in order to have a label for each pulse [9]. The flow chart in Fig. 1 illustrates how the IPI series are processed after their extraction.

2.3. Short measurements

The objective of this study was to investigate how well a model trained with short measurements in a stable setting would perform when tested with data from an ambulatory setting. For selecting the short segments in a stable setting, the periods when the subject was resting and sleeping were identified. This was done manually by using the self-reported sleep and wake times and looking at the accelerometer data.

After the period during sleep was identified, a 5-minute segment was selected randomly from each subject. IPIs shorter than 200 ms and longer than 2200 ms were removed as outliers. After outlier removal, the sum of the remaining IPIs needed to be at least 90% of the length of the segment in order to consider the pulses being detected reliably and the setting being stable. If the segment did not meet the constraint, a new segment was selected.

2.4. Feature computation

Three features were used for the IPI-irregularity assessment in order to classify AF: the percentage of interval differences of successive intervals greater than 70 ms (pNN70), Shannon Entropy (ShE), and Sample Entropy (SampEn). They have been previously used for AF detection from PPG [4, 9, 12].

ShE and SampEn quantify in different ways the likelihood that a similar or regular pattern would not repeat in the sequence. For calculating ShE, first the IPI-values in the time series are divided into bins. The probability of the values in each bin is calculated as

$$p(i) = \frac{n_{(i)}}{l}, \tag{1}$$

$n(i)$ being the number of values in the bin i, l length of the sequence. ShE can be calculated from the probabilities as follows

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

where N is the number of bins. The number of bins used was 16 as in [9].

SampEn is the negative natural logarithm of the conditional probability that two sequences that match with each other at m points, the sequences also match when $m+1$ points are compared. The match is defined as a difference between two sequences being smaller than tolerance r. SampEn was calculated following [14] as

$$SampEn = -\ln(A/B) = -\ln(A) + \ln(B). \quad (3)$$

A is the number of matches with template length $m+1$ and B is the number of matches with length m, that was set to 1, and r was 0.25 times the standard deviation of the sequence as in [4].

The data was segmented into 30-second time windows of IPI-sequences with 20-second overlap and the three features were calculated for every time window.

The features were computed by using two different constraints and performances with the constraints were compared. In the first case, every window needed to include at least 20 IPIs, meaning that the heart rate could not be lower than 42 bpm. In the second case, no constraint was set for the window itself. However, SampEn would be only calculated if there were nine consecutive IPIs in the window [15] and therefore feature vectors without a value for SampEn were removed from the analysis. In addition, if more than half of the beats in the window were labeled as artifact according to the ECG, the features were not calculated.

2.5. Leave-one-subject-out cross-validation

The model was trained and the classification performance was tested by using leave-one-subject-out cross-validation. The classification was made by combining the features with logistic regression to give a probability for AF in each time window:

$$p_{AF}(t) = \frac{e^{X(t)\cdot b}}{1 + e^{X(t)\cdot b}}, \quad (4)$$

where t is the index of the window, $X(t)$ a vector containing the feature values for the window at time t, and b a vector of the model coefficients. The threshold for the probability to classify a window as AF was defined by the Youden index [16].

The model parameters were trained first with the 5-minute data of 31 subjects, leaving data of one subject for testing. The trained model was tested both with the 5-minute data of this subject as well as with the 24-hour data of this subject. This was repeated 32 times using data of each subject for testing once. For comparison, the same procedure was repeated by using the entire 24-hour data for training the model and testing the performance.

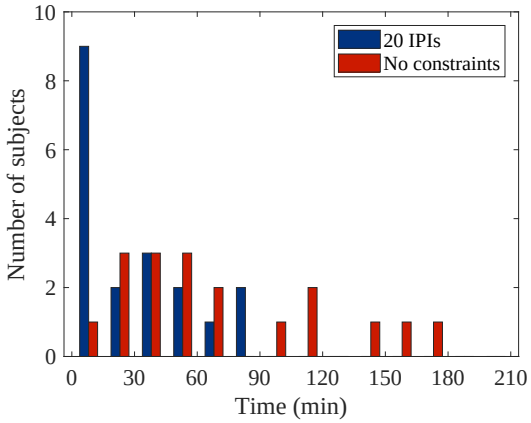

Figure 2. Durations of false positive detections.

3. Results

The aggregated results of the classification performance from the cross-validation are presented in Table 1. The measurement coverage for the 24-hour data when using the constraint of 20 IPIs was 49.5% and in the case of no constraints 63.3%.

The classification performance decreased when the model was tested with 24-hour data. With less constraints and higher coverage, the decrease was larger. Specificity and positive predictive value (PPV) were affected more than sensitivity and negative predictive value (NPV). When the models were tested with the 5-minute data segments, there were false positive AF detections with only two of the subjects. With the data from daily life, all of the subjects had false positive detections. On average there were 172 $\pm$ 92 false positives per subject when using the 20 IPIs constraint and 467 $\pm$ 328 when no constraint was used. Figure 2 shows how the total durations of the false positives per subject are distributed.

4. Discussion

The classification model achieved a high performance with 5-minute measurements independent of the constraint used. When testing the model with the data from daily life, the constraint used for calculating the features influenced both the classification performance and the measurement coverage. When more data was judged as analyzable, the performance decreased. This is in line with other studies [9, 10]. The difference in performance depending on the selected constraint was only observed with the data

Table 1. Classification performance. All metrics are presented as percentages (%).

	20 IPIs			No constraints		
	M_{5min} - 5 min	M_{5min} - 24 hours	M_{24h} - 24 hours	M_{5min} - 5 min	M_{5min} - 24 hours	M_{24h} - 24 hours
Sensitivity	98.82	97.74	96.48	98.90	96.94	94.58
Specificity	99.01	96.28	96.79	99.03	91.99	93.91
NPV	99.20	98.57	97.81	99.23	97.80	96.36
PPV	98.53	94.19	94.88	98.63	88.79	91.05
Accuracy	98.93	96.84	96.67	98.98	93.91	94.17

M_{5min} = model trained on 5-min data, M_{24h} = model trained on 24-h data, NPV = negative predictive value, PPV = positive predictive value.

from daily life. Therefore, measurements from daily life are needed in order to optimize the constraints.

5. Conclusion

The classification performance of an AF detection model developed with short measurements from a stable setting decreased and especially the number of false positives increased when tested with data from daily life. Therefore, testing AF detection models with daily life data is essential to have a realistic estimate of the detection accuracy and understanding the constraints required in order to achieve a high performance.

Acknowledgements

This research was performed within the framework of the Eindhoven MedTech Innovation Center (e/MTIC) in collaboration with Eindhoven University of Technology, Catharina Hospital Eindhoven, and Philips Research, and was partly funded by ITEA3 project, 15032, eWatch. The authors would like to thank everybody who contributed to the data collection.

References

[1] Kirchhof P, et al. 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. Europace 2016;18(11):1609–1678.

[2] Sanna T, et al. Cryptogenic stroke and underlying atrial fibrillation. New England Journal of Medicine 2014; 370(26):2478–86. ISSN 1533-4406.

[3] Reiffel JA, Verma A, Kowey PR, Halperin JL, Gersh BJ, Wachter R, Pouliot E, Ziegler PD. Incidence of previously undiagnosed atrial fibrillation using insertable cardiac monitors in a high-risk population: The REVEAL AF study. JAMA Cardiology 2017;2(10):1120–1127.

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

[5] Shashikumar SP, Shah AJ, Li Q, Clifford GD, Nemati S. A deep learning approach to monitoring and detecting atrial fibrillation using wearable technology. In IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). ISBN 9781509041794, 2017; 141–144.

[6] Fan YY, Li YG, Li J, Cheng WK, Shan ZL. Diagnostic performance of a smart device with photoplethysmography technology for atrial fibrillation detection: pilot study (Pre-mAFA II registry). JMIR mHealth and uHealth 2019; 7(3):1–11.

[7] Dörr M, Nohturfft V, Brasier N, Bosshard E, Djurdjevic A, Gross S, Raichle CJ, Rhinisperger M, Stöckli R, Eckstein J. The WATCH AF Trial: SmartWATCHes for detection of atrial fibrillation. JACC Clinical electrophysiology 2019; 5(2):199–208. ISSN 2405-5018.

[8] Bonomi AG, et al. Atrial fibrillation detection using a novel cardiac ambulatory monitor based on photoplethysmography at the wrist. Journal of the American Heart Association 2018;7(15). ISSN 20479980.

[9] Eerikäinen LM, Bonomi AG, Schipper F, Dekker LRC, Vullings R, de Morree HM, Aarts RM. Comparison between electrocardiogram- and photoplethysmogram-derived features for atrial fibrillation detection in free-living conditions. Physiological Measurement 2018;39(8).

[10] Sološenko A, Petrėnas A, Sörnmo L, Paliakaitė B, Marozas V. Detection of atrial fibrillation using a wrist-worn device. Physiological Measurement 2019;40:025003.

[11] Wasserlauf J, You C, Patel R, Valys A, Albert D, Passman R. Smartwatch performance for the detection and quantification of atrial fibrillation. Circulation Arrhythmia and Electrophysiology 2019;12(6):1–9. ISSN 1941-3149.

[12] Eerikäinen LM, Dekker L, Bonomi AG, Vullings R, Schipper F, Margarito J, de Morree HM, Aarts RM. Validating features for atrial fibrillation detection from photoplethysmogram under hospital and free-living conditions. Computing in Cardiology 2017;3–6.

[13] Tison GH, et al. Passive detection of atrial fibrillation using a commercially available smartwatch. JAMA Cardiology 2018;3(5):409–416. ISSN 23806591.

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

[15] De Mazumder D, Lake DE, Cheng A, Moss TJ, Guallar E, Weiss RG, Jones SR, Tomaselli GF, Moorman JR. Dynamic analysis of cardiac rhythms for discriminating atrial fibrillation from lethal ventricular arrhythmias. Circulation Arrhythmia and Electrophysiology 2013;6(3):555–561.

[16] Youden WJ. Index for rating diagnostic tests. Cancer 1950; 3(1):32–35. ISSN 10970142.

Address for correspondence:

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