JACC: CLINICAL ELECTROPHYSIOLOGY VOL. 10, NO. 2, 2024

ORIGINAL RESEARCH

AI/ML - ARRHYTHMIA MONITORING

Continuous Atrial Fibrillation Monitoring From Photoplethysmography

Comparison Between Supervised Deep Learning and Heuristic Signal Processing

Pavel Antiperovitch, MD,[a] David Mortara, PhD,[a] Joshua Barrios, PhD,[a,b] Robert Avram, MD, MSc,[a,c,d] Kimberly Yee, BS,[a] Armeen Namjou Khaless, MS,[a] Ashley Cristal, MD,[a] Geoffrey Tison, MD, MPH,[a,b,*] Jeffrey Olgin, MD[a,*]

ABSTRACT

BACKGROUND Continuous monitoring for atrial fibrillation (AF) using photoplethysmography (PPG) from smartwatches or other wearables is challenging due to periods of poor signal quality during motion or suboptimal wearing. As a result, many consumer wearables sample infrequently and only analyze when the user is at rest, which limits the ability to perform continuous monitoring or to quantify AF.

OBJECTIVES This study aimed to compare 2 methods of continuous monitoring for AF in free-living patients: a well-validated signal processing (SP) heuristic and a convolutional deep neural network (DNN) trained on raw signal.

METHODS We collected 4 weeks of continuous PPG and electrocardiography signals in 204 free-living patients. Both SP and DNN models were developed and validated both on holdout patients and an external validation set.

RESULTS The results show that the SP model demonstrated receiver-operating characteristic area under the curve (AUC) of 0.972 (sensitivity 99.6%, specificity: 94.4%), which was similar to the DNN receiver-operating characteristic AUC of 0.973 (sensitivity 92.2, specificity: 95.5%); however, the DNN classified significantly more data (95% vs 62%), revealing its superior tolerance of tracings prone to motion artifact. Explainability analysis revealed that the DNN automatically suppresses motion artifacts, evaluates irregularity, and learns natural AF interbeat variability. The DNN performed better and analyzed more signal in the external validation cohort using a different population and PPG sensor (AUC, 0.994; 97% analyzed vs AUC, 0.989; 88% analyzed).

CONCLUSIONS DNNs perform at least as well as SP models, classify more data, and thus may be better for continuous PPG monitoring. (J Am Coll Cardiol EP 2024;10:334–345) © 2024 The Authors. Published by Elsevier on behalf of the American College of Cardiology Foundation. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

From the [a]Division of Cardiology, Department of Medicine and Cardiovascular Research Institute, University of California-San Francisco, San Francisco, California, USA; [b]Bakar Computational Health Sciences Institute, University of California-San Francisco, San Francisco, California, USA; [c]Montreal Heart Institute, Department of Medicine, University of Montreal, Montreal, Quebec, Canada; and the [d]Heartwise.ai Laboratory, Montreal, Quebec, Canada. *Drs Tison and Olgin contributed equally to this work and are co-senior authors.

The authors attest they are in compliance with human studies committees and animal welfare regulations of the authors' institutions and Food and Drug Administration guidelines, including patient consent where appropriate. For more information, visit the Author Center.

Manuscript received August 16, 2023; revised manuscript received October 19, 2023, accepted October 24, 2023.

ISSN 2405-500X https://doi.org/10.1016/j.jacep.2024.01.008

Atrial fibrillation (AF) is the most common sustained cardiac arrhythmia, which can impair patients' quality of life, and is a major cause of hospital admissions, stroke, and mortality.[1,2] Even after initiation of treatment, ongoing objective monitoring for recurrence of AF is important for patient management, keeping in mind that one-third of individuals with AF are asymptomatic and in general there is poor correlation between symptoms and onset/duration of AF episodes.[3,4]

Smartwatches have emerged as a potential tool for continuous AF monitoring and detection. These wearable devices are able to evaluate pulse wave and rate data using photoplethysmography (PPG)—a noninvasive sensor that measures reflected light from the surface of the skin that captures volumetric variations in capillary blood flow.[5] PPG-based AF detection technology utilizing PPG signal and other onboard sensors for opportunistic sampling or noise rejection was developed by several smartwatch manufacturers including Samsung,[6] Apple,[7] and Fitbit[8] and validated with various degrees of rigor. However, the reported performance characteristics are affected by selection bias: studies use highly selected patients for validation and discard a large proportion of signals or sampling time due to noise. For example, the Apple Heart Study evaluated heart rhythm using a noncontinuous tachogram measure, which performs 1-minute recordings every 2 hours, and only when the user is still—this equates to about 12 min/d of AF monitoring (<1% of the day).[8] The Fitbit (Google) algorithm only sampled when the user was still, which was an average of ~8 hours each day.[8] Our previous report using Samsung watch data excluded 32% of data due to poor signal quality, which is lower than for any previously published smartwatch.[6,9] Other published signal processing (SP) heuristics and deep learning models were also challenged by limited sampling,[6-8,10-13] low specificity,[14-16] and the high percent of signals rejected due to noise.[6,8,17-19] Many of these studies were limited by selection bias: the Apple and Fitbit studies only evaluated patients already flagged positive by their AF algorithm, leading to an unknown amount of missed patients with AF, and many other published studies only evaluated patients asleep or at rest undergoing cardioversion.[13,14,20-22] Overall, none of the previously published models evaluated enough user-level signal to be considered "near continuous," nor did they include free-living patients for an unbiased validation, both of which are required for highly accurate continuous around-the-clock outpatient monitoring for AF using a smartwatch.

Here, we report the largest known dataset of simultaneous continuous ECG and PPG recordings from ambulatory free-living patients that are annotated by cardiologists. From this, we developed a SP algorithm based on a well-validated best-performing heuristic that measures irregularity (entropy).[12] We compared the SP model with a convolutional deep neural network (DNN) that is trained on raw PPG signal and externally validated these models against a sample of ambulatory PPG data simultaneously obtained with ECG telemetry. We report the performance of each model as well as the amount of signal data analyzed with the goal of near-continuous monitoring of AF (see **Central Illustration**).

ABBREVIATIONS AND ACRONYMS

AF = atrial fibrillation
DNN = deep neural network
IBI = interbeat interval
ECG = electrocardiography
IRN = irregular rhythm notification
NPV = negative predictive value
PPG = photoplethysmography
PPV = positive predictive value
ROC-AUC = receiver-operating characteristic-area under the curve
SP = signal processing

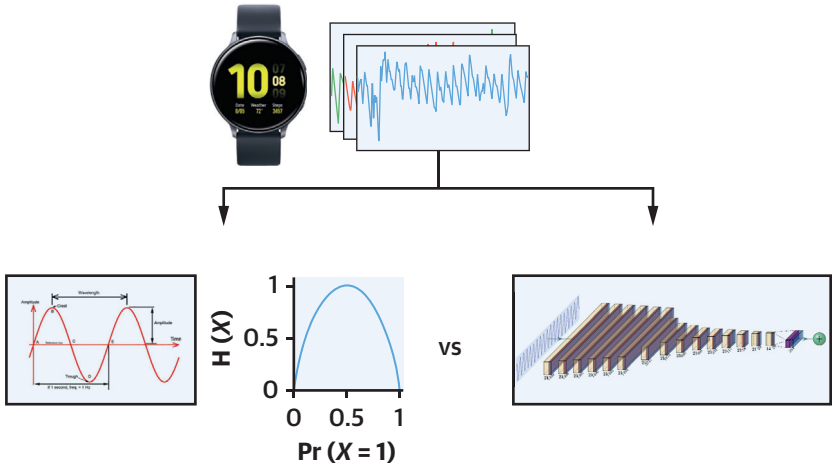
METHODS

STUDY SAMPLE. The protocol for patient recruitment and data collection was reported previously.[6] The study was approved by the University of California-San Francisco Institutional Review Board, and consent was obtained from all participants. Participants with a self-reported history of AF or at least 1 risk factor for AF were invited electronically using the Eureka Research Platform,[23] from the Health eHeart Study, which is a large observational cohort of >300,000 participants (see the Supplemental Methods for details). Upon enrollment, participants were shipped a study kit containing two 14-day BioTel ePatches and a Samsung Galaxy Active 2 smartwatch along with instructions. Instructions were also provided using the Eureka Research app, along with push notifications and SMS messages reminding them to wear the watch and ECG patch during the entire 4-week period. The total monitoring time was up to 28 days (2× 14-day patches), after which the ePatches were returned by mail. Upon return of the ePatches, the ECG data were converted to ISHNE format for input into the University of California-San Francisco's Signal Processing Core using CER-S (Continuous ECG Recording Suite) (AMPS-LLC). AF episodes were identified by a technician and overread by a board-certified cardiologist. Atrial flutter episodes were considered non-AF.

TRAINING DATASET GENERATION. ECG and PPG signals were time aligned and each 5-minute segment labeled as AF, normal rhythm, or other based on the ECG reading. ECG labels with no available time-aligned PPG recording within a 3-minute search window or those with >80% of missing signal were

JACC: CLINICAL ELECTROPHYSIOLOGY VOL. 10, NO. 2, 2024
FEBRUARY 2024:334–345

CENTRAL ILLUSTRATION **Comparison of Performance of a Robust Signal Processing Algorithm to a DNN Machine Learning Algorithm for Detection of AF From PPG**



	Signal Processing Heuristic	Deep Neural Network
% Signal Classified	62%-88%	95%-97%
ROC-AUC	0.97-0.99	0.97-0.99
Sensitivity	**91%-99%**	90%-94%
Specificity	95%-98%	**96%-99%**

Antiperovitch P, et al. J Am Coll Cardiol EP. 2024;10(2):334–345.

A robust signal processing heuristic and convolutional deep neural network were compared in classification of atrial fibrillation on photoplethysmography smartwatch tracings. Overall performance was similar between the 2 approaches; however, the deep neural network was able to classify substantially more signals, demonstrating its superior tolerance of motion artifact. AF = atrial fibrillation; DNN = deep neural network; PPG = photoplethylsmography; ROC-AUC = receiver-operating characteristic area under the curve.

discarded (~15% of data). This yielded a total of 847,033 5-minute ECG-labeled PPG samples that were used for training/validation on 204 participants. The data were then split by participants into training, validation, and holdout groups in a 70:15:15 split. All data analysis was performed on participants assigned to the holdout group.

EXTERNAL DATASET GENERATION. We randomly selected 50 ambulatory inpatients who had telemetry recordings containing simultaneous ECG and PPG telemetry data between July 1, 2022, and July 15, 2022, at the University of California, San Francisco Medical Center (Philips IntelliVue MX40 device). Up to 5 days of telemetry data were selected per patient depending on availability. Telemetry data were processed using CER-S and manually annotated by a cardiac electrophysiologist. The opinion of a second cardiac electrophysiologist was obtained if the cardiac rhythm was not clear at initial review. Tracings were partitioned into 5-minute PPG segments, resampled to 25 Hz, scaled, and paired with ECG-adjudicated labels. Segments with >80% of missing signal were discarded (ie, PPG sensor was off).

DNN DEVELOPMENT AND VALIDATION. The DNNs were developed using a Python 3.9 development environment (Python Software Foundation) using PyTorch library suite (Meta AI). A large number of DNN architectures were tested; however, convolutional neural networks provided superior performance, similar to prior findings.[13] A series of high-density architectural and hyperparameter sweeps were conducted using Weights & Biases software.[14] The best architecture for AF detection with minimal memory footprint involved 15 blocks of convolutional layers, batch normalization, and dropout. The model was trained as a binary classifier using Binary Cross Entropy Loss and Adam optimizer, with sigmoid output (0 = "no AF" and 1 = "AF"). For the purposes of comparison, the DNN cutoffs were selected in order to maximize the amount of data

TABLE 1 Baseline Characteristics of Patients (N = 202)

Age	
<65 y	104 (51.5)
65-75 y	67 (33.2)
>75 y	31 (15.3)
Sex	
Male	108 (53.5)
Female	94 (46.5)
Race	
Caucasian	181 (89.6)
Hispanic	8 (4.0)
Black or African American	8 (4.0)
Asian	8 (4.0)
American Indian or Alaska Native	3 (1.5)
Other	4 (2.0)
History of AF	
No history	9 (4.4)
Paroxysmal	175 (86.6)
Persistent	16 (7.9)
Symptoms of AF	
Symptoms of AF at time of study	27 (13.4)
Daily	9 (4.5)
Weekly	50 (24.8)
Monthly	40 (19.8)
Within 1 y	39 (19.3)
More than a year	15 (7.4)
Never aware	13 (6.4)
Comorbidities	
History of HTN	105 (52.0)
History of CHF	31 (15.3)
Previous MI	19 (9.4)
History of coronary artery disease	40 (19.8)
History of TIA or stroke	20 (9.9)
Diabetes	28 (13.9)
Obstructive sleep apnea	82 (40.6)
CHA_2DS_2-VASc score	
0-1	69 (34.2)
2-4	110 (54.5)
5-7	16 (7.9)

Values are n (%).
AF = atrial fibrillation; CHF = congestive heart failure; HTN = hypertension; MI = myocardial infarction; TIA = transient ischemic attack.

classified while achieving ideal sensitivity/specificity. The DNN-standard model classified AF if sigmoid output was >0.9, no-AF if sigmoid output was <0.1, and the rest uncertain. The DNN-lower uncertainty threshold had cutpoints of 0.99 and 0.01.

SP ALGORITHM DEVELOPMENT. The SP algorithm was created based on Zhao at al.[24] The signal was processed using a bandpass filter to reduce the impact of low- and high-frequency artifact. Pulse detection was accomplished by locating peaks in filtered PPG signal. Timing of pulses was refined fitting a quadratic to 3 samples surrounding the peak and noting the peak time of the quadratic curve to an accuracy of 1 millisecond. Peaks with amplitudes more than twice the running average and interbeat interval (IBI) less than the average IBI by more than 160 milliseconds or with a baseline shift larger than twice the running average pulse amplitude were discarded as being artifact. Segments were classified as artifact if they contained fewer than 100 IBIs or if the sum of valid IBIs was <150 seconds. Entropy of IBIs over a 5-minute sliding window was calculated using the Shannon technique that was previously reported[24]; Shannon entropy values >100 (S=100) or >80 (S=80) were used as thresholds for AF. Because atrial/ventricular ectopy is a common cause for false positives, a specific ectopy detection function was implemented to reject these signals as non-AF (Supplemental Methods, SP Ectopy Detection).

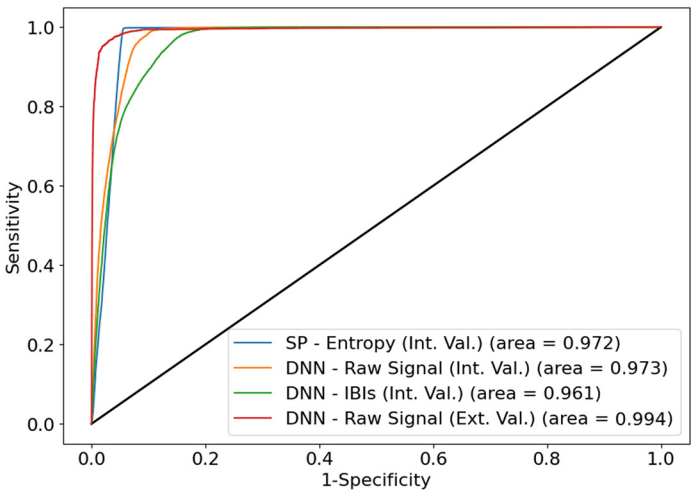
DNN EXPLAINABILITY. Model explainability analysis for the DNN was performed using 3 methods: input augmentation, input generation, and LIME (Local Interpretable Model-Agnostic Explanations).[25] For input augmentation, the model was trained using IBIs obtained from identified peaks while excluding waveform information (SciPy.signal module). For input generation, we generated synthetic signal segments with engineered features and performed inference using a pretrained DNN to understand the relative importance of each feature. Features tested include random vs nonrandom IBIs, random IBIs drawn from a normal probability distribution of varying mean and standard deviation, and PPG segments containing various proportions/order of both sinus rhythm and AF.

RESULTS

A total of 847,033 labels and corresponding PPG tracings were used for testing and training; a separate 120,900 samples from 31 (~15%) participants were withheld for final validation and were not used in the design, optimization, or training of models. The holdout samples had 18% AF; baseline characteristics are outlined in **Table 1**.

INTERNAL VALIDATION. The SP model demonstrated an entropy receiver-operating characteristic-area under the curve (ROC-AUC) of 0.972 (sensitivity of 99.6%, specificity of 94.4%, positive predictive value [PPV] 76.2% and negative predictive value [NPV] 99.9%) for AF detection, and it classified 62% of the signal; the remaining signal was labeled as uncertain (**Figure 1**, **Table 2**). The DNN model trained on raw PPG signal revealed a similar ROC-AUC of 0.973 (sensitivity 92.2%, specificity 95.5%, PPV 76.0%, NPV 98.7%) but classified a larger portion of data (95%).

FIGURE 1 Receiver-Operating Characteristic Curves



Performance of 2 deep neural network (DNN) models on internal validation (Int. Val.) and external validation (Ext. Val.). The DNN–raw signal model was trained on the raw signal, whereas the DNN–interbeat interval (IBI) model was trained on IBIs only. SP = signal processing.

The DNN cutpoints were then adjusted to increase performance at the expense of amount of data classified (DNN-lower uncertainty threshold) and revealed a sensitivity of 98.0% and specificity of 99.9% (PPV 87.0%, NPV 99.9%) (Table 2) but reduced the amount of data classified to 79%; notably, even with this stricter cutpoint, the DNN still classified considerably more data than the SP model (Table 2).

EXTERNAL VALIDATION. A total of 29,878 segments were generated from 50 ambulatory inpatients on telemetry with simultaneous PPG and ECG signals, of which 5,843 (19.0%) demonstrated AF in 21 patients (7 paroxysmal AF, 14 permanent AF). The SP model classified AF with a ROC-AUC of 0.989 (sensitivity 96.9%, specificity 96.3%, PPV 97.0%, NPV 99.2%), and 88% of samples were analyzed (Table 3). Evaluating the performance of the DNN on this external dataset revealed a ROC-AUC of 0.994 (sensitivity 90.1%, specificity 99.7%, PPV 98.6%, NPV 97.9%), and 97% of data were classified. Lowering the threshold for uncertainty (DNN-lower uncertainty threshold) raised the sensitivity to 94.0% and the specificity to 99.9% (PPV 99.2%, NPV 99.3%), at the expense of a lower amount of data classified (86%). Analyzing the results grouped by patient, both the DNN-raw and SP models identified AF in every patient who had experienced AF. However, the DNN-raw falsely detected AF in 1 of 29 participants who did not have AF, whereas SP falsely detected AF in 16 of 29. The only false positive classified by the DNN was a patient with periods of sinus arrhythmia and frequent premature atrial contractions. False positives from the SP method included tracings with poor signal quality (11 users), frequent premature atrial contractions/premature ventricular contractions (4 users), and sinus arrhythmia (2 users).

TABLE 2 Operating Characteristics of Models on Holdout Data (Internal Validation)

	% of Data Classified	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	ROC-AUC
Signal processing (S = 100)	62	94.2	94.9	77.1	99.9	0.972
Signal processing (S = 80)[a]	62	99.6	94.4	76.2	99.9	
DNN–standard[b]	95	92.2	95.5	76.0	98.7	0.973
DNN–lower uncertainty threshold[c]	79	98.0	99.9	87.0	99.9	
DNN–IBI	88	82.6	97.4	79.2	97.9	0.961

[a]S is the entropy cutpoint, and with a lower entropy cutpoint there is increased sensitivity of AF detection. [b]Model was evaluated using cutoffs >0.9 for AF, <0.1 for non-AF, and outputs 0.1 to 0.9 as uncertain. [c]Same as the DNN–standard model, except that outputs were adjusted to maximize classification accuracy in preference to % of data classified: sigmoid output of >0.99 for AF and <0.01 for non-AF, and outputs between these values are uncertain.

AF = atrial fibrillation; DNN = deep neural network; IBI = interbeat interval; NPV = negative predictive value; PPV = positive predictive value; ROC-AUC = receiver-operating characteristic area under the curve.

DNN EXPLAINABILITY ANALYSIS. Several approaches were used to understand how the DNN functions. First, we trained the same DNN with just IBI input, which was derived by preprocessing the signal with a peak detection algorithm. This DNN-IBI revealed a ROC-AUC of 0.961 on internal validation and 0.988 on external validation, which was less performant than the DNN trained on raw signal (0.973 and 0.994, respectively), and created more uncertain classifications (12% vs 5%). This suggests that the DNN trained on raw signal is better at rejecting noise and/or identifying relevant parts of the signal. Implementing other heuristics such as second-order derivatives and filtering of the signal did not improve DNN performance.

Second, LIME explanation analysis was used to identify regions of the DNN that were important for prediction. We found that the model learned to distinguish noise from biological signal and focused on noise-free areas of the tracing while suppressing areas of high-frequency motion artifact (Figure 2A). When performing LIME evaluation on tracings that were noise-free, the model used the majority of the tracing to generate a prediction (Figure 2B).

Third, we investigated model performance on tracings that contain both AF and sinus rhythm by splicing the PPG signal containing AF and sinus rhythm in different ratios. The model needed at least 20% of a 5-minute tracing to be AF to pass the

threshold of AF detection, and whether the AF was in the beginning, end, or middle of the tracing had no impact (Figure 3A). When we replaced the AF portion of the tracing with a more challenging AF rhythm (ie, AF with noise artifact), the DNN required longer stretches of AF (~40%) to be convinced (Figure 3B).

To investigate the effect of IBI distribution on the DNN model performance, we performed inference on synthetic tracings created from randomly ordered IBIs (simulating AF) of varying mean and standard deviation in a normal distribution (Figure 4A). This revealed that the model requires an IBI SD of ~8 to 20 beats/min to detect AF most accurately. Below this range generates non-AF prediction, and signals above 15 to 20 beats/min have a high rate of "uncertain" decisions. It is clear that when plotting predictions of DNN-raw on the internal validation set, the majority of false negatives arose from heart rates outside of 45 to 120 beats/min (blue dots in Figure 4B); however, the false positives did not follow a clear heart rate pattern and seem to cluster in the 50- to 60-beats/min band (yellow dots in Figure 4C). Of note, the DNN model had very few training examples outside of 45 to 120 beats/min (Supplemental Figure 1). We also explored the impact of the degree of normality; we found no relationship between the normality of the distribution and probability of AF decisions (Supplemental Figure 2).

TABLE 3 Operating Characteristics of Models on External Validation

	% of Data Classified	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	ROC-AUC
Signal processing (S = 100)	88	91.2	98.0	92.3	97.8	0.989
Signal processing (S = 80)[a]	88	96.9	96.3	97.0	99.2	
DNN-standard[b]	97	90.1	99.7	98.6	97.9	0.994
DNN-lower uncertainty threshold[c]	86	94.0	99.9	99.2	99.3	
DNN-IBI	91	86.6	99.8	98.4	97.7	0.988

[a]S is the entropy cutpoint, and with lower entropy cutpoint there is increased sensitivity of AF detection and a better balance between sensitivity and specificity. [b]Model was evaluated using cutoffs >0.9 for AF, <0.1 for non-AF, and outputs 0.1 to 0.9 as uncertain. [c]Same as the DNN-standard model, except that outputs were adjusted to maximize classification accuracy in preference to % of data analyzed (>0.99 for AF and <0.01 for non-AF), with everything in between as uncertain.

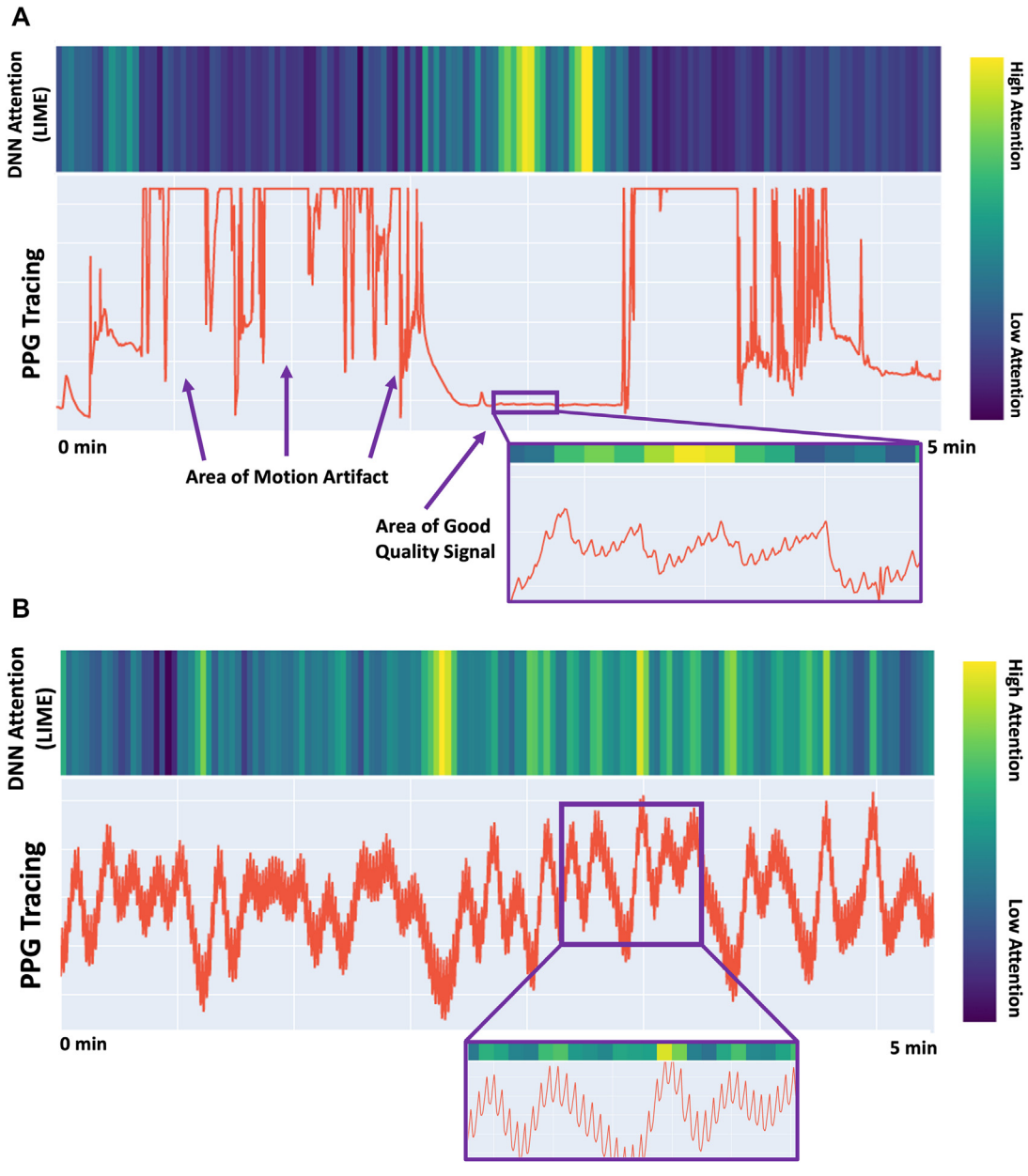
Abbreviations as in Table 2.

DISCUSSION

We have developed and validated 2 models for detecting AF using PPG: a SP model based on a proven high-performance AF detection heuristic, and a custom convolutional DNN trained on an extensive collection of raw PPG signals from individuals in their natural living environments. Overall, the SP model had similar performance with ROC-AUC of 0.972 (SP sensitivity 99.6%, specificity 94.4%) compared with the DNN, which demonstrated an ROC-AUC of 0.973 (DNN sensitivity 92.2%, specificity 95.5%); however, the DNN classified significantly more samples (95% vs 62%) while maintaining accuracy. With stricter uncertainty cutpoints, the DNN can perform even better (sensitivity 98.0%, specificity 99.9%) but at the cost of fewer tracings classified (79% classified). However, the DNN still classifies more tracings than SP. The robustness of our models was further confirmed by testing on an external dataset in a distinct patient group using a different PPG recording device (hospitalized patients by pulse oximeter vs free-living patients by smartwatch). The results demonstrated the DNN's exceptional performance, with an ROC-AUC of 0.994, compared with 0.989 for the SP model, underscoring its potential to generalize to various devices and populations, possibly with even higher efficacy. Moreover, the capacity of the DNN to analyze the most extensive duration of signal data in real-world conditions stands as a notable strength over pre-existing approaches to PPG analysis for rhythm detection, enabling continuous free-living monitoring.

Our SP model is a PPG implementation of one of the most high performing ECG-based AF detectors, originally developed by Zhou et al.[24] Like most AF detection heuristics, this algorithm was originally developed for ECGs and adapted for PPG input.[26] It was previously well validated on ECG datasets, and found to be highly tolerant of noise, making it ideal for PPG use. The PPG signal differs in that it does not contain P waves, is gathered by smartwatches in free-living ambulating patients, and is more prone to amplitude changes, noise, and artifact. We previously reported a robust SP detector that utilized a multi-sensor noise detection algorithm (PPG signal + accelerometer) to automatically reject an entire 5-minute segment and a hierarchical decision model based on IBI variability and entropy.[6] We were able to demonstrate a sensitivity of 89.7% and specificity of 97.0% with an AUC of 0.933 while classifying 32% of PPGs as indeterminate. Our SP model demonstrated significantly improved performance on the same dataset (AUC: 0.972) without the need for additional sensor input (accelerometer). This is the only other study that validated a SP algorithm on continuous signal from free-living patients. In comparison, other models include McManus et al,[22] who combined 3 independently validated techniques that included root mean square of successive RR differences, Shannon entropy and Poincaré plots, and was able to achieve a sensitivity of 97.0% and a specificity of 93.5% but discarded an unreported amount of poor-

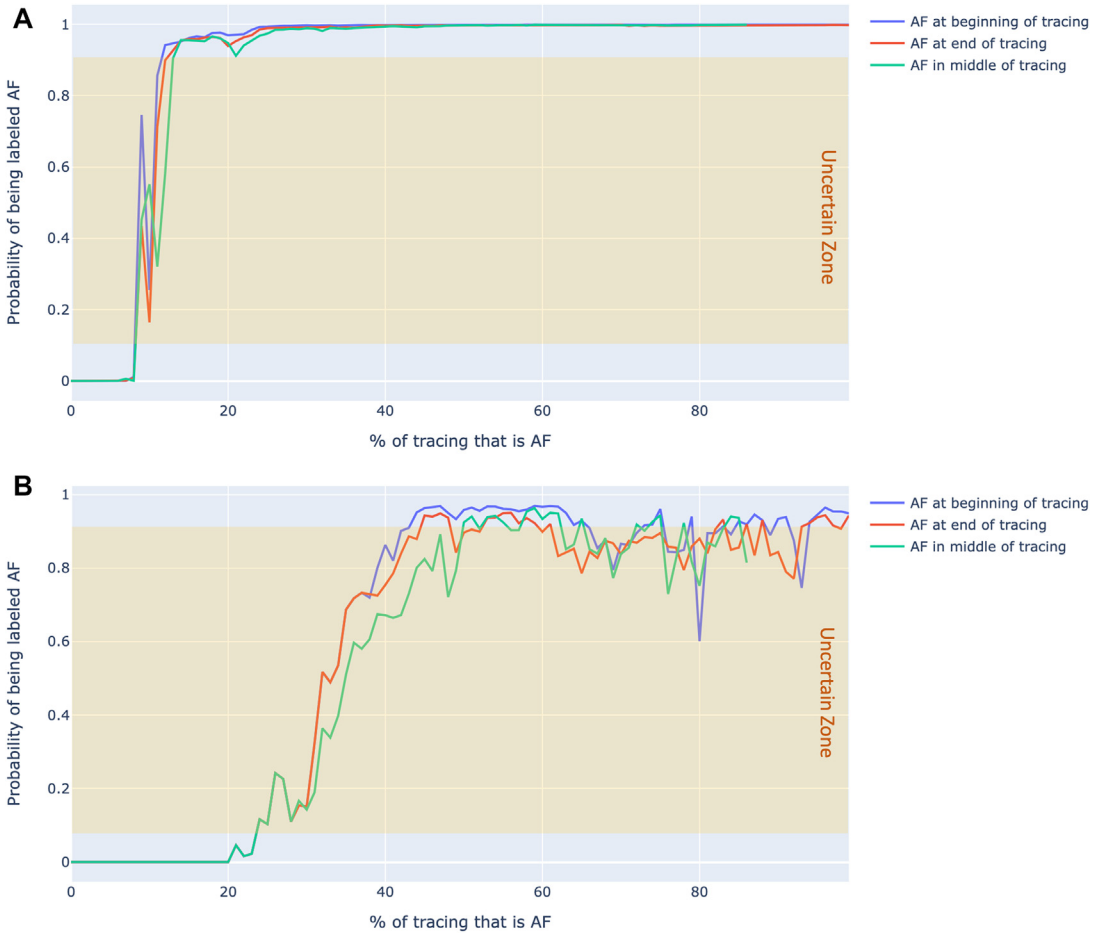
FIGURE 2 DNN Interpretation Using LIME



Area at the top demonstrates deep neural network (DNN)-determined relative importance of the respective part of the tracing to model prediction (yellow = high importance, blue = low importance). (A) Tracing with substantial motion artifact showing the DNN suppressed areas of high-frequency noise, and focused on areas of clean biological signal. (B) A tracing with minimal motion artifact—the DNN weighed the entire signal relatively equally. PPG = photoplethysmography. LIME = Local Interpretable Model-Agnostic Explanations.

quality recordings and asked the user to reacquire until a high-quality recording was gathered. Conroy et al[21] also developed a model based on heart rate variability measures, resulting in a sensitivity and specificity of 90.9%, but did not mention noise rejection rate. Zaman et al[20] also reported one of the most accurate AF detection heuristics to date (sensitivity 97%, specificity 98%), but ~59% of data were discarded as corrupted. Most of these previously reported SP models (except Avram et al)[6] are limited by

FIGURE 3 **Probability of AF Class Plotted Against Percent of AF on a 5-Minute Segment**

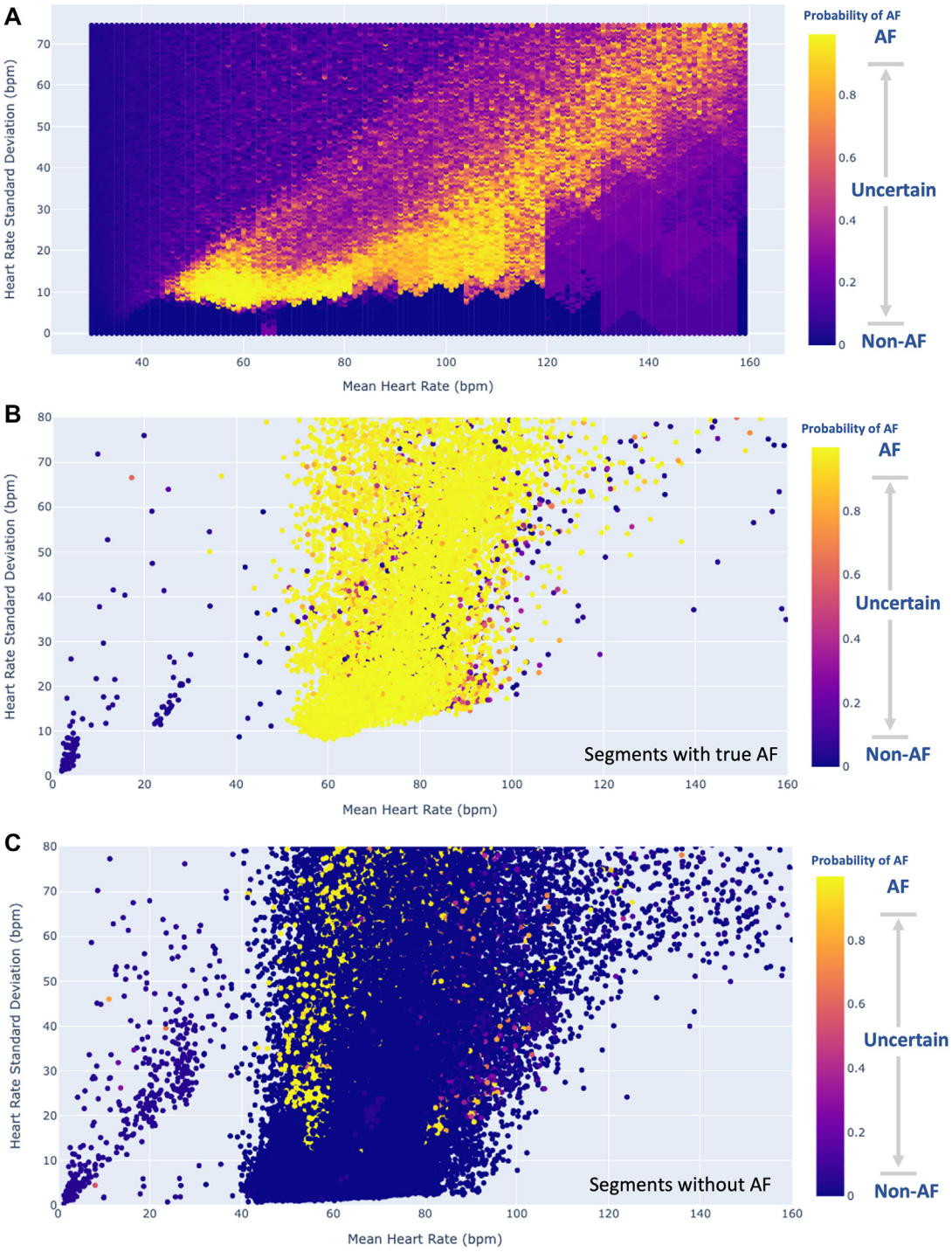


The model was evaluated by performing inference on spliced atrial fibrillation (AF)/sinus photoplethysmography tracings at various ratios. (A) Probability of AF label plotted against the % of tracing that is AF. (B) Same as panel A, except a more challenging AF tracing was used (ie, AF with noise).

selection bias: they were validated in highly selected patients—those undergoing a cardioversion, during which the patient is still and the device has optimal skin contact to minimize noise. These are not typical PPG tracings found on consumer wearable devices (**Figure 2A**) in a free-living environment, and as such, these models cannot be used for continuous outpatient monitoring of AF without more robust validation.

The commercial models, developed by Fitbit/Google and Apple, used data from free-living patients but only interpreted tracings at certain times of day and when the accelerometer revealed no motion.[6-8] For example, the Apple Heart Study evaluated the Apple irregular rhythm notification (IRN) algorithm by screening 419,297 participants and enrolling 450 (0.5%) with a positive IRN. Subsequent IRNs were compared against a gold standard of ECG patch recordings, resulting in a PPV of 71% for an individual tachogram, compared with 76% to 99% in our study. The IRN algorithm sampled 1-minute tachograms every 2 hours, equaling at most 12 minutes of AF monitoring per day, of which an unknown proportion is further discarded if the accelerometer detects motion. The study was also limited in assessing the true performance of the algorithm because the gold-standard assessment of AF using ECG monitoring was done only in a small convenience sample of those

FIGURE 4 **Probability of AF Class Plotted Against Mean Heart Rate and Heart Rate Standard Deviation ("Range")**



(A) Predictions of the DNN-IBI model on randomly ordered IBI sets of various means and standard deviations. (B) The heart rate distributions of the validation dataset that had true AF—blue dots on this graph indicate false negatives. (C) Heart rate distributions of the validation dataset that did not have AF—yellow dots represent false positives. The dot color represents model prediction (yellow = AF, blue = non-AF, shades of orange = uncertain). Abbreviations as in Figures 2 and 3.

with an IRN alert who chose to wear and return the patch ECG monitor, rather than in the entire population or a random sample; as a result, the false negative rate could not be assessed and the true positive rate was in a potentially biased sample. The newer algorithm from the Apple Watch, AF History, attempts to detect AF as frequently as every 15 minutes but still only analyzes when the user is completely at rest; no validation studies are published on this newer algorithm. The Fitbit study attempted to improve sampling time by analyzing every 5 minutes with 50% overlap between segments. However, they also only collected data when users were stationary for the entire collection period, which resulted in monitoring for an average 8 hours/day, 76% of which was during sleep. They screened 455,699 participants and only enrolled participants flagged positive by their algorithm (~1%) into simultaneous validation with ECG, reporting a PPV of 97.0%. However, much like the Apple Heart Study, they enrolled highly selected patients—those already flagged positive by their strict algorithm, which was 11 consecutive PPG segments positive for AF (30 minutes), rendering an assessment of false negatives impossible. Even in this selected sample, they report a sensitivity of 68.0%, compared to a sensitivities of 91.2% to 99.6% for the signal processing heuristic and 90.1% to 98.0% for the DNN in our study reported herein.[8] Of note, the noise rejection technique, the proportion rejected due to signal quality, or those with indeterminate classifications were not reported in either Apple or Fitbit studies.

Many other DNNs were developed for AF detection on PPG signal, but were all limited by selection bias because they were validated in an artificial environment on patients who were asleep or supine and still, minimizing or eliminating noisy signals. For example, Gotlibovych et al,[27] Kwon et al,[13] and Tison et al[14] reported DNNs with exceptional performance in identifying AF with an AUC of 0.999, an AUC of 0.996, and a C-statistic of 0.97, respectively. However, the validation and training sample includes patients around the time of sedation/cardioversion under supervision or healthy volunteers during sleep, limiting the generalizability to ambulatory patients.

Another finding of this study is the exceptional performance of DNNs when trained on raw PPG signal. Many previously published DNNs were paired with various noise-filtering algorithms,[17,19] or used IBIs as an input instead of raw signal.[6-8,14,18] In our study, we found that a DNN trained on raw signal outperformed the DNN trained on IBIs, allowing the DNN to develop its own noise rejection and interval/rhythm detection, which led to classification of more signals (95%) than our SP model (62%) and other published DNNs trained on IBIs. The LIME explainability analysis indicated automatic suppression of nonbiologic signal with impressive accuracy (**Figure 2A**). As a result, the DNN seems to use as much of the signal as it can extract to make its prediction (**Figure 2B**). In fact, user-level data indicated that only 1 patient of 29 who did not have AF received a false positive using our DNN classifier, compared with 16 patients with the SP model. The false positives in the SP model were mostly on segments containing significant noise, which highlights DNN's superior performance in noisy signal (**Table 4**).

TABLE 4 False Positive Results Based on User-Level Tracings

	SP False Positives	DNN False Positives
ECG/PPG tracing review	N users N total = 16 (segments)[a]	N users N total = 1 (segments)
Noise/artifact[b]	11 (63)	0
Frequent PACs/PVCs	4 (126)	0
Sinus arrhythmia	2 (184)	1 (25)

Values are n (%). [a]Numbers do not add to 16 because 1 user had PPG tracing with both poor quality and frequent PACs. [b]Based on visual assessment of PPG waveform.

DNN = deep neural network; ECG = electrocardiography; PAC = premature atrial contraction; PPG = photoplethysmography; PVC = premature ventricular contraction; SP = signal processing.

It is worth emphasizing that the DNN outlined in this study has a small memory footprint (~725 kb), and thus can be deployed on wearable devices to continuously monitor for AF. The thresholds of the DNN can also be adjusted to ensure higher certainty when screening patients without a history of AF to avoid false alarms, and modified to increase sensitivity of AF detection in patients with known AF to characterize the burden of AF (% of time spent in AF). This fine-tuning is more difficult with SP due to a separate noise-rejection algorithm.

STUDY LIMITATIONS. First of all, the study included generally younger patients (52.0% <65 years of age) with mostly Caucasian ethnicity, which may introduce some bias. Of note, these patients have self-identified AF, which represents a higher risk group.

Explainability analysis using synthetic data reveals that the DNN most accurately identifies AF between heart rates of 45 and 120 beats/min (**Figure 4A**). PPG tracings with heart rates outside of this range result in high false negative rates (**Figures 4B and 4C**), likely due to a paucity of training data with extreme heart rates (Supplemental Figure 2). This is similar to the operating range of current commercial AF detection algorithms.[10-12] Performance can be improved by enriching training with examples of AF at those heart rates; however, PPG-based AF detection at fast heart

rates is challenging due to smaller difference in IBI, and may not be necessary, as the smartwatch can recommend seeking medical attention at persistently extreme heart rates.

The model also requires at least 1 minute of AF for detection; however, the clinical significance of very short AF episodes is unknown especially if patients can be continuously monitored.[4] Understanding the inner workings of the DNN is also challenging. All explainability techniques have limitations as they can only evaluate some features that are potentially used by the model, and only provide limited insight into how the DNN functions. In comparison, the SP model was specifically engineered to use certain features on PPG based on known characteristics of AF, and is easier to understand/troubleshoot. DNNs are also prone to overfitting—to overcome this, we externally validated our DNN, which revealed an impressive ability to generalize to other patients and PPG recording devices (pulse oximeter vs smartwatch).

CONCLUSIONS

This report highlights the strength of DNNs in PPG-based AF detection by demonstrating that it is able to accurately classify almost all available PPG signal in free-living patients, which is significantly more than the best available SP heuristic. Automatic feature detection of DNNs can be maximized by training on raw signal, as performance can degrade if trained on preprocessed signal (IBIs or filtering). DNNs are able to suppress motion artifact/noise, infer IBIs, evaluate their stochastic nature, and incorporate information about the distribution of IBIs to identify AF. Future DNN-based models that are trained on raw signal can provide minimally invasive continuous AF monitoring with accuracy that approaches that of ambulatory ECG monitoring.

FUNDING SUPPORT AND AUTHOR DISCLOSURES

Dr Olgin is supported by a grant from the Mark Cuban Foundation. Dr Tison is supported by National Institutes of Health grant K23HL135274. Dr Avram is supported by the Fonds de la recherche en santé du Québec (grant no. 312758-Montreal Heart Institute Research Centre, the Montreal Heart Institute Foundation). All other authors have reported that they have no relationships relevant to the contents of this paper to disclose.

ADDRESS FOR CORRESPONDENCE: Dr Jeffrey Olgin, Division of Cardiology, Department of Medicine and Cardiovascular Research Institute, University of California-San Francisco, 505 Parnassus Avenue, San Francisco, California 94117, USA. E-mail: jeffrey.olgin@ucsf.edu.

PERSPECTIVES

COMPETENCY IN SYSTEMS-BASED PRACTICE: The authors outline technology for remote minimally invasive monitoring of AF in free-living outpatients using wearable PPG sensors (ie, smartwatches). This is a new paradigm for outpatient monitoring of heart rhythm abnormalities.

TRANSLATIONAL OUTLOOK: The next steps involve translational research to deploy the AF detection model to wearable PPG-capable devices and perform prospective validation.

REFERENCES

1. Kirchhof P, Camm AJ, Goette A, et al. Early rhythm-control therapy in patients with atrial fibrillation. *N Engl J Med*. 2020;383:1305-1316.

2. Andrade JG, Wells GA, Deyell MW, et al. Cryoablation or drug therapy for initial treatment of atrial fibrillation. *N Engl J Med*. 2021;384:305-315.

3. Rienstra M, Lubitz SA, Mahida S, et al. Symptoms and functional status of patients with atrial fibrillation. *Circulation*. 2012;125:2933-2943.

4. Andrade JG, Aguilar M, Atzema C, et al. The 2020 Canadian Cardiovascular Society/Canadian Heart Rhythm Society comprehensive guidelines for the management of atrial fibrillation. *Can J Cardiol*. 2020;36:1847-1948.

5. Castaneda D, Esparza A, Ghamari M, Soltanpur C, Nazeran H. A review on wearable photoplethysmography sensors and their potential future applications in health care. *Int J Biosens Bioelectron*. 2018;4:195-202.

6. Avram R, Ramsis M, Cristal AD, et al. Validation of an algorithm for continuous monitoring of atrial fibrillation using a consumer smartwatch. *Heart Rhythm*. 2021;18:1482-1490.

7. Perez MV, Mahaffey KW, Hedlin H, et al. Large-scale assessment of a smartwatch to identify atrial fibrillation. *N Engl J Med*. 2019;381:1909-1917.

8. Lubitz SA, Faranesh AZ, Selvaggi C, et al. Detection of atrial fibrillation in a large population using wearable devices: the Fitbit Heart Study. *Circulation*. 2022;146:1415-1424.

9. Zhu L, Nathan V, Kuang J, et al. Atrial fibrillation detection and atrial fibrillation burden estimation via wearables. *IEEE J Biomed Health Inform*. 2022;26:2063-2074.

10. Olson L. FDA Filing: Photoplethysmograph analysis software for over-the-counter use. Food and Drug Administration. 2018. Accessed April 13, 2023. https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180042.pdf

11. Olson L. AF history FDA. 510(k) premarket notification. 2022. Accessed April 11, 2023. https://www.accessdata.fda.gov/cdrh_docs/pdf21/K213971.pdf

12. Parry R. Fitbit irregular rhythm notifications. Food and Drug Administration. 2022. Accessed April 11, 2023. https://www.accessdata.fda.gov/cdrh_docs/pdf21/K212372.pdf

13. Kwon S, Hong J, Choi E-K, et al. Deep learning approaches to detect atrial fibrillation using photoplethysmographic signals: algorithms development study. *JMIR Mhealth Uhealth*. 2019;7:e12770.

14. Tison GH, Sanchez JM, Ballinger B, et al. Passive detection of atrial fibrillation using a commercially available smartwatch. *JAMA Cardiol*. 2018;3:409–416.

15. Lee J, Nam Y, McManus DD, Chon KH. Time-varying coherence function for atrial fibrillation detection. *IEEE Trans Biomed Eng*. 2013;60:2783–2793.

16. Nemati S, Ghassemi MM, Ambai V, et al. Monitoring and detecting atrial fibrillation using wearable technology. *Annu Int Conf IEEE Eng Med Biol Soc*. 2016;2016:3394–3397.

17. Aliamiri A, Shen Y. Deep learning based atrial fibrillation detection using wearable photoplethysmography sensor. *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE; 2018:442–445.

18. Poh M-Z, Poh YC, Chan P-H, et al. Diagnostic assessment of a deep learning system for detecting atrial fibrillation in pulse waveforms. *Heart*. 2018;104:1921–1928.

19. Liang Y, Chen Z, Ward R, Elgendi M. Hypertension assessment via ECG and PPG signals: an evaluation using MIMIC database. *Diagnostics*. 2018;8:65.

20. Zaman R, Chong JW, Cho CH, Esa N, McManus DD, Chon KH. Motion and noise artifact-resilient atrial fibrillation detection using a smartphone. *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE; 2016:366–369.

21. Conroy T, Guzman JH, Hall B, Tsouri G, Couderc J-P. Detection of atrial fibrillation using an earlobe photoplethysmographic sensor. *Physiol Meas*. 2017;38:1906–1918.

22. McManus DD, Chong JW, Soni A, et al. PULSE-SMART: pulse-based arrhythmia discrimination using a novel smartphone application: automated arrhythmia discrimination using a smartphone. *J Cardiovasc Electrophysiol*. 2016;27:51–57.

23. Peyser ND, Marcus GM, Beatty AL, Olgin JE, Pletcher MJ. Digital platforms for clinical trials: the Eureka experience. *Contemp Clin Trials*. 2022;115:106710.

24. Zhou X, Ding H, Ung B, Pickwell-MacPherson E, Zhang Y. Automatic online detection of atrial fibrillation based on symbolic dynamics and Shannon entropy. *Biomed Eng Online*. 2014;13:18.

25. Ribeiro MT, Singh S, Guestrin C. Model-agnostic interpretability of machine learning. Preprint. Posted online June 16, 2016. *arXiv*. 2016:1606.05386. https://doi.org/10.48550/arXiv.1606.05386

26. Pereira T, Tran N, Gadhoumi K, et al. Photoplethysmography based atrial fibrillation detection: a review. *NPJ Digit Med*. 2020;3:3.

27. Gotlibovych I, Crawford S, Goyal D, et al. End-to-end deep learning from raw sensor data: atrial fibrillation detection using wearables. Preprint. Posted online July 27, 2018. *arXiv*. 2018;1807.10707. https://doi.org/10.48550/arXiv.1807.10707

KEY WORDS **atrial fibrillation, deep learning, rhythm monitoring, signal processing, smartwatch**

APPENDIX **For an expanded supplemental Methods section as well as a table and figures, please see the online version of this paper.**