Atrial fibrillation detection from raw photoplethysmography waveforms: A deep learning application

Kirstin Aschbacher, PhD,[*][†][1] Defne Yilmaz, BA,[*][1] Yaniv Kerem, MD,[‡]
Stuart Crawford, PhD,[‡] David Benaron, MD,[‡] Jiaqi Liu, BS,[‡]
Meghan Eaton, MSN, PNP-BC,[‡] Geoffrey H. Tison, MD, MPH,[*][§] Jeffrey E. Olgin, MD,[*]
Yihan Li, DPhil,[‡][2] Gregory M. Marcus, MD, MAS, FHRS[*][2]

From the [*]Division of Cardiology, Department of Medicine, University of California, San Francisco, San Francisco, California, [†]Department of Psychology, University of California, San Francisco, San Francisco, California, [‡]Jawbone Health, San Francisco, California, and [§]Bakar Computations Health Sciences Institute, University of California, San Francisco, San Francisco, California.

BACKGROUND Atrial fibrillation (AF), a common cause of stroke, often is asymptomatic. Smartphones and smartwatches can detect AF using heart rate patterns inferred using photoplethysmography (PPG); however, enhanced accuracy is required to reduce false positives in screening populations.

OBJECTIVE The purpose of this study was to test the hypothesis that a deep learning algorithm given raw, smartwatch-derived PPG waveforms would discriminate AF from normal sinus rhythm better than algorithms using heart rate alone.

METHODS Patients presenting for cardioversion of AF (n = 51) were given wrist-worn fitness trackers containing PPG sensors (Jawbone Health). Standard 12-lead electrocardiograms over-read by board-certified cardiac electrophysiologists were used as the reference standard. The accuracy of PPG signals to discriminate AF from sinus rhythm was evaluated by conventional measures of heart rate variability, a long short-term memory (LSTM) neural network given heart rate data only, and a deep convolutional-recurrent neural net (DNN) given the raw PPG data.

RESULTS From among 51 patients with persistent AF (age 63.6 ± 11.3 years; 78% male; 88% white), we randomly assigned 40 to train

and 11 to test the algorithms. Whereas logistic regression analysis of heart rate variability yielded an area under the receiver operating characteristic curve (AUC) of 0.717 (sensitivity 0.741; specificity 0.584), the LSTM model given heart rate data exhibited AUC of 0.954 (sensitivity 0.810; specificity 0.921), and the DNN model given raw PPG data yielded the highest AUC of 0.983 (sensitivity 0.985; specificity 0.880).

CONCLUSION A deep learning model given the raw PPG-based signal resulted in AF detection with high accuracy, performing better than conventional analyses relying on heart rate series data alone.

KEYWORDS Artificial intelligence; Atrial fibrillation; Heart rate sensor; Machine learning; Mobile health; Smartwatch; Photoplethysmography; Wearable

Introduction

Atrial fibrillation (AF) is a leading cause of stroke and increases the risk of myocardial infarction, chronic kidney disease, dementia, and mortality.[1–5] Although anticoagulant therapy may mitigate these risks, clinically occult AF frequently conceals evidence of the disease until one of

---

KEY FINDINGS

■ A machine learning model fed raw photoplethys-mography (PPG) waveform data seems to more accurately discriminate atrial fibrillation from sinus rhythm compared to conventional heart rate variability measurements.

■ A machine learning model fed raw PPG waveform data seems to more accurately discriminate atrial fibrillation from sinus rhythm compared to a machine learning model using heart rate information alone.

■ Restriction to sedentary individuals undergoing cardioversion in this study may not apply to ambulatory free-living individuals in the community.

---

those complications first becomes apparent.[6] Therefore, approaches to accurately detect asymptomatic AF are needed.

Digital technology now offers tools to facilitate AF detection outside of traditional clinical settings. Many smartwatches and fitness trackers measure continuous heart rate data based on photoplethysmography (PPG). PPG uses optical sensors to detect changes in the blood volume of tissue microvasculature in the finger (eg, using a smartphone camera) or wrist (eg, using a wearable wrist-band). Because AF has a characteristic irregularly irregular pulse, it may be particularly amenable to detection using such a sensor. Whereas the initial studies relied on conventional measures of heart rate variability (similar to what is used in clinically applicable electrocardiographic [ECG] algorithms),[7] the application of deep learning algorithms to PPG waveforms is particularly promising, as deep learning algorithms can learn highly predictive models from raw data.[8]

We previously demonstrated that a deep neural net fed serial heart rate measurements derived from the PPG sensor on Apple Watches could accurately discriminate AF from sinus rhythm.[9] However, a low positive predictive value was observed, particularly in an ambulatory population, which would translate into a substantial number of false-positive results if applied broadly to the population. Therefore, further efforts to enhance the accuracy of AF detection are needed. A major potential advantage of deep learning algorithms is that they can ingest "raw" signals (such as the complete PPG waveform), thereby omitting the need for extensive preprocessing and feature extraction. Deep learning models have the potential to learn representations from PPG waveforms across multiple domains, including time, frequency, and morphology, without being given explicit formulas. Hence, we sought to determine how well a wrist-worn fitness tracker using a deep learning algorithm fed continuous raw waveform PPG data could reliably detect AF.

Methods

Study design

This study obtained PPG waveforms using a wearable wristband device (Jawbone Health, San Francisco, CA) from an in-person cohort of patients receiving electrical cardioversion, and these waveforms were used to build an AF prediction model. The model was supervised by labels that were first generated by application of standard Muse (GE Healthcare, Chicago, IL) software to ECG data and confirmed by board-certified cardiac electrophysiologists. ECGs were recorded in 10-second intervals before and immediately after the shock(s) administered during each electrical cardioversion procedure. All participants provided written informed consent before enrollment. This study was approved by the institutional review board of the University of California, San Francisco (UCSF).

Study sample

We enrolled 51 consecutive, consenting, English-speaking patients scheduled to undergo electrical cardioversion for AF at UCSF between December 15, 2017, and July 20, 2018. Patients were excluded if they exhibited atrial arrhythmias other than AF at the time of enrollment or had ventricular pacing >80% of the time.

Data collection procedure

Patients undergoing electrical cardioversion were sedated and remained supine during the study. Before cardioversion, a 12-lead ECG was obtained, and an application-activated Jawbone wrist-worn prototype fitness tracker was applied to the participant's wrist for at least 20 continuous minutes before electrical cardioversion. During the procedure, a study coordinator recorded the exact time of shock administered, total energy delivered (in joules), success of the procedure, and, if successful, the time of transition from AF to normal sinus rhythm as determined by a concomitant continuous 12-lead ECG. After the procedure, the Jawbone wrist device was removed at least 20 minutes after the final shock was administered.

We sought to enroll at least 50 patients based on our past experience validating a smartwatch to detect AF among patients undergoing cardioversion (at which conventional approaches such as those discussed later also were used).[9] We used an 80%/20% train/test split for the machine learning algorithm. One additional patient was enrolled during the study period, for a total of 51 patients. Data from 40 randomly selected patients (80% of the minimum planned of 50) were used to train the models, and data from the remaining 11 patients were used to test the models. In total, this resulted in 72 total hours of continuous PPG data (47 hours of AF and 25 hours of sinus rhythm). In addition, to better differentiate AF from normal sinus rhythm, a second de-identified dataset from 13 individuals without known arrhythmia during sleep, yielding 91 hours of PPG recordings, was incorporated into the training model. We contrasted 3 models using the area under the receiver operating characteristic curve (AUC) as the primary performance metric: (1) a

"traditional model" using heart rate variability as a predictor in a logistic regression model; (2) a single-layer, long short-term memory (LSTM) neural network fed a series of 35 consecutive heartbeat periods as the sole input; and (3) a deep convolutional-recurrent neural net (DNN) using the raw PPG waveform as the input.

Statistical analysis

Model 1: Conventional approach
To evaluate a "conventional approach" used in previous studies,[7] we computed the root mean square of the successive interval differences (RMSSD), an established measure of heart rate variability. R-R intervals were identified using a previously validated adaptive multiscale peak detection algorithm applied to a bandpass-filtered optical signal.[10] The normality of the distribution was verified through visual inspection. RMSSD was calculated on each sequence of 35 heartbeats, constituting a time window of approximately 30 seconds. In order to maximize the comparability of performance metrics across all 3 models, we fit the logistic regression coefficients in the training set and reported the final performance indicators from the test set of UCSF patients. Model performance was based on the AUC (primary) and sensitivity and specificity (secondarily) in the test set.

Model 2: LSTM neural net
A single-layer LSTM neural network was given a training sample consisting of a sequence of 35 consecutive heartbeat periods, which roughly corresponded to a 30-second time window, and required fitting of exactly 4385 parameters.[11] Heartbeat periods were extracted from the bandpass-filtered waveform using a modified multiscale peak detection algorithm,[10] which measured the distance between consecutive minima in the optical absorption.

Model 3: DNN
To evaluate the predictive value of a deep learning approach utilizing the raw PPG waveform data, we used a convolutional-recurrent neural network. An earlier iteration of this approach has been previously described in detail.[12] In brief, the architecture (Supplemental Figure 1) involved multiple convolutional layers, each followed by max-pooling, then an LSTM[11] layer, and finally several flattened layers, resulting in roughly 10,000 parameters to estimate. Model hyperparameters were chosen with minimal cross-validation. Notably, no specifically extracted amplitude or morphologic features were used, although in theory, this DNN model has the capacity to "learn" features within the time, frequency, or morphologic domains. As an input, the model ingested raw PPG signals sampled at 20 Hz and then outputted a sequence of calibrated, instantaneous probabilities.

Results
Participant characteristics
Table 1 characterizes the 51 patients with AF who underwent cardioversion. There were no substantial differences in

Table 1 Baseline characteristics of participants undergoing cardioversion

Baseline characteristics	Train (n = 40)	Test (n = 11)	P value
Mean age (y)	62.8 ± 11.0	65.6 ± 14.4	.48
Male	30 (75%)	10 (91%)	.26
White	36 (90%)	9 (82%)	.46
Body mass index (kg/cm^2)	30.2 ± 6.2	30.0 ± 6.3	.91
Medical characteristics			
Hypertension	20 (50%)	8 (73%)	.18
Diabetes mellitus	10 (25%)	1 (9%)	.26
Coronary artery disease	4 (10%)	0 (0%)	.27
Congestive heart failure	4 (10%)	1 (9%)	.91
Obstructive sleep apnea	17 (43%)	4 (36%)	.71
Myocardial infarction	3 (8%)	1 (9%)	.86
Cardiomyopathy	4 (10%)	1 (9%)	.93
Valvular heart disease	2 (5%)	0 (0%)	.45
Chronic obstructive pulmonary disease	1 (3%)	0 (0%)	.60
Previous cardioversion	19 (48%)	6 (55%)	.68
Stroke	2 (5%)	1 (9%)	.61
Treatment characteristics			
Beta-blocker	22 (55%)	9 (82%)	.11
Antiarrhythmic drug	25 (63%)	4 (36%)	.12
Anticoagulant drug	38 (95%)	11 (100%)	.45
Procedural characteristics			
No. of shocks	1.3 ± 0.9	1.1 ±0.3	.22
Successful cardioversion	34 ± 0.4	9 ± 0.4	.80
Joules delivered	306.5 ± 312	242.7 ± 117	.30

Demographic, medical, and procedural characteristics for all participants in train and test sets undergoing cardioversion are listed. Values are given as mean ± SD or n (%), unless otherwise indicated.

patient or procedural characteristics between those selected for training and those selected for testing. Figure 1 illustrates PPG waveforms from a randomly chosen participant during representative sections of the pre- vs post-cardioversion periods. Specific rhythm characteristics after cardioversion for each patient are shown in Supplementary Tables 1 and 2. Figure 2 illustrates the raw PPG waveform from 1 patient against the DNN-generated probability that a given PPG segment was classified as AF. It also illustrates the power spectra across component frequencies.

Model selection
The AUC for discrimination of AF from sinus rhythm in the test dataset was higher in the DNN model using the raw PPG waveform (AUC = 0.983) than in the traditional logistic regression model using RMSSD (AUC = 0.717) or the LSTM model fed only heart rate data (AUC = 0.954) (Table 2 and Figure 3). Sensitivity and specificity values were computed using a discrimination cutoff of 0.5 (Table 2). Figure 4 offers qualitative insight into how the potential morphology differences between PPG segments labeled as AF vs sinus rhythm might potentially combine with heart rate information to inform prediction.

Discussion
A decade ago, the National Heart, Lung, and Blood Institute recommended that one of the most important challenges in cardiovascular research was the use of emerging technologies, such as wearables, for early AF detection.[13] Herein, we differentiate episodes of AF vs sinus rhythm with near-perfect prediction by applying a convolutional-recurrent

neural network algorithm to raw, wearable-based PPG data gathered from patients undergoing cardioversion. Moreover, the ability to obtain strong predictions with the passively collected, raw waveform is both novel and potentially advantageous. Therefore, PPG waveforms derived from a wearable wristband device, combined with deep learning algorithms, may provide useful tools for population screening of occult AF.

The DNN model using raw PPG waveforms performed substantially better than a traditional approach using heart rate variability among a set of UCSF patients presenting for cardioversion for persistent AF. Moreover, the DNN also performed somewhat better than a single-layer LSTM neural network algorithm using the beat-to-beat heart rate data alone,[7] which was derived from the PPG waveforms. Previous studies utilizing PPG data via either a smartphone camera or a smartwatch to detect AF used only an analysis of the heart rate to identify AF.[7,9] In both cases, the model performance was somewhat similar to that observed in the current study when constrained to the heart rate data only, albeit lower than the performance reported here when compared to an analysis using the raw waveform data. Of note, a previous study of the Cardiio Rhythm Mobile application, which used a support vector machine classification of rhythm irregularity features derived from PPG waveforms, was not as high-performing as the neural network-based models reviewed here.[14] Because PPG-based tools may one day provide a population screening tool for occult AF, it is important to note that the positive predictive value of the DNN in this study (95%) was higher than reported in previous studies.[15,16] It is particularly crucial for artificial intelligence–based AF screening tools to have a high positive predictive
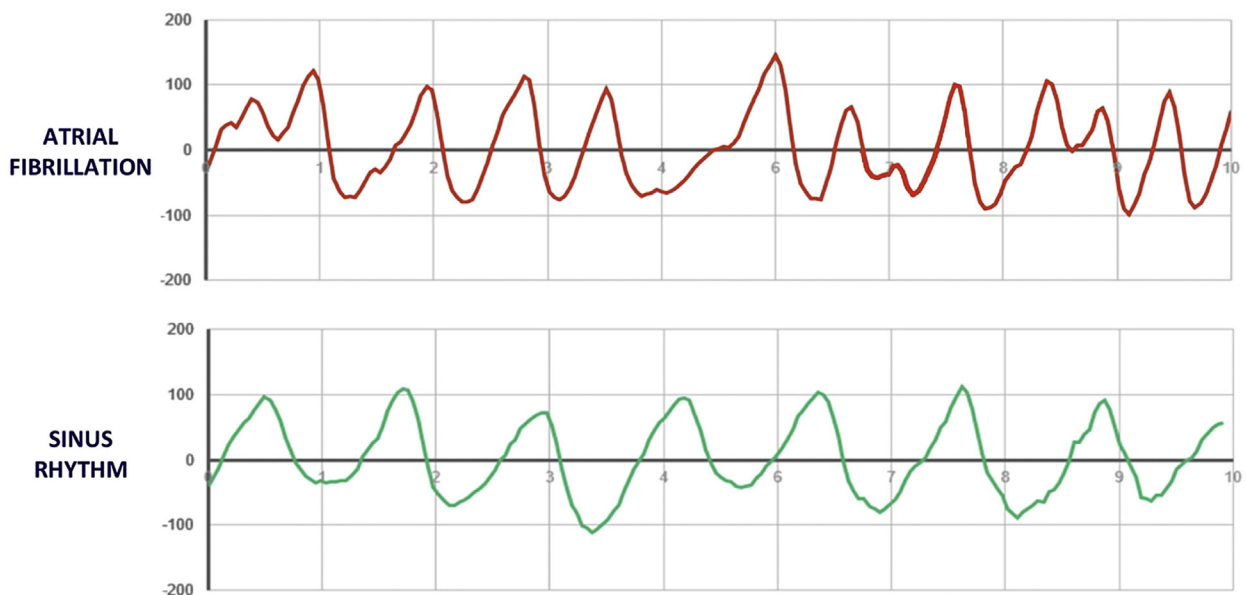

Figure 1 Raw photoplethysmography (PPG) waveforms of a randomly selected patient in both atrial fibrillation (AF) and normal sinus rhythm. Shown are waveform tracings from the raw PPG output of a single participant in AF before cardioversion and in normal sinus rhythm after cardioversion. **Top:** The morphology of the AF waveform displays the characteristically "irregularly irregular" pattern, with irregularities in both wave amplitude and period, which underlies all AF arrhythmias as measured on standard 12-lead electrocardiograms. **Bottom:** In contrast, the morphology of the sinus rhythm waveform is more uniform, both in amplitude and period.
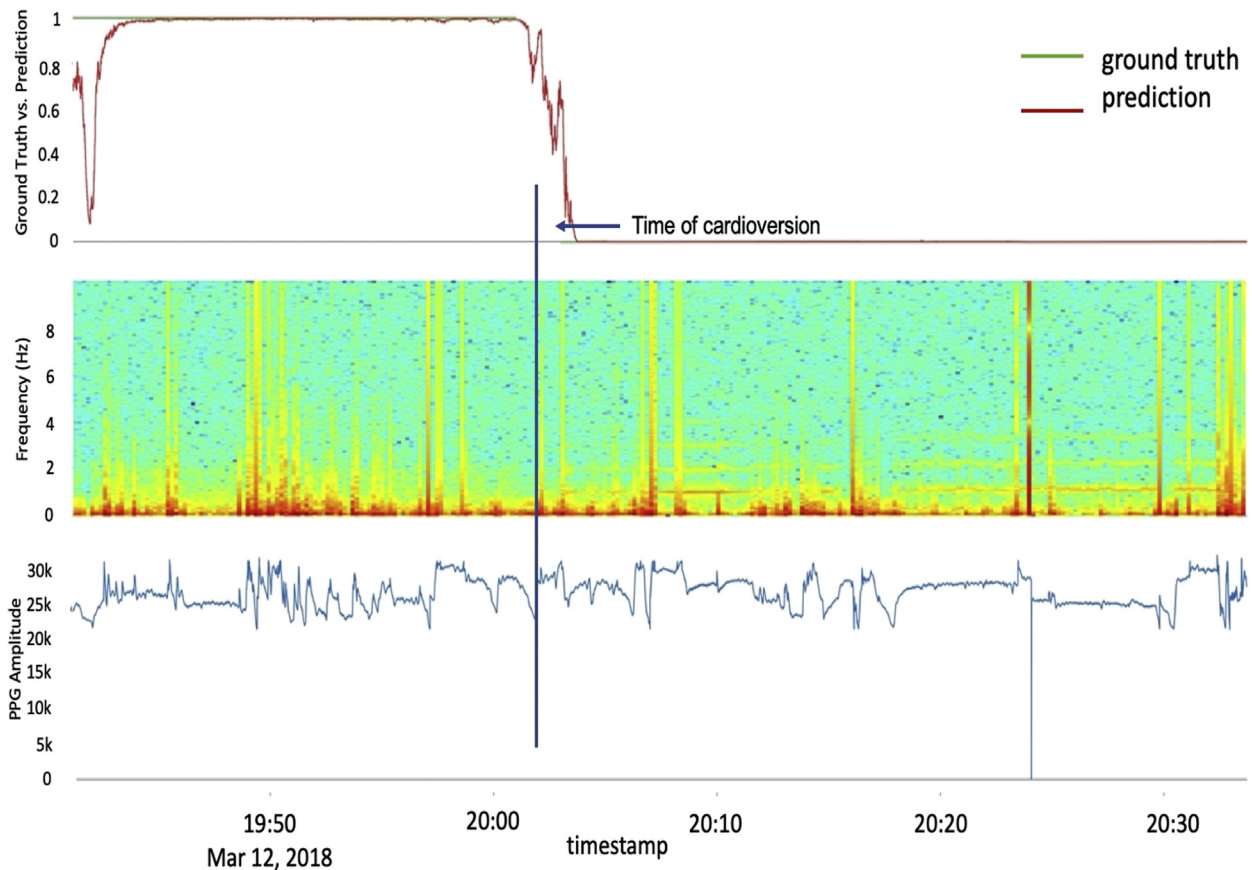
Figure 2   Deep neural network model predictions plotted against the actual photoplethysmography (PPG) waveform and frequency analysis. **Top:** Model prediction outcome for a sample participant, using the deep convolutional-recurrent neural network, in which outcomes at or above 0.5 are classified as atrial fibrillation and those below 0.5 as normal sinus rhythm. **Middle:** Spectrogram with consecutive Fourier transformation on the raw optical data, illustrating the power spectra across component frequencies. These spectra illustrate (as expected) the greater preponderance of higher frequencies and spectral variability pre- vs post-cardioversion. **Bottom:** Raw PPG optical stream at 20 Hz.

value in order to minimize the number of false positives and thereby reduce unnecessary patient visits, anxiety, and costs.

Furthermore, the ability to utilize the raw PPG waveform is an important and novel discovery of this study, which could have other important implications beyond model performance. Whereas the LSTM is limited to learning features within the time domain, the DNN has the theoretical capability to identify other features, such as amplitude or morphology, as well as high-level interactions between various domains. In addition, previous methods for analyzing PPG have often required exten-

sive signal preprocessing and feature engineering (eg, peak detection and extraction of heart rate series); therefore, the ability to analyze raw PPG data is a novel discovery.

In terms of platforms and technology, both wearable and smartphone camera-derived PPG waveforms have yielded strong AF detection models.[7,9] Although smartphone cameras initially might seem preferable for PPG waveform acquisition due to convenience and ubiquity, wristbands may have distinct advantages. AF detection studies have observed relatively high rates of

Table 2   Performance characteristics of 3 models in the cardioversion patient test set

Algorithm type	AUC	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
Conventional heart rate variability* (model 1)	0.717	74.1	58.4	80.8	48.8
Machine learning fed heart rate only data[†] (model 2)	0.954	81.0	92.1	96.0	67.1
Machine learning fed raw waveform data[‡] (model 3)	0.983	98.5	88.0	95.1	96.2

Model evaluation indices are given for each of the 3 models applied to the test set of patients undergoing cardioversion.

AUC = area under the receiver operating characteristic curve; NPV = negative predictive value; PPV = positive predictive value.

*Using the root mean square of the successive interval differences.

[†]Using a long short-term memory algorithm.

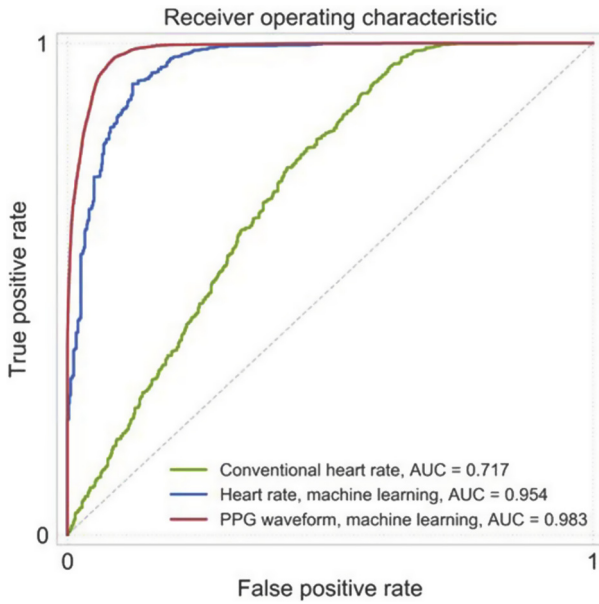[‡]Using a deep convolutional-recurrent neural network algorithm.

Figure 3   Model performance characteristics for conventional heart rate, heart rate analyzed by deep learning, and the raw photoplethysmography (PPG) waveforms analyzed by deep learning. The area under the receiver operating characteristic curve (AUC), measured in the test set, is reported for each of the 3 models. The *green line* represents the conventional logistic regression model fed the root mean square of successive differences, a conventional assessment of heart rate variability. The *blue line* represents the long short-term memory model fed a heart rate data series. The *red line* represents the deep learning model (a convolutional-recurrent neural network) fed the raw PPG waveform data.

nonadherence or dropout,[17] so passive collection methods that minimize engagement efforts may be crucial. The fact that the neural networks tested herein performed well with relatively few patients relies partially on the ability to collect more data per patient, which is facilitated by wristband-enabled passive collection.

Study limitations

These results were gathered using a wearable wristwatch, which may not be economically accessible to a subset of the population at high risk for undetected AF due to poor access to medical care; nonetheless, the use of smartphones and wearables with PPG technology is becoming more ubiquitous over time, even in developing nations.[18] The majority of study participants were white, and the optical PPG sensors may behave differently in patients of other races or ethnicities with different skin colors or tones. In addition, the study only recruited individuals with a known diagnosis of AF; therefore, we made no identifications of new diagnoses of AF with the model, and its accuracy in detecting other arrhythmias cannot currently be determined. Finally, although excellent test characteristics were observed among sedentary cardioversion patients, these results were obtained in a controlled setting and may not be applicable to ambulatory individuals. Because the model was supervised to distinguish AF from sinus rhythm, further testing is needed to identify the extent to which the model may tend to falsely classify noise, premature beats, and/or atrial tachyarrhythmias as AF. Indeed, premature atrial contractions were fairly commonly observed in these patients, although they were not quantified specifically to determine the precise influence on the accuracy of each AF detection method. Future studies will be important to elucidate the effects of premature atrial contractions, particularly on false-positive results, before these algorithms can be reliably deployed in ambulatory free-living individuals.

Conclusion

This study demonstrates that a deep learning algorithm can use a raw PPG signal, without feature engineering or extensive preprocessing, to detect AF with very high accuracy. Moreover, this novel approach was superior to standard
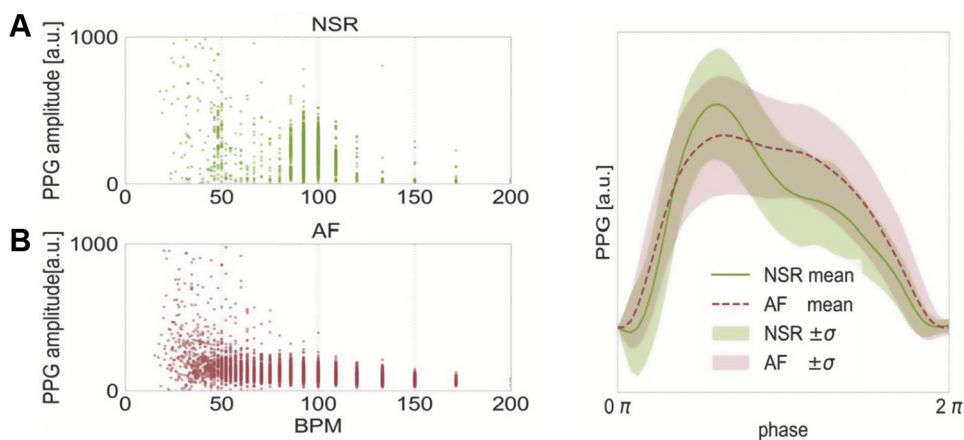

Figure 4   Waveform morphology of a single patient in atrial fibrillation (AF) and normal sinus rhythm (NSR). Morphologic differences are shown. **Left:** Variability in heart rate and photoplethysmography (PPG) amplitude for a randomly selected patient who was successfully cardioverted, during AF (**A**) and NSR (**B**). AF (*red* in **B**) is characterized by greater heart rate variability and lower amplitude. The x-axis represents beats per minute (BPM), which is calculated as the inverse of the time between consecutive minima. The y-axis is the PPG amplitude in arbitrary units (a.u.). **Right:** Morphologic characteristics of the patient's mean heart rate cycle, using resampling to equate the time domain. It is visually evident that the morphology differed during AF and NSR for this representative patient.

methods that rely on heart rate variability and standard statistical approaches and outperformed a machine learning algorithm that was fed PPG-derived heart rate data alone. The success of the deep learning algorithms exhibited herein suggests that AF detection tools are now ready for the challenge of bedside-to-field translation. AF detection tools, fueled by artificial intelligence and wearables, have the potential to mitigate cardiovascular morbidity risks through the identification and delivery of early treatment to millions of individuals suffering from undiagnosed AF.

Appendix
Supplementary data
Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.hroo.2020.02.002.

References

1. Soliman EZ, Safford MM, Muntner P, et al. Atrial fibrillation and the risk of myocardial infarction. JAMA Intern Med 2014;174:107–114.
2. Dukes JW, Marcus GM. Atrial fibrillation begets myocardial infarction. JAMA Intern Med 2014;174:5–7.
3. Bansal N, Fan D, Hsu CY, Ordonez JD, Marcus GM, Go AS. Incident atrial fibrillation and risk of end-stage renal disease in adults with chronic kidney disease. Circulation 2013;5:569–574.
4. Jacobs V, Cutler MJ, Day JD, Bunch TJ. Atrial fibrillation and dementia. Trends Cardiovasc Med 2015;25:44–51.
5. Benjamin EJ, Muntner P, Alonso A, et al. Heart disease and stroke statistics—2019 update: a report from the American Heart Association. Circulation 2019;139:e56–e528.
6. Sanna T, Diener H-C, Passman RS, et al. Cryptogenic stroke and underlying atrial fibrillation. N Engl J Med 2014;370:2478–2486.
7. McManus DD, Lee J, Maitas O, et al. A novel application for the detection of an irregular pulse using an iPhone 4S in patients with atrial fibrillation. Heart Rhythm 2013;10:315–319.
8. LeCun Yann, Yoshua B, Hinton G. Deep learning. Nature 2015;521:436–444.
9. Tison GH, Sanchez JM, Ballinger B, et al. Passive detection of atrial fibrillation using a commercially available smartwatch. JAMA Cardiol 2018;3:409–416.
10. Scholkmann F, Boss J, Wolf M. An efficient algorithm for automatic peak detection in noisy periodic and quasi-periodic signals. Algorithms 2012;5:588–603.
11. Hochreiter S, Urgen Schmidhuber J. Long short-term memory. Neural Comput 1997;9:1735–1780.
12. Gotlibovych I, Crawford S, Goyal D, et al. End-to-end deep learning from raw sensor data: atrial fibrillation detection using wearables. ArXiv 2018: arXiv:1807.10707.
13. Benjamin EJ, Chen PS, Bild DE, et al. Prevention of atrial fibrillation: report from a National Heart, Lung, and Blood Institute workshop. Circulation 2009;119:606–618.
14. Rozen G, Vaid J, Hosseini SM, et al. Diagnostic accuracy of a novel mobile phone application for the detection and monitoring of atrial fibrillation. Am J Cardiol 2018;121:1187–1191.
15. Chan PH, Wong CK, Poh YC, et al. Diagnostic performance of a smartphone-based photoplethysmographic application for atrial fibrillation screening in a primary care setting. J Am Heart Assoc 2016;5:e003428.
16. Nemati S, Ghassemi MM, Ambai V, et al. Monitoring and detecting atrial fibrillation using wearable technology. Proc Annu Int Conf IEEE Eng Med Biol Soc EMBS 2016;3394–3397.
17. Mela T. Smartwatches in the fight against atrial fibrillation: the little watch that could. J Am Coll Cardiol 2018;71:2389–2391.
18. Poushter J. Smartphone ownership and internet usage continues to climb in emerging economics. Pew Research Center. February 22, 2016. Available from: https://www.pewresearch.org/global/2016/02/22/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies/. Accessed August 13, 2018.