Atrial Fibrillation Detection from PPG Interbeat Intervals via a Recurrent Neural Network

Jérôme Van Zaen[1], Elsa Genzoni[1,2], Fabian Braun[1], Philippe Renevey[1], Etienne Pruvot[3], Jean-Marc Vesin[2], Mathieu Lemay[1]

[1] Swiss Center for Electronics and Microtechnology (CSEM), Neuchâtel, Switzerland
[2] Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
[3] Lausanne University Hospital (CHUV), Lausanne, Switzerland

Abstract

Atrial fibrillation (AF) affects millions of individuals worldwide and can lead to serious complications such as stroke or heart failure. This arrhythmia is difficult to diagnose with ambulatory electrocardiogram monitors in the early stages due to its transient nature. Recent advances in wearable photoplethysmographic (PPG) devices are promising for screening AF in large populations as they are relatively comfortable and can be worn over long periods of time. Herein, we propose a system to detect AF from PPG recordings. This system is composed of a beat detector to extract interbeat intervals and a classifier for detection. We trained the classifier on a large public database of interbeat intervals and then evaluated the whole system on PPG recordings collected during catheter ablation procedures. We achieve an accuracy of 0.986 for the detection of AF with a sensitivity and specificity of 1.0 and 0.978 respectively. These metrics compare favorably with existing systems.

1. Introduction

Atrial fibrillation (AF) is the most common cardiac arrhythmia affecting 1–2% of the population [1], and whose prevalence increases with age, ranging from less than 0.5% at 40–50 years to 5–15% at 80 years. AF can lead to severe complications [2] and is associated with a 3–5-fold increased risk of stroke, a 2-fold increased risk of mortality [3] and a 3-fold increased risk of heart failure [4]. This arrhythmia is caused by disorganized atrial activation resulting in rapid and irregular heart rate. This irregularity may cause palpitations, shortness of breath and fainting. However, approximately one third of the patients are asymptomatic [5], impeding early diagnosis and treatment which might protect the patient from the consequences of AF and stop its progression. Indeed, AF causes structural and electrical remodeling of the atria which promotes future episodes, i.e. AF begets AF [6].

The gold standard for the diagnosis of AF is the 12-lead electrocardiogram (ECG). After reviewing ECG and patient's history, a trained electrophysiologist can select the most appropriate therapy. However, the transient and sometimes asymptomatic nature of early stage AF limits the detection performance of ambulatory ECG recording systems. Holter monitors can collect long-term ECG recordings but they are usually limited to 24–48 hours due to their cumbersomeness. Wearable photoplethysmographic (PPG) devices appear promising for AF screening in large populations. Indeed, they can be worn as simple wrist watches for several days, and are relatively low-cost. Nevertheless, reviewing long recordings is time-consuming and several approaches tend to automatize AF detection on both ECG [7] and PPG [8] signals. Herein, we propose a system to detect AF from PPG recordings that combines a beat detector to extract interbeat intervals (IBIs) and a recurrent neural network (RNN) to classify AF to be embedded in a wearable device. This RNN was trained on a large dataset of IBIs from PhysioNet [9].

2. Methods

2.1. Datasets

We used two datasets to develop and evaluate our system for detecting AF form PPG data. The first one is the Long-Term AF Database [5] from PhysioNet. This database includes 84 long-term ECG records of patients with paroxysmal or sustained AF. Each record contains a two-lead ECG signal, beat labels, and rhythm annotations. As this dataset does not contain PPG data, it is used solely to train the classifier. We used the beat labels to compute IBIs and divided them into 30-second windows without overlap. We excluded windows that were not exclusively labeled with normal sinus rhythm (NSR) or AF. We also discarded the windows with less than 12 IBIs indicative of a heart rate

 ISSN: 2325-887X DOI: 10.22489/CinC.2019.084

under 24 bpm. This resulted in a dataset of 206380 windows. Note that due to IBI variations, the number of values per window was not constant. The next step was to split the dataset into subsets for training, validation, and testing. All windows from the last 28 records (identifiers larger than or equal to 100) were put aside for testing ($n = 70022$). A 80%/20% split stratified by label was used to divide the remaining windows in a training set ($n = 109086$) and a validation set ($n = 27272$).

The second dataset was collected during a clinical study at the Lausanne University Hospital (CHUV) and was used to evaluate our system. This study included 21 patients referred for catheter ablation of cardiac arrhythmias. For each patient, a 12-lead ECG was recorded at 2 kHz with a commercial electrophysiology system (Siemens Sensis) during the ablation procedure. Simultaneously, green light PPG and tri-axis accelerometer signals were sampled at 21.33 Hz with a proprietary wrist-worn device. An expert reviewed the 12-lead ECG recordings and labeled the heart rhythms in two classes: AF and no-AF. The no-AF class includes NSR, sinus bradycardia, sinus tachycardia, and atrial tachycardia. Other rhythms, such as bigeminy, frequent ectopic beats, atrial fluttering, and ventricular tachycardia, were excluded from the analysis. All activities in this study were carried out in compliance with local regulations and the Declaration of Helsinki.

2.2. Beat Detection

The beat detector is composed of two finite impulse response filters. The first filter is a low-pass filter with a cutoff at 2 Hz to smooth the PPG signal and is used to detect segments of interest between waveform maxima. The second filter is a differentiator to compute the first derivative of the PPG signal. Then, pulse times are extracted by searching for derivative maxima in the segments identified by the first filter. Finally, IBIs are computed from the pulse times. In addition, a flag is associated with each IBI to indicate if it is an outlier depending on waveform morphology and motion estimated from accelerometer signals.

2.3. AF Classification

We selected an RNN to classify AF from IBIs as this class of networks can process sequences with different lengths. As the objective was to develop a system that can be embedded into a wearable device, we used a simple architecture with two layers: a gated recurrent unit (GRU) layer [10] with 8 units to take into account windows with different numbers of IBIs and a sigmoid layer to output an estimated probability of AF. This RNN has a total of 249 parameters (240 for the GRU layer and 9 for the sigmoid layer). To improve generalization performance, we applied dropout [11] and L_2 regularization to the GRU layer.

To facilitate convergence, we centered and scaled the IBIs with mean and standard deviation measured on the training set. We trained the RNN by minimizing the cross-entropy with the Adam algorithm [12] for 30 epochs. When the cross-entropy evaluated on the validation set did not decrease for a given number of epochs, the learning rate was reduced by a factor.

After defining the RNN architecture and the training procedure, we applied Bayesian optimization to tune the following hyper-parameters: the dropout rate, the factor for L_2 regularization, the batch size, the initial learning rate, and the factor and the number of epochs without improvements to wait for reducing the learning rate. The best hyper-parameters were selected by monitoring the classification accuracy evaluated on the validation set. The RNN was then trained with the tuned hyper-parameters and evaluated by computing the accuracy, sensitivity, specificity, and F_1 score on the training, validation, and test sets. The training pipeline was implemented in Python with the Keras package for neural networks [13].

2.4. AF Detection

Our system for detecting AF from PPG data combines the beat detector and the RNN classifier described above. The beat detector extracts IBIs and associated flags to indicate outliers from a PPG signal. The cumulative sum of successive IBIs is then computed until it exceeds 30 seconds. At this point, the corresponding sequence of IBIs without outliers is fed to the classifier. This corresponds to IBI windows slightly longer than 30 seconds similar to the ones used to train the RNN. If more than 20% of the IBIs are outliers, no prediction is made to avoid errors due to poor signal quality or motion and the window is marked as undecidable. Otherwise, the probability of AF estimated by the RNN, p, is used to make a prediction with the following rule:

$$\text{Prediction} = \begin{cases} \text{AF} & \text{if } p > 0.7, \\ \text{no-AF} & \text{if } p < 0.3, \\ \text{undecidable} & \text{otherwise.} \end{cases}$$

The rationale for this rule is to output a prediction only when the RNN clearly favors one of the two classes. Once a prediction is made (or not depending on the number of IBI outliers and the estimated AF probability), the RNN is reset and a new cycle of around 30 seconds is started.

Discarding windows due to IBI outliers or because the AF probability is between 0.3 and 0.7 means that the PPG signal is not fully used. However, it should not compromise the detection of AF in long recordings. Indeed, such recordings include many 30-second windows and AF typically lasts for several minutes. The thresholds on the number of outliers and the estimated AF probability were cho-

Table 1. RNN classification performance evaluated on the Long-Term AF Database.

Metric	Training	Validation	Test
Accuracy	0.9891	0.9924	0.9784
Sensitivity	0.9881	0.9999	0.9894
Specificity	0.9907	0.9805	0.9676
F_1 score	0.9911	0.9939	0.9784

sen as they provide a good trade-off between reliable prediction and the number of undecidable windows.

The AF detection system was evaluated on the clinical dataset of PPG and accelerometer data described above by measuring the accuracy, sensitivity, specificity, and F_1 score. To assess the importance of outlier rejection, we also computed the performance metrics without discarding windows including more than 20% of IBI outliers. However, the windows with AF probability p between 0.3 and 0.7 were still marked as undecidable. The proportion of undecidable windows was recorded for both cases.

3. Results

The first step was to train the RNN to classify AF from IBIs on the Long-Term AF Database. After running Bayesian optimization to tune the hyper-parameters, the dropout rate was set to 0.1224, the factor for L_2 regularization to 10^{-9}, the batch size to 10, and the initial learning rate to 0.0057. In addition, when the cross-entropy did not decrease for 2 epochs on the validation set, the learning rate was reduced by a factor of 0.8. After training the RNN with these hyper-parameters, it achieved an AF classification accuracy of 0.9784 on the test set. All performance metrics evaluated on the training, validation, and test sets are reported in Table 1. It is worth noting that dropout and regularization helped to limit overfitting as the performance on the test set is only slightly degraded compared to the training and validation sets.

After training the RNN, we applied the whole AF detection system to PPG and accelerometer signals from the second dataset recorded during catheter ablation procedures. This resulted in 1473 windows of IBIs. Of these windows, 360 were labeled as AF and 1113 as no-AF. An example of beat detection from a PPG signal during an AF episode is shown in Figure 1. More than 50% of the windows were marked as undecidable by our system. However, the accuracy on the remaining windows was 0.9860. Without outlier rejection, the proportion of undecidable windows dropped to around 5% but the accuracy decreased to 0.9294. There were still a few undecidable windows corresponding to the case where the AF probability p was between 0.3 and 0.7. All performance metrics, obtained with and without outlier rejection, are reported in Table 2. One

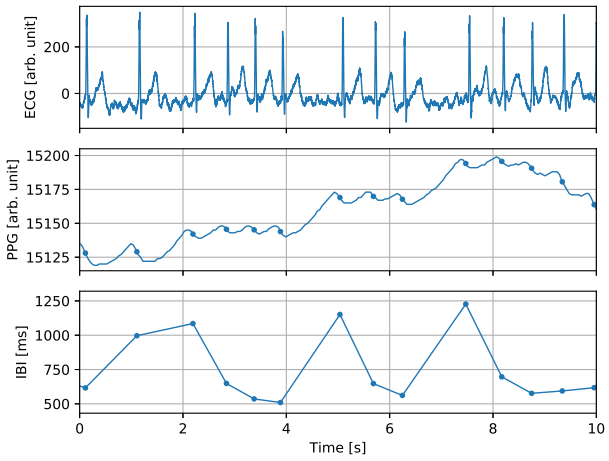

Figure 1. Beat detection example: (top) lead I ECG signal, (middle) PPG signal with detected beats denoted by dots, (bottom) IBIs derived from PPG beats.

Table 2. AF detection system performance evaluated on PPG and accelerometer data.

Metric	With outlier rejection	Without outlier rejection
Accuracy	0.9860	0.9294
Sensitivity	1.0	0.9801
Specificity	0.9781	0.9122
F_1 score	0.9809	0.8753
Undecidable	0.5153	0.0577

can observe that rejecting outliers has a considerable impact on the classification accuracy. This highlights the importance of using robust signal quality indicator and avoid predicting AF or no-AF in the case of motion or insufficient signal quality.

4. Discussion

Despite the limited number of parameters of our model, the RNN could classify AF and NSR with an accuracy just below 0.98 on the test set. The sensitivity and specificity were also close to 0.99 and 0.97, meaning that these two heart rhythms can be reliability discriminated using IBIs time series. The two strategies implemented to prevent overfitting, dropout and regularization, proved effective as the decrease in performance observed between the training and validation sets on the one hand and the test set on the other hand was limited. Bayesian optimization was also helpful to tune the hyper-parameters of the classifier.

Once the RNN was trained, it was included in the system for detecting AF from PPG signals. When evaluating this system on a dataset of PPG and accelerometer data, all performance metrics were close to one. In particular, the accuracy was just below 0.99 and there was no false

no-AF detection. Rejecting IBIs outliers appears crucial to obtain good AF detection performance. Indeed, the accuracy dropped below 0.93 and the specificity was around 0.91 when outliers were not rejected. In other words, the number of false positives increased significantly. There is certainly an application-dependent trade-off to find between processing all IBIs and rejecting too many IBIs. Importantly, although the RNN was trained with two heart rhythms (AF and NSR), the no-AF class in the PPG dataset included sinus bradycardia, sinus tachycardia, and atrial tachycardia in addition to NSR. These different rhythms did not seem to affect the performance of the system.

Taken together, these results suggest that it is possible to train a classifier for AF on a large dataset of IBIs extracted from ECG data and then apply it to IBIs extracted from PPG data. Such an approach is advantageous since many ECG databases with arrhythmias are freely available while similar PPG databases are scarce. Herein, only IBIs were used for classification, while other features extracted from PPG signals, such as waveform morphology, might also be essential for reliable detection. Our system bears some limitations. First, it is limited to the detection of two heart rhythms: NSR and AF. While AF is the most common arrhythmia, there are many other arrhythmias that could be treated more efficiently if they were detected early. Second, the sizes of the two datasets we used for development were limited. The first one includes a large number of 30-second windows but the number of different patients is still relatively limited while the second dataset is even smaller with 21 patients and shorter recordings. One of the main obstacles to reliable arrhythmia detection from PPG is the lack of large databases containing various types of heart rhythms.

5. Conclusion

We propose a system composed of a beat detector and a classifier to detect AF from PPG and accelerometer data. The classifier was trained on a large dataset of IBIs extracted from ECG and the whole system was evaluated on a smaller dataset of PPG data. The high performance metrics indicate that such a system can be used to screen AF in large populations as it can easily be embedded in a wearable device. Furthermore, the reported sensitivity and specificity compare favorably with the values obtained by FibriCheck (0.953 and 0.965) [14] and AliveCor (0.944 and 0.994) [15].

References

[1] Camm AJ, et al. Guidelines for the management of atrial fibrillation. European Heart Journal 2010;31(19):2369–2429.

[2] January CT, Wann LS, Alpert JS, Calkins H, Cigarroa JE, Cleveland JC, Conti JB, et al. 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the ACC/AHA task force on practice guidelines and the HRS. Journal of the American College of Cardiology 2014;64(21):e1–e76.

[3] Kannel WB, Wolf PA, Benjamin EJ, Levy D. Prevalence, incidence, prognosis, and predisposing conditions for atrial fibrillation: population-based estimates. The American journal of cardiology 1998;82(7):2N–9N.

[4] Wang TJ, Larson MG, Levy D, Vasan RS, Leip EP, Wolf PA, D'Agostino RB, et al. Temporal relations of atrial fibrillation and congestive heart failure and their joint influence on mortality: the Framingham heart study. Circulation 2003;107(23):2920–2925.

[5] Petrutiu S, Sahakian AV, Swiryn S. Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans. Europace 2007;9(7):466–470.

[6] Wijffels MCEF, Kirchhof CJHJ, Dorland R, Allessie MA. Atrial fibrillation begets atrial fibrillation. Circulation 1995; 92(7):1954–1968.

[7] De Chazal P, O'Dwyer M, Reilly RB. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. IEEE Transactions on Biomedical Engineering 2004;51(7):1196–1206.

[8] Lemay M, Fallet S, Renevey P, Solà J, Leupi C, Pruvot E, Vesin JM. Wrist-located optical device for atrial fibrillation screening: a clinical study on twenty patients. In Computing in Cardiology. 2016; 681–684.

[9] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 2000; 101(23):e215–e220.

[10] Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv e prints 2014;arXiv:1406.1078.

[11] Gal Y, Ghahramani Z. A theoretically grounded application of dropout in recurrent neural networks. In Advances in Neural Information Processing Systems. 2016; 1019–1027.

[12] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv e prints 2014;arXiv:1412.6980.

[13] Chollet F, et al. Keras. https://keras.io, 2015.

[14] Proesmans T, Mortelmans C, Van Haelst R, Verbrugge F, Vandervoort P, Vaes B. Mobile phone–based use of the photoplethysmography technique to detect atrial aibrillation in primary care: diagnostic accuracy study of the FibriCheck app. JMIR Mhealth Uhealth 2019;7(3):e12284.

[15] Haberman ZC, Jahn RT, Bose R, Tun H, Shinbane JS, Doshi RN, Chang PM, Saxon LA. Wireless smartphone ECG enables large-scale screening in diverse populations. Journal of Cardiovascular Electrophysiology 2015;26(5):520–526.

Address for correspondence:

Jérôme Van Zaen
CSEM, Rue Jaquet-Droz 1, 2002 Neuchâtel, Switzerland
jerome.vanzaen@csem.ch