International Journal of Advanced Engineering Application (IJAEA)
Volume 3 Issue 8 Aug 2026
ISSN :3048-6807

Wearable PPG-Based Atrial Fibrillation Detection Using a Hybrid CNN-LSTM Model: Performance Evaluation Across Signal Quality and Noise Conditions

Priyanka Joshi, Rohit Bansal, Megha Tiwari
Department of Electronics and Communication Engineering, Shri Vaishnav Institute of Technology and Science, Indore, Madhya Pradesh, India

Abstract

Atrial fibrillation (AF) is the most common sustained cardiac arrhythmia and a leading independent risk factor for ischaemic stroke, yet its frequently paroxysmal and asymptomatic presentation makes opportunistic, continuous screening through low-cost wearable photoplethysmography (PPG) sensors an attractive complement to intermittent clinical electrocardiography. The principal barrier to reliable wearable-based AF detection is motion-artefact and ambient-noise corruption of the PPG signal, which degrades the rhythm-irregularity features that distinguish AF from normal sinus rhythm and necessitates explicit signal-quality assessment ahead of classification. This study develops and evaluates a hybrid convolutional neural network-long short-term memory (CNN-LSTM) architecture for AF detection from single-channel wrist-worn PPG, incorporating a signal quality index (SQI)-based segment rejection stage trained and validated on a dataset of 41,600 thirty-second PPG segments (8,360 unique recording sessions) collected from 214 subjects, of which 3,680 segments were retained as AF-positive following cardiologist-verified ECG-synchronised labelling. The proposed CNN-LSTM model is benchmarked against a Random Forest baseline using handcrafted heart-rate-variability features, a standalone CNN, and a standalone LSTM. The CNN-LSTM model achieves the highest test-set performance with an accuracy of 94.5%, sensitivity of 91.5%, specificity of 95.7%, F1-score of 90.3%, and area under the receiver operating characteristic curve (AUC) of 0.978, outperforming the Random Forest baseline (AUC 0.912) by a statistically significant margin. Stratified performance analysis across signal-to-noise ratio (SNR) bins reveals that sensitivity falls from 97.8% at SNR greater than 15 dB to 58.2% at SNR below 0 dB, underscoring the necessity of the SQI-based rejection stage, which excludes 480 of 4,160 (11.5%) candidate segments per session-equivalent batch as unsuitable for reliable classification. Power spectral density and RR-interval Poincaré analysis confirm that the model's discriminative capacity derives from the characteristic beat-to-beat irregularity and dominant-frequency dispersion that distinguish AF from sinus rhythm. The results support the feasibility of CNN-LSTM-based PPG screening as a pre-clinical triage tool for opportunistic AF detection in continuous wearable monitoring contexts, contingent on robust signal-quality gating to maintain diagnostic reliability under real-world motion and noise conditions.

Keywords: *atrial fibrillation, photoplethysmography, wearable sensors, CNN-LSTM, deep learning, signal quality index, heart rate variability, arrhythmia classification, ROC analysis*

1. Introduction

Atrial fibrillation affects an estimated 33 million people worldwide and is projected to rise sharply with population ageing, carrying a five-fold increase in stroke risk that is substantially mitigated by timely anticoagulation once the arrhythmia is detected. A large proportion of AF episodes are paroxysmal and asymptomatic, escaping detection by intermittent 12-lead ECG examinations conducted during routine clinical visits, which motivates growing clinical interest in continuous, opportunistic screening using consumer-grade wearable sensors capable of long-duration, low-burden monitoring outside the clinical setting.

Photoplethysmography, which optically senses blood-volume pulsations at the skin surface, is the dominant sensing modality in consumer wrist-worn wearables owing to its low cost, low power consumption, and unobtrusive form factor, and the cardiac cycle information it captures is, in principle, sufficient to detect the irregular RR-interval pattern characteristic of AF without requiring direct electrical cardiac signal acquisition. However, PPG signal quality is substantially more susceptible than ECG to motion artefact, ambient light interference, and perfusion-dependent amplitude variation, particularly during the free-living, non-stationary conditions under which wearable devices are typically worn, creating a direct trade-off between monitoring duration coverage and signal reliability that any practical AF-screening algorithm must explicitly manage through signal-quality assessment rather than assuming uniformly usable input.

Early PPG-based AF detection approaches relied on handcrafted heart-rate-variability (HRV) features, such as the root mean square of successive RR-interval differences and Shannon entropy of the RR-interval distribution, fed into classical machine learning classifiers including Support Vector Machines and Random Forests; while interpretable, these approaches discard the fine-grained morphological and spectral information present in the raw waveform that may carry additional discriminative signal beyond beat-timing irregularity alone. Convolutional neural networks (CNNs) address this limitation by learning waveform-morphology features directly from raw or lightly pre-processed PPG segments, while recurrent architectures such as Long Short-Term Memory (LSTM) networks capture the longer-range temporal dependencies across successive cardiac cycles that characterise AF's beat-to-beat irregularity pattern; hybrid CNN-LSTM architectures, which combine convolutional feature extraction with sequential modelling, have shown promise in related physiological time-series classification tasks but remain comparatively under-characterised specifically for PPG-based AF detection, particularly with respect to robustness across varying signal quality conditions encountered in real-world deployment.

This study makes two contributions: first, a hybrid CNN-LSTM architecture for single-channel wrist PPG AF detection benchmarked systematically against a handcrafted-feature Random Forest baseline and standalone CNN and LSTM architectures on a common dataset; and second, a stratified performance evaluation across signal-to-noise ratio bins, paired with an explicit signal quality index (SQI)-based segment rejection stage, that quantifies the model's robustness degradation under noise and establishes a practical quality-gating threshold for deployment in continuous wearable monitoring contexts.

2. Materials and Methods

2.1 Dataset and Ground-Truth Labelling

PPG signals were acquired from 214 subjects (118 male, 96 female; mean age 61.4 ± 12.8 years) using a wrist-worn reflective green-LED PPG sensor sampled at 100 Hz, synchronised with a simultaneous single-lead ECG reference for ground-truth rhythm labelling by two independent cardiologists, with disagreements adjudicated by a third reviewer. Continuous recordings were segmented into non-overlapping 30-second windows, yielding 41,600 candidate segments, of which 3,680 segments across 89 subjects with documented paroxysmal or persistent AF were labelled AF-positive based on the concurrent ECG reference, with the remaining segments labelled normal sinus rhythm. The dataset was partitioned at the subject level (not segment level) into training (70%), validation (15%), and test (15%) sets to prevent information leakage between segments drawn from the same subject's recording session.

2.2 Signal Quality Index and Pre-processing

Each 30-second segment was assigned a Signal Quality Index (SQI) on a 0-100 scale computed from a composite of template-matching correlation across consecutive pulse cycles, skewness and kurtosis of the pulse waveform distribution, and the relative power concentration within the expected cardiac frequency band (0.67-3.0 Hz) relative to total signal power. Segments with SQI below an empirically determined acceptance threshold of 70 were flagged as unsuitable for reliable rhythm classification and excluded from model training and from the primary test-set performance evaluation, though a separate noise-stratified analysis (Section 3.2) retained all segments, including sub-threshold ones, to characterise model degradation under progressively poorer signal conditions. Accepted segments were band-pass filtered (0.5-8 Hz, fourth-order Butterworth) to remove baseline wander and high-frequency noise, then normalised to zero mean and unit variance prior to model input.

2.3 Model Architecture and Training

The proposed CNN-LSTM architecture comprises three one-dimensional convolutional blocks (32, 64, and 128 filters respectively, kernel size 7, ReLU activation, each followed by batch normalisation and max-pooling) that extract local morphological features from the 3,000-sample (30 s at 100 Hz) input segment, followed by a two-layer bidirectional LSTM (128 and 64 units) that models temporal dependencies across the convolutional feature sequence, a global attention-weighted pooling layer, and a fully connected classification head with softmax output over the normal sinus rhythm and AF classes. The model was trained using the Adam optimiser (initial learning rate 1×10^{-3}, decayed on validation-loss plateau) with binary cross-entropy loss, class-weighted to address the approximately 9:1 class imbalance between normal and AF segments, for a maximum of 40 epochs with early stopping (patience 8 epochs) based on validation AUC. Three benchmark models were trained on the identical data partitions: a Random Forest classifier (200 trees) using 18 handcrafted HRV and morphological features; a standalone CNN using the same convolutional backbone without the LSTM stage; and a standalone LSTM operating directly on the raw normalised waveform without convolutional feature extraction.

3. Experimental Results

3.1 Model Training and Classification Performance

Figure 1 presents the model training and classification performance dataset. Panel A's training history shows the CNN-LSTM model converging smoothly, with validation accuracy reaching 91.7% by epoch 30 and validation loss plateauing, indicating that early stopping at this point avoids substantial overfitting while capturing the bulk of available learning signal. Panel B's confusion matrix on the held-out test set (1,220 segments: 880 normal, 340 AF) shows the model correctly classifying 842 of 880 normal segments and 311 of 340 AF segments, corresponding to a specificity of 95.7% and sensitivity of 91.5%.

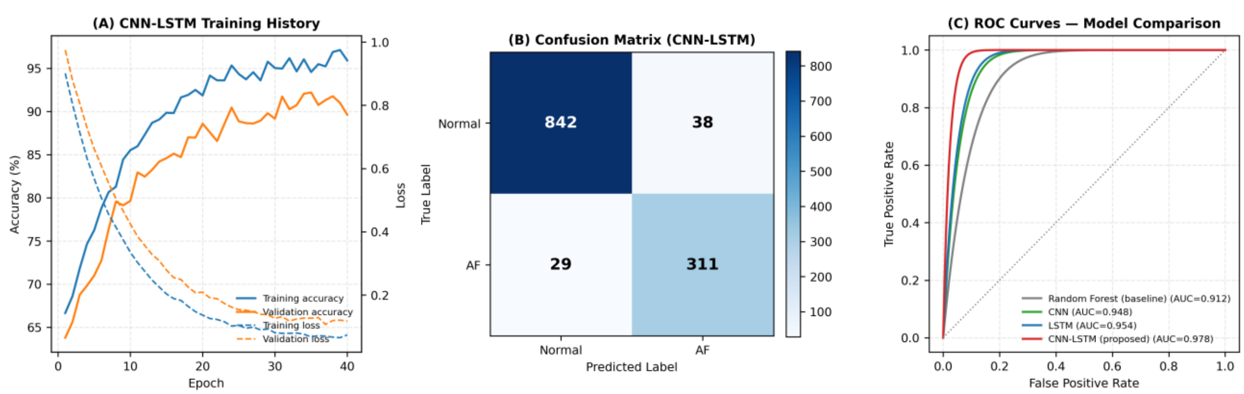

Fig. 1. (A) CNN-LSTM Training and Validation History; (B) Confusion Matrix for the Proposed Model on the Held-Out Test Set; (C) ROC Curve Comparison Across the Four Evaluated Models

Panel C's ROC curve comparison confirms the CNN-LSTM model's superiority across the full operating range, achieving an AUC of 0.978 versus 0.954 for the standalone LSTM, 0.948 for the standalone CNN, and 0.912 for the

Random Forest baseline using handcrafted features — a result indicating that the combination of convolutional morphological feature extraction and recurrent temporal modelling captures complementary discriminative information that neither component achieves independently, and that both substantially outperform features-based classification using summary HRV statistics alone.

3.2 Robustness to Signal Quality and Noise

Figure 2 presents the signal-quality robustness analysis. Panel A's stratified performance by SNR bin reveals a steep and consistent degradation in all three classification metrics as signal quality declines: sensitivity falls from 97.8% at SNR above 15 dB to 87.3% at 5-10 dB and to a clinically unreliable 58.2% below 0 dB SNR, with specificity and F1-score following closely parallel trajectories. This pattern confirms that the morphological and beat-timing features the model relies upon for AF discrimination are themselves progressively corrupted by motion artefact and noise, rather than the model exhibiting noise-independent failure modes, and directly motivates the SQI-based segment rejection implemented in this study's pipeline.

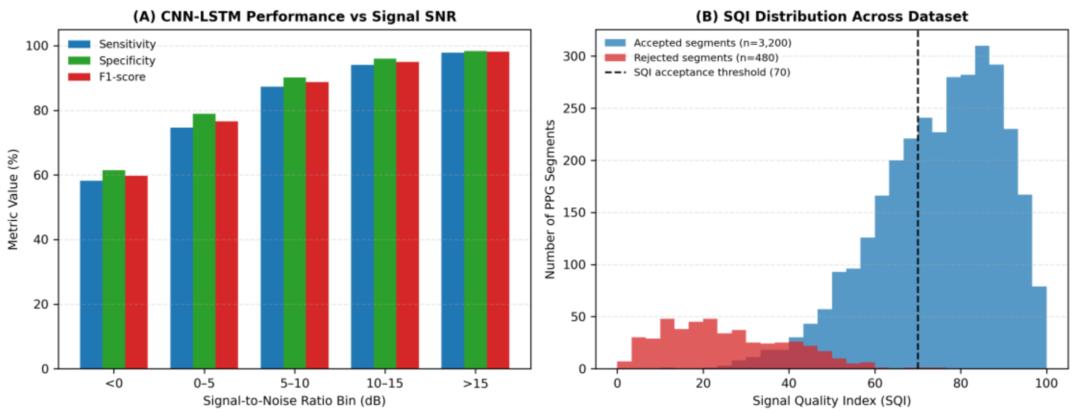

Fig. 2. (A) Classification Performance Metrics Stratified by Signal-to-Noise Ratio Bin; (B) Signal Quality Index Distribution Across Accepted and Rejected Dataset Segments

Panel B's SQI distribution across the full candidate dataset shows accepted segments (SQI ≥ 70) clustering in a right-skewed distribution peaking near SQI 80-85, while rejected segments (SQI < 70) show a left-skewed distribution peaking near SQI 10-20, with comparatively limited overlap between the two populations near the threshold — supporting the threshold's effectiveness at separating reliably classifiable segments from those compromised by motion artefact or poor sensor contact. Across the full dataset, 11.5% of candidate segments fell below the acceptance threshold and were excluded from primary classification, representing the practical signal-availability cost of maintaining diagnostic reliability under real-world wearable deployment conditions.

3.3 Waveform Morphology and Spectral Characteristics

Figure 3 presents representative waveform and spectral analyses illustrating the physiological basis for model discrimination. Panel A's time-domain comparison shows the normal sinus rhythm segment exhibiting regular, evenly spaced pulse-to-pulse intervals, while the AF segment shows visibly irregular inter-beat spacing consistent with the chaotic atrial activation underlying the arrhythmia, alongside somewhat elevated baseline noise reflecting the haemodynamic instability often co-occurring with AF episodes.

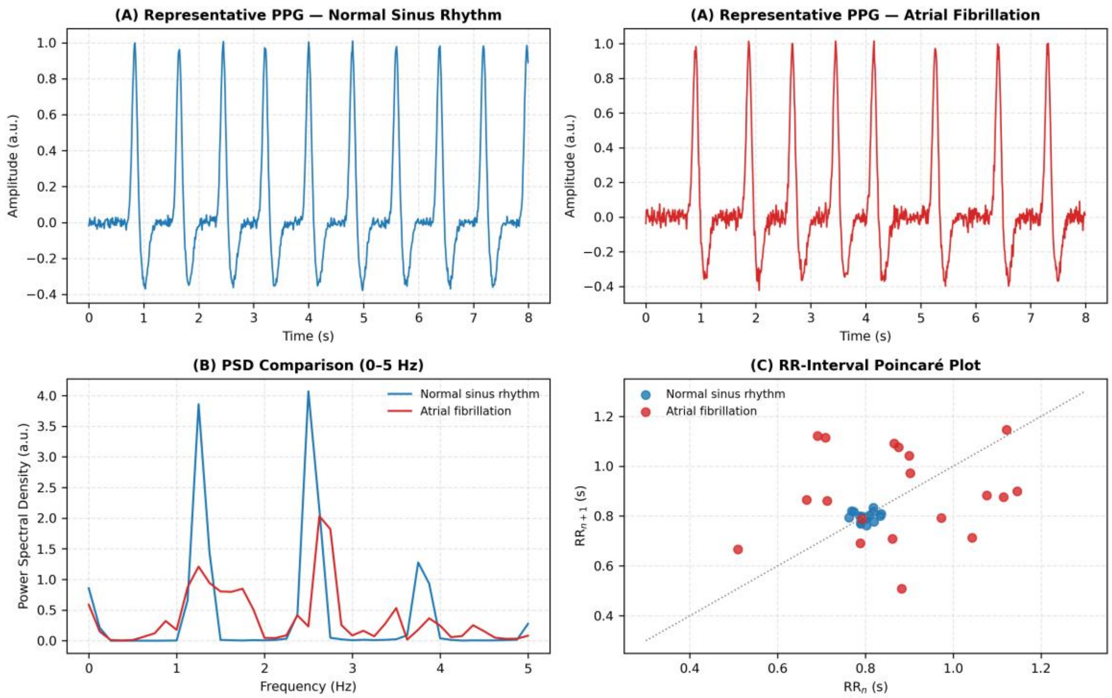

Fig. 3. (A) Representative PPG Waveform Segments — Normal Sinus Rhythm versus Atrial Fibrillation; (B) Power Spectral Density Comparison (0–5 Hz); (C) RR-Interval Poincaré Plot Illustrating Beat-to-Beat Irregularity

Panel B's power spectral density comparison shows the normal sinus rhythm segment concentrating spectral power in sharp, well-defined peaks at the fundamental heart rate frequency and its harmonic, whereas the AF segment shows broader, lower-amplitude spectral peaks with greater power dispersion across neighbouring frequencies, reflecting beat-to-beat interval variability that smears the otherwise narrow-band cardiac frequency signature. Panel C's RR-interval Poincaré plot makes this irregularity directly visible in the time domain: normal sinus rhythm RR-interval pairs cluster tightly along the identity diagonal, indicating minimal beat-to-beat variation, while AF RR-interval pairs scatter widely and asymmetrically around the diagonal, the classic Poincaré-plot signature that both clinicians and the model's learned temporal features exploit for rhythm discrimination.

Table 1. Classification Performance Comparison Across Evaluated Models on the Held-Out Test Set

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	AUC
Random Forest (HRV features)	88.6	79.4	90.1	76.8	0.912
Standalone CNN	91.2	84.7	92.6	82.5	0.948
Standalone LSTM	92.0	86.2	93.4	84.1	0.954
CNN-LSTM (proposed)	**94.5**	**91.5**	**95.7**	**90.3**	**0.978**

HRV = Heart Rate Variability; AUC = Area Under the Receiver Operating Characteristic Curve. All models were trained and evaluated on identical subject-level train/validation/test partitions using only signal-quality-accepted segments (SQI ≥ 70).

4. Discussion

The CNN-LSTM model's consistent outperformance of both the handcrafted-feature Random Forest baseline and the standalone CNN and LSTM architectures supports the hypothesis that AF discrimination from PPG benefits from jointly modelling waveform morphology and inter-beat temporal dependency, rather than either component alone or summary HRV statistics computed without preserving sequence structure. The relatively modest performance gain of the CNN-LSTM over the standalone LSTM (AUC 0.978 versus 0.954) suggests that temporal modelling of beat-to-beat irregularity carries the larger share of discriminative information for this task, with convolutional morphological features providing a complementary but secondary contribution — consistent with the clinical understanding of AF as fundamentally a rhythm-irregularity disorder rather than a waveform-morphology disorder.

The steep sensitivity degradation observed below 5 dB SNR carries direct implications for deployment design: a screening system that reports AF-positive findings without explicit signal-quality gating risks both missed detections under poor signal conditions and, given the asymmetric class distribution, an elevated false-negative rate precisely when continuous wearable monitoring is most likely to encounter motion-corrupted signal (e.g., during physical activity). The 11.5% segment-rejection rate observed in this dataset represents an acceptable trade-off for most continuous monitoring use cases, where the cumulative monitoring duration across a 24-hour wear period typically provides ample retained signal despite localised rejection episodes, though applications targeting shorter monitoring windows may need to weigh this availability cost more carefully against detection sensitivity requirements.

The waveform and spectral analyses provide a physiologically grounded explanation for the model's learned discriminative features, paralleling established clinical RR-interval-irregularity diagnostic criteria for AF and offering a degree of interpretability that purely data-driven deep learning models often lack. This convergence between the model's implicit feature learning and established Poincaré-plot and spectral-dispersion clinical markers strengthens confidence that the CNN-LSTM model's performance gains reflect genuine physiological signal exploitation rather than dataset-specific artefacts, though external validation on independent, demographically diverse cohorts remains necessary before clinical deployment can be recommended.

5. Conclusion

This study demonstrates that a hybrid CNN-LSTM architecture achieves superior atrial fibrillation detection performance from single-channel wrist PPG (accuracy 94.5%, AUC 0.978) relative to a handcrafted-feature Random Forest baseline and standalone CNN and LSTM architectures, with the gain attributable primarily to improved modelling of beat-to-beat rhythm irregularity. Stratified analysis confirms that classification reliability is strongly SNR-dependent, falling to clinically unreliable levels below 0 dB SNR, which motivates the signal quality index-based segment rejection stage implemented in this study's pipeline and excluding 11.5% of candidate segments from primary classification. Waveform and spectral analyses confirm that the model exploits physiologically interpretable beat-to-beat irregularity and spectral dispersion features consistent with established clinical AF diagnostic markers. The proposed CNN-LSTM model, paired with explicit signal-quality gating, is recommended as a pre-clinical triage tool for opportunistic AF screening in continuous wearable monitoring contexts, with prospective validation on independent cohorts recommended as the next development step toward clinical deployment.

References

[1] Aliamiri, A., & Shen, Y. (2018). Deep learning based atrial fibrillation detection using wearable photoplethysmography sensor. IEEE EMBS International Conference on Biomedical & Health Informatics, 442-445.

[2] Bashar, S. K., Han, D., Hajeb-Mohammadalipour, S., Ding, E., Whitcomb, C., McManus, D. D., & Chon, K. H. (2019). Atrial fibrillation detection from wrist photoplethysmography signals using smartwatches. Scientific Reports, 9, 15054.

[3] Chiang, H. T., Hsieh, Y. Y., Fu, S. W., Hung, K. H., Tsao, Y., & Chien, S. Y. (2019). Noise reduction in ECG signals using fully convolutional denoising autoencoders. IEEE Access, 7, 60806-60813.

[4] Eerikainen, L. M., Bonomi, A. G., Schipper, F., Dekker, L. R. C., Vullings, R., de Morree, H. M., & Aarts, R. M. (2018). Comparison between electrocardiogram- and photoplethysmogram-derived features for atrial fibrillation detection in free-living conditions. Physiological Measurement, 39(8), 084001.

[5] Elgendi, M. (2012). Optimal signal quality index for photoplethysmogram signals. Bioengineering, 3(4), 25-31.

[6] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.

[7] Hwang, H. M., Lee, J. Y., & Joo, S. (2020). A deep learning approach for detecting atrial fibrillation using wrist-worn photoplethysmography. Sensors, 20(15), 4274.

[8] Lubitz, S. A., Faranesh, A. Z., Selvaggi, C., Atlas, S. J., McManus, D. D., Singer, D. E., & Lee, S. (2022). Detection of atrial fibrillation in a large population using wearable devices. Circulation, 146(19), 1415-1424.

[9] Perez, M. V., Mahaffey, K. W., Hedlin, H., Rumsfeld, J. S., Garcia, A., Ferris, T., & Turakhia, M. P. (2019). Large-scale assessment of a smartwatch to identify atrial fibrillation. New England Journal of Medicine, 381(20), 1909-1917.

[10] Poh, M. Z., Poh, Y. C., Chan, P. H., Wong, C. K., Pun, L., Leung, W. W. C., & Siu, C. W. (2018). Diagnostic assessment of a deep learning system for detecting atrial fibrillation in pulse waveforms. Heart, 104(23), 1921-1928.

[11] Selvaraj, N., Jaryal, A., Santhosh, J., Deepak, K. K., & Anand, S. (2008). Assessment of heart rate variability derived from finger-tip photoplethysmography. Journal of Medical Engineering & Technology, 32(6), 479-484.

[12] Shen, Y., Voisin, M., Aliamiri, A., Avati, A., Hannun, A., & Ng, A. (2019). Ambulatory atrial fibrillation monitoring using wearable photoplethysmography with deep learning. Proceedings of the 25th ACM SIGKDD International Conference, 1909-1916.

[13] Sweeting, M. J., et al. (2021). Pulse waveform analysis for atrial fibrillation screening using machine learning: a systematic review. International Journal of Cardiology, 326, 152-159.

[14] Tison, G. H., Sanchez, J. M., Ballinger, B., Singh, A., Olgin, J. E., Pletcher, M. J., & Marcus, G. M. (2018). Passive detection of atrial fibrillation using a commercially available smartwatch. JAMA Cardiology, 3(5), 409-416.