Real-time Monitoring and Early Warning of Atrial Fibrillation

Marino E. Gavidia[1†], Hongling Zhu[2†], Arthur N. Montanari[1], Jesús Fuentes[1], Cheng Cheng[3], Sergio Dubner[4], Martin Chames[5], Yinuo Jiang[3], Shengjun Zhang[3], Hai-Tao Zhang[3], Xin He[6], Basi Teng[7], Guohua Wan[8], Ye Yuan[3*], Xiaoyun Yang[2*], and Jorge Goncalves[1,7*]

[1]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Belvaux L-4367, Luxembourg
[2]Division of Cardiology, Department of Internal Medicine, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China.
[3]School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China
[4]Clinica y Maternidad Suizo Argentina, Buenos Aires 1461, Argentina
[5]Centro Integral Cardiovascular, Gualeguaychú, Entre Ríos, Argentina
[6]School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China
[7]Department of Plant Sciences, Cambridge University, Cambridge CB2 3EA, United Kingdom
[8]Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai 200052, China
[†]Equal contributor, [*]Corresponding authors: yye@hust.edu.cn, yangxiaoyun321@126.com, jmg77@cam.ac.uk

ABSTRACT

Atrial Fibrillation (AF) is the most common cardiac rhythm disorder. Advance knowledge of an imminent switch from sinus rhythm (SR) to AF could prompt patients to take preventive actions to avoid AF, like taking oral antiarrhythmic drugs. The question is whether there is information, even if subtle, in the minutes prior to AF to indicate an imminent switch from SR. This paper shows that, for the vast majority of patients, the answer is affirmative. On test data, our algorithm can predict the onset of AF on average 31 minutes before it appears, with an accuracy of 83% and an F1-score of 85%. The predictions were based on deep learning and data from 350 patients, plus an external validation of 48 patients. Overall, the proposed method has low computational complexity and can be embedded in common wearable devices for continuous heart monitoring and early warning of AF onset.

Atrial fibrillation (AF) is the most common cardiac arrhythmia worldwide; the estimated number of individuals with AF in 2010 was 33.5 million[1]. Hand in hand with the growing prevalence of AF, healthcare costs continue to increase, mainly because of hospitalization and treatment costs[2]. AF episodes contribute to emergency department presentations because of high symptom burden and heart failure decompensation from tachycardiomyopathy. Moreover, as the aging of the population increases, the number of patients with AF is expected to further increase[3].

Clinicians can use catheter ablation, rate control therapy, and anticoagulation therapy to reduce the occurrence of stroke, systemic embolism, hospitalization, and deaths for patients with persistent or permanent AF, normally diagnosed by routine electrocardiography. Nevertheless, many patients with paroxysmal AF are misdiagnosed with resting electrocardiogram (ECG) or even when they are monitored with Holter devices for 24 hours, and are at high risk of stroke and systemic embolism due to repeated occurrences of AF[4,5]. Even though implantable ECG monitoring recorder can be used to detect paroxysmal AF, the method is invasive and difficult to popularize and apply[6].

Maintaining sinus rhythm (SR) on patients can reduce symptoms of AF and further prevent the atrial remodeling that enhances susceptibility to future episodes[7]. Consequently, an early prediction of AF episodes in patients with paroxysmal AF can prevent emergency department presentations and associated healthcare costs. Yet the appropriate identification of patients with a high likelihood of AF onset, and its early-warning prediction within the near future, are challenging problems in the clinical setting[8]. It is the need to overcome these difficulties that motivates this study: the aim is to develop a method that continuously monitor patients to predict and give early warnings of imminent AF onset, using only simple wearable devices such as smartwatches.

The automated *detection* of AF regimes from recorded time-series data is a well-studied problem in the literature[9–11], with recent approaches based on machine learning and neural networks achieving over 99% accuracy in the classification task[12–16].

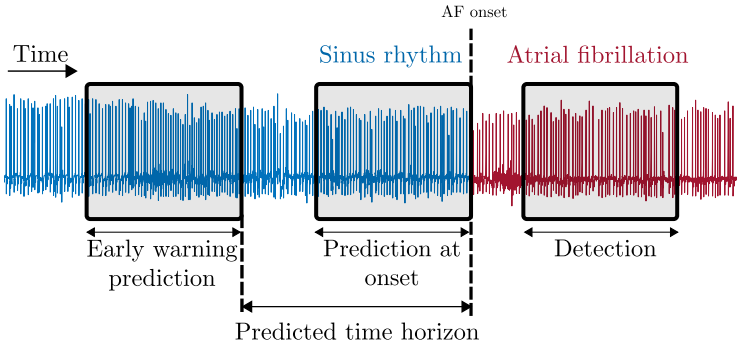

Figure 1. Detection versus prediction. Early-warning AF prediction (left), AF prediction at onset (middle), and AF detection (right). All methods are based on time-series windows sampled at different instants with respect to AF onset.

Nonetheless, *prediction* of the onset of AF is still an open problem[17]. Several well-established methods make long-term risk assessment of AF in the order of months and years[18–21], without the capability of real-time predictions. Past work on short-term predictions can be classified into two main categories. The first category includes methods applied to one or multiple ECG leads, involving features extracted from premature atrial contractions, atrial or ventricular ectopic beats, and spectral analysis[22–31]. The second category focuses on R to R intervals (RRI) analysis, often consisting of handcrafted feature extraction and power spectral analysis[32–38]. Or a combination of both, which typically involves ECG morphological features and RRI analysis combined with convolutional neural networks (CNN) and automatic feature extraction algorithms[18,39–50]. Such handcrafted features entail serious limitations: they are computationally expensive, can lead to human bias, and require long-term samples, making them unsuitable for real-time monitoring applications[51].

Overall, despite these recent advances, important constraints remain. The main one is that all these methods provided "predictions of AF at the onset". In other words, the data window used for prediction comprises the time-series signal sampled right up to the onset of AF and, hence, provides *zero* early warning (Fig. 1, middle). Standard detection algorithms use AF data (Fig. 1, right). This paper, in contrast, proposes a method for accurate early-warning prediction of AF far ahead of the onset of AF (Fig. 1, left). Table S1 shows the distinction between preceding works and the proposed method.

A limitation of ECG data is that long-term recordings are difficult to acquire, requiring the inconvenient and impractical use of Holter devices or patches. Though simple smartwatches and smart bands can continuously capture RRI during the individuals' daily life, existing methods using RRI tend to require large time-series windows of up to 30min length for prediction at onset[29,37,38]. This limits its application for short-term predictions. Moreover, when the RRI duration was reduced to less than 30min in previous methods, a poor prediction accuracy was obtained[51].

This paper presents a method, based on artificial intelligence (AI), for real-time monitoring and early warning of AF from moving RRI windows with short-term duration, entitled AFiRMo (Atrial Fibrillation Real-time Monitoring). To circumvent the above limitations, we provide a variable time horizon of prediction, with RRI samples of 30s to train a deep CNN. Hence, we do not require feature selection and use significantly shorter RRI samples compared to previous work. Moreover, our method does not require ECG data, and uses instead RRI signals that can be sampled by affordable and easy to wear pulse signal recorders, such as smartwatches or smart fitness bands. These devices can be used continuously by patients, paving the way for real-time monitoring algorithms that learn and monitor long-term cardiac dynamics.

Results

Data description

The original dataset from Tongji Hospital (Huazhong University of Science and Technology, Wuhan, China) consists of long-term 12-lead ECG Holter from 595 patients, where each ECG is recorded in SR at baseline and includes at least one AF episode. This study was approved by the Ethical Committee of Tongji Hospital with Institutional Review Board Approval number of TJ-IRB20220423. The beginning and end of individual AF episodes were labeled by experienced cardiologists at Tongji Hospital. The records have an average duration of 22.2 ± 2.2 hours, with a sampling frequency of 128 Hz and a resolution of 12 bits. We excluded records that did not have SR or AF episode. Records starting from AF were also excluded since the section of ECG preceding AF cannot be segmented. We considered only AF episodes with a duration of 10min or longer. Finally, records that have significant noise artifacts before AF onset were excluded (by checking if the percentage of missing R peaks within a 5min sliding window is above a threshold of 15%). After the exclusion criteria are applied, the remaining 350 records were used in this study. The cohort was divided into two groups following a chronological order between

Table 1. Characteristic of the patients.

Characteristic	Training Cohort (total = 280)	Test Cohort (total = 70)
Age $<$ 65 (mean)	115 (55 years old)	31 (54 years old)
Age $\geq$ 65 (mean)	165 (73 years old)	39 (73 years old)
Male	163	26
Female	117	44

2014 and 2019: the first 80% (280 patients) were used for the training/cross-validation of the model (252 for training and 28 for validation) and the last 20% (70 patients) for testing (Table 1).

To externally validate the performance of AFiRMo on "out-of-distribution" datasets, we considered additional ECG data collected from patients with AF from two healthcare centers in different countries: the Clínica y Maternidad Suizo Argentina (53 patients of 24h ECG) and the open-access data Atrial Fibrillation Prediction Database (AFPDB) from Physionet[52] (75 patients of 30min ECG). Applying the same exclusion criteria described above resulted in a total of 48 patients for external validation: 8 from Argentina and 40 from Physionet. The Physionet database consisted of 50 healthy controls (SR) and 25 AF patients (with 30min ECG just before AF onset and 5min right after). Of those, 5 AF patients presented AF with duration shorter than 1min and were excluded. Also, 2 records from healthy controls presented AF and heavy distortion due to artifact noise and were excluded. Hence, overall, there were 20 records to predict AF plus an equal number of 20 randomly selected records from healthy subjects for control.

AFiRMo algorithm for early warning prediction of atrial fibrillation

The AFiRMo algorithm for early warning of the onset of AF is divided into two stages. First, we train AFiRMo to detect three cardiac rhythms: SR, AF, and Pre-AF (the instances just before AF onset). Second, this model sequentially analyzes RRI data to monitor the probability of an imminent switch to AF. When this probability crosses a predefined threshold, it triggers a warning. Here, we provide an overview of the algorithm. Further details can be found in Methods and Supplementary Material.

Figure 2 illustrates the first stage of AFiRMo's algorithm. The ECG Holter recordings are segmented into three classes: SR, Pre-AF, and AF segments (Fig. 2a). The AF segments were labeled by cardiologists. The Pre-AF segments are labeled as the ECG data preceding the AF onset characterized by high RRI variability (Fig. 5), in contrast with SR segments that typically have lower variability[53]. A sliding window in the ECG data extracts segments of 30s which are converted to RRI data (Fig. 2b,c) and then to recurrence plots (Fig. 2d), which serve as inputs to the CNN. Finally, the prediction is performed by a deep CNN (Fig. 2e). The outputs of the network are the probabilities of the input belonging to each of the three regimes: $P(\text{SR})$, $P(\text{Pre-AF})$, and $P(\text{AF})$ (Fig. 2f).

The second stage of AFiRMo's algorithm computes the probability of a patient switching to AF in the near future from the outputs of the trained CNN. Define the probability of danger as $P(\text{danger}) = P(\text{Pre-AF}) + P(\text{AF})$, which represents the probability that a sliding window is in either Pre-AF or AF states. From the time-series RRI data, the algorithm sequentially generates a sliding window of 30s every 15s. For each new window, the recorded 30s time-series are converted into a recurrence plot and fed to the CNN for classification. Figure 3a illustrates the probability of danger computed by AFiRMo for a representative example. Since the probability of danger has very high variability, we implemented a non-anticipative moving average window to filter this high-frequency noise and smooth the output (Fig. 3b). A binary early-warning indicator ("danger" or "no danger") also requires the selection of a particular threshold of the probability. These two hyperparameters, the moving average window length and probability threshold, can be optimized to maximize different performance metrics, depending on the needs of particular patients.

The performance of AFiRMo was evaluated on a time series of 60min of sequential data before AF onset and also far from AF (in SR), selected patient wise. The selection of samples far from AF is performed randomly, at least 2 hours before AF, to guarantee that the median value of the coefficients of variation of the RRI signals (computed over the selected 60min sample) is close to the median value associated with the SR distribution computed over all patients (Fig. S2a). For the validation data (corresponding to the cross-validation with the highest performance in the CNN), as expected, it is not possible to simultaneously maximize all performance measures. For example, Figs. S2a-c show that the maximum predicted time horizon until AF onset (that is, the instant of the first early warning until AF onset, see Fig. 3b) is achieved at low thresholds (Fig. S2b,c). However, the accuracy is very low for those values (bottom of Figs. S2a). Likewise, when the accuracy is maximized at 88.3%, the predicted time horizon is relatively short.

To achieve a tradeoff in the validation set, we searched for hyperameters where the accuracy, sensitivity, specificity, and F1 score are all greater than 80%, and mean and median predicted time horizon are above 30min. There was a total of 34

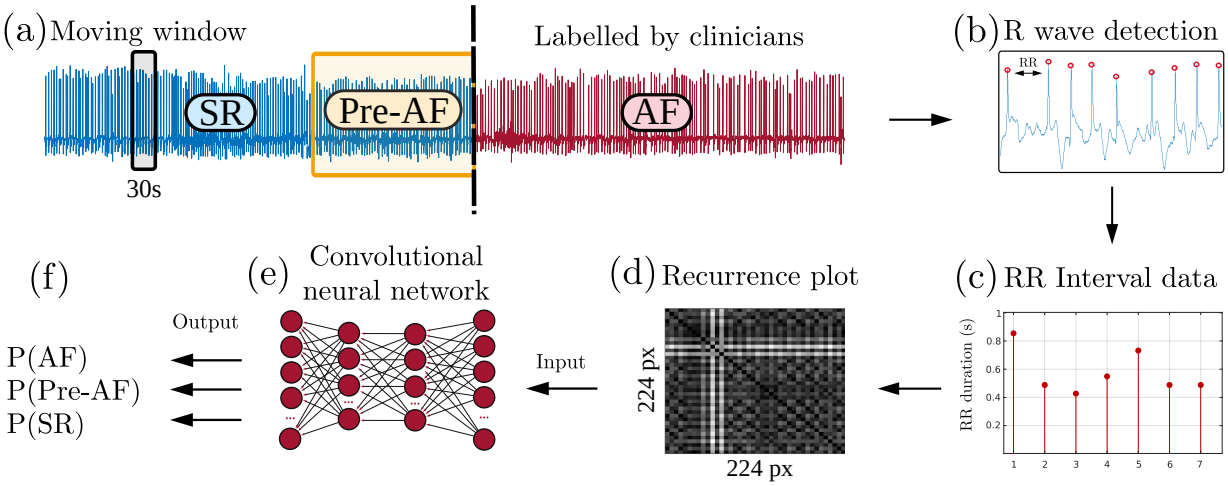

Figure 2. Pipeline of the first stage of AFiRMo. (a) Each ECG record is split into three classes: SR, Pre-AF, and AF. (b) R peaks are detected from a sliding window of 30s in the ECG data. (c) RRI signal is generated from the R peaks. (d) A recurrence plot is generated from the RRI signal. (e) A deep CNN is trained using the recurrence plots as inputs. (f) Network output are the probabilities of the sampled data to belong to each of the three classes (SR, Pre-AF, and AF).

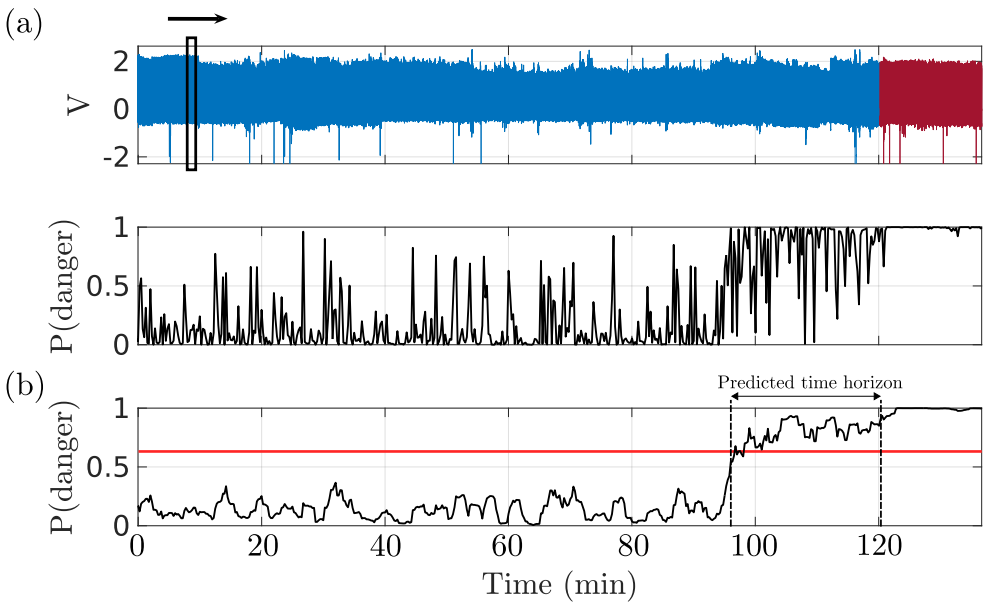

Figure 3. Second stage of AFiRMo: indicator of early warning. (a) Probability of danger computed by AFiRMo as a function of time for a representative patient from the test dataset. The sample images are generated by taking a sliding window of 30s every 15s and the probability of danger is computed for every sampled window. (b) Average probability of danger, computed from a non-anticipative moving average window of 7 samples to smooth the probability fluctuation. The red line is a threshold of 0.57 that will trigger a warning of an imminent AF onset.

hyperparameters satisfying this criterion. We picked the smallest moving average window, since this leads to lower computation and memory usage in smart devices, and, among those, the one that maximizes accuracy, F1 score, and mean and median predicted time horizon. The resulting hyperparameters are a moving average window size of 7 samples (corresponding to a 1.5min window) and a threshold of 0.57. This produced an accuracy of 86.7%, F1 score of 87.5%, sensitivity of 93.3%, and specificity of 80%. The mean (median) predicted time horizon until onset of AF is 31.4min (36.3min). Figure 3b illustrates the

moving average of the probability of danger and the threshold for a representative patient. In this example, an early-warning signal is alerted to the patient by AFiRMo with 22min in advance to the AF onset. These hyperparameters can be adjusted according to different objectives by giving more importance to, for example, sensitivity, specificity or size of the predicted time horizon.

Performance on test data

AFiRMo was evaluated on the test data set as an early warning of the onset of AF. The results are summarized in Fig. 4. For a fixed moving average window, the choice of threshold depends on the needs of a patient. For example, for a healthy user, the threshold could be higher and more specific to avoid false positives. For a patient at risk, the threshold could be smaller and more sensitive to reduce false negatives. Figure 4 presents results for different thresholds (the default 0.57, plus 0.73 and 0.88) to contrast tradeoffs between sensitivity and specificity. Figure 4a shows the mean and median values of the predicted time horizon of the AF onset across all patients, i.e., the time from the first early warning until AF onset (Figs. 1 and 3b). With a threshold of 0.57, AFiRMo predicts AF with an average (median) of 31min (38min) before its onset and an accuracy of 82.7%. AF in younger patients (less than 65 years old) can be predicted slightly earlier than for older patients: means of 32.5min versus 29.4min, respectively (Fig. 4a).

As seen in Fig. 4a, the predicted horizon depends on the threshold. Overall, a smaller threshold leads to a larger predicted horizon and, hence, a higher number of patients predicted to be in danger before the onset of AF. Figure 4b shows the fraction of patient to be predicted in danger as a function of the time until AF onset. For a threshold of 0.57, around 50% of patients are predicted to be in danger at least 37min before AF onset. To better observe the tradeoffs, Fig. 4c shows the performance of AFiRMo as a function of the threshold for different metrics (see Supplementary Material for definitions of performance metrics). The curves cross at a threshold of 0.74 (with a value of 83.6%).

The area under the receiver operating characteristic curve (AUROC) and the precision-recall curve (AUPRC) are 0.90 and 0.88, respectively (Fig. S3). This shows that the performance is balanced among the "danger" and "SR" samples. Figure 4d display the confusion matrices computed on 75 episodes of AF out of the 70 patients in the test set for different values of the threshold. It contains the true and false number of predictions by AFiRMo with respect to positive (danger) and negative (SR) episodes. Finally, Fig. 4e presents trade-offs between AFiRMo accuracy and the predicted time horizon as a function of the threshold. As expected, an increase in the predicted time horizon typically results in a decrease in accuracy, and vice-versa.

Performance on ECG data

The above results were shown for predictions based on RRI data (computed from ECG data). Here, we investigate by how much the performance increases when using the original ECG data. As explained, RRI data are preferable since they can be obtained from simple wearable devices that can be worn 24/7. In contrast, 24/7 continuous measurements of ECG using Holter or patches devices are impractical for long-term usage.

The maximum accuracy of the validation set of the ECG data is 89.3%, which results in a relatively low mean predicted time horizon. Following a similar tradeoff decision as described above, we obtained 8 hyperparameters that satisfy the criterion. The smallest moving average is 6 samples (corresponding to 1.25min windows), and the threshold of 0.48 maximizes accuracy, F1 score, and mean and median predicted time horizon. The mean predicted time horizon is 32.5min and the median 43.4min (Fig. S2d-f). With these hyperparameters, Fig. S4 summarizes the performance of AFiRMo on the test ECG data. The mean (median) predicted time horizon before onset of AF were 32.1min (35.6min), with a similar accuracy of 82.4% of the RRI data. F1, sensitivity, and specificity were 84.5%, 96.0%, and 68.9%, respectively.

Both the AUROC and AUPRC increased to 0.95 and 0.96. The cut-off balance between all the performance metrics is reached at a threshold of 0.76 with a value of 86.4%. These values are about 3.3% higher than the model performance using RRI data alone. It is not surprising that the ECG data has improved performance. However, it is surprising that the improvement was relatively small, considering the richer highly-sampled continuous-time ECG data compared to the simpler low-sampled discrete-time RRI data. This shows that prediction of AF onset can be performed with only RRI data.

Performance on external center data

To further validate the performance of AFiRMo on "out-of-distribution" datasets, we considered ECG data collected from patients with AF from two healthcare centers in Argentina (8 patients). With the same hyperparameters from the RRI validation set of 7 samples moving window and 0.57 threshold, we obtained an accuracy of 77.8% and a mean (median) predicted time horizon until AF onset of 47.8min (58.0min). Figure S5 summarizes the model performance for this external center dataset. The AUROC and AUPRC are 0.75 and 0.62, respectively. A higher threshold leads to slightly more balanced results. For example, for a threshold of 0.71, the accuracy increases to 83.3%, with mean (median) predicted horizon of 32.6min (41.1min). Although there were only 8 patients in this dataset, the performance is similar to the above RRI test dataset.

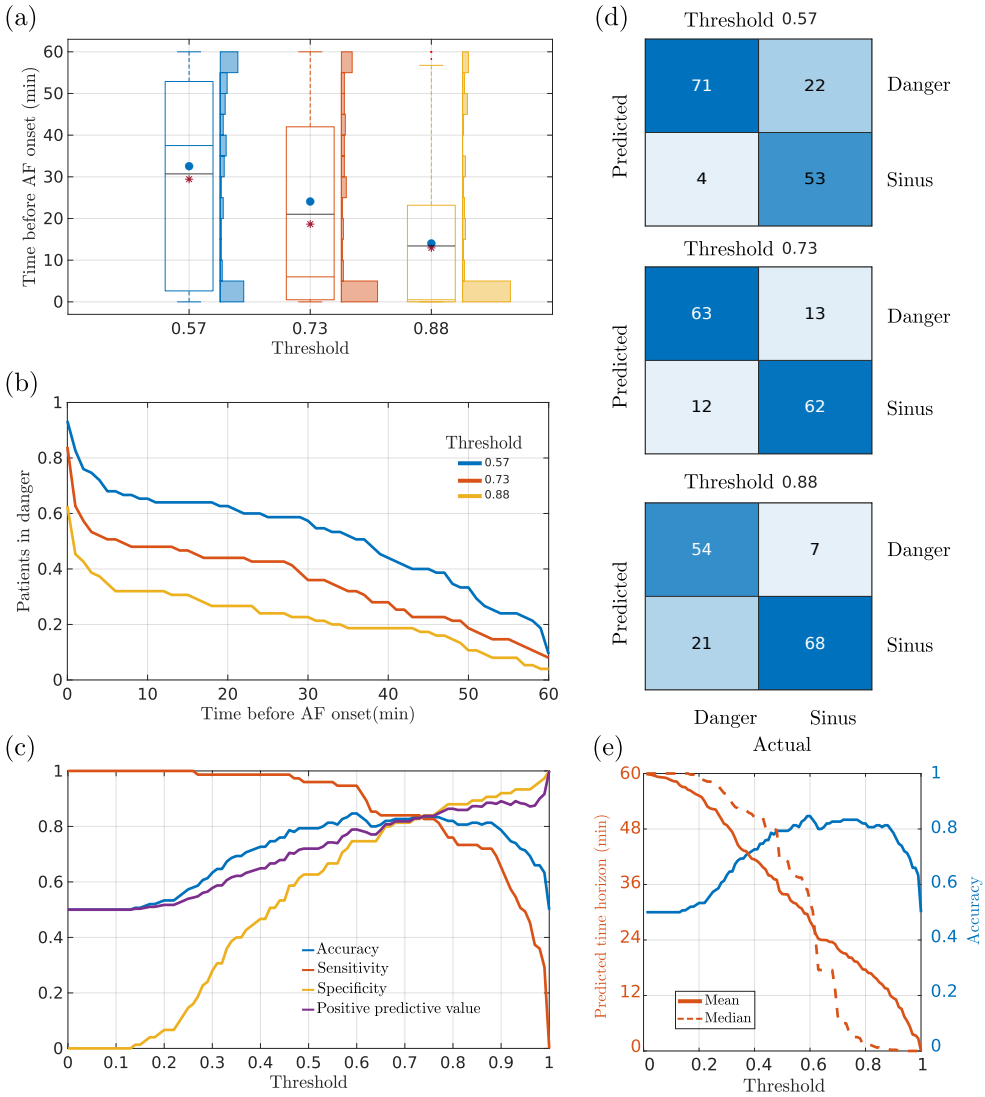

Figure 4. Performance of AFiRMo with 7 samples moving average. (a) Box plots of the predicted time horizon until AF onset for different probability thresholds across all patients. Median and mean values are marked by coloured and black lines, respectively. Blue circles and red asterisks represent the means of patients younger and older younger than 65 years, respectively. Histograms of the predicted time horizon are shown on the right side of the box plots. (b) Fraction of patients predicted to be in danger as a function of time before the AF onset for different thresholds. (c) Performance metrics as a function of the probability threshold. (d) Confusion matrices for different values of the threshold. (e) Mean (solid line) and median (dashed line) of the predicted time horizon and model accuracy as a function of the probability threshold.

Performance on Physionet AF prediction challenge

Finally, we further tested AFiRMo's performance on the open-access data Atrial Fibrillation Prediction Database (AFPDB) from Physionet[52] (20 AF patients and 20 healthy patients). With the same hyperparameters from the RRI validation set of 7 samples moving window and threshold of 0.57, we obtained an accuracy of 70.0% and a mean (median) predicted time horizon of 12.9min (9.1min) out of 30min (the ECG data collected in this dataset is 30min long and, hence, the predicted horizon time is limited to 30min). Figure S6 summarizes the AFiRMo performance for this dataset. The AUROC and AUPRC are 0.76 and 0.79 respectively.

Discussion

This paper has developed AFiRMo, an automated prediction method for early warning of AF onset based on deep CNN and short-term RRI signals. Our method takes 30s RRI samples every 15s and computes the probability of danger of imminent AF onset. A moving average of 7 samples, corresponding to 1min30s, smooths this probability to filter out noise. The key feature is the early and continuous rise of the probability of danger when approaching AF, providing an early warning when this probability crosses 0.57. On the test data (70 patients) and two external validation sets (8 and 40 patients), AFiRMo predicted AF onset on average 31min, 48min, and 13min in advance with an accuracy of 83%, 78%, and 70%, respectively. AF onset for younger patients (less than 65 years old) were consistently predicted earlier than older patients by an average of about 3.1min. If AFiRMo is implemented in smartwatches or other wearable devices that can record RRI signals, which can be worn continuously by patients, this early warning of AF could give sufficient time for patients to take oral antiarrhythmic drugs in advance to try to prevent the onset of AF. Moreover, anticoagulation therapy could also be used simultaneously. Finally, AFiRMo can significantly improve the positive screening rate for paroxysmal AF.

This paper is a proof of concept, showing that it is possible to predict and give early warnings to a cardiovascular disease. Hence, it opens the door to further develop tools, devices and apps that can continuously monitor both patients and healthy subjects. Moreover, the more data devices obtain from users, the more they can be personalized by learning unique individual disease traits. We expect that in the near future these prediction devices will be widely used by the general population to alert us to the onset of not only AF but also other more catastrophic cardiovascular diseases such as cardiac arrest, with the potential to save numerous lives.

AFiRMo introduced two parameters that can be tuned by physicians depending on the clinical application: the probability threshold (danger indicator) and the moving average. These two parameters are roughly inverse to each other: lower (higher) moving averages required higher (lower) thresholds (Fig S2). Our choice in this paper was based on a simple tradeoff decision to keep both accuracy, F1 score, and prediction horizon relatively high. Placing greater weight on one of these objectives would lead to different results. For example, for a particular moving average, smaller thresholds yield more sensitive models, which can be used for high risk patients. On the other hand, higher thresholds lead to more specific models and reduce false positives, which may be more suitable to monitor healthy patients.

We analyzed all false predictions, negative and positive, to gain insight on the algorithm. We focused on the RRI data with the chosen hyperparameters of 7 samples moving average and threshold 0.57 (Fig. 4). Table S4 summarizes the observations on the incorrectly classified patients. Of the 4 false negatives, one patient had a sudden AF onset with a very stable SR beforehand. The other 3 patients had a combination of tachycardia, bradycardia, unstable base lines, and noisy signals before AF onset. There were 22 false positives. Among those, 13 had premature atrial contractions (PACs), 5 had premature ventricular contractions (PVCs), 6 had unstable baselines, 4 had sinus tachycardia and one had atrial flutter. Moreover, 15 records were very noisy. We speculate that some of these false positive events correspond to moments where the heart was close to switch from SR to AF and, for some reason, it did not. Due to a number of conditions (e.g., stress or stimulants), heart dynamics can be pushed towards the tipping point that leads to a dynamical transition from SR to AF. It is possible that in some false positives the heart was close to switch to AF, but reverted back to SR. Especially those patients with PACs (13 out of 22), which are well-known precursors of AF[54], and highly connected to the occurrence or even trigger of AF. Thus, monitoring PACs seems significant to predict AF. Noise was also a major factor in most false predictions. Hence, it is necessary to treat patients' skin with saline or disinfectant before wearing ECG devices to ensure the electrodes are well connected to the skin and decrease noise during recording.

Compared with ECG data, results using RRI data had a slightly reduced performance. On the test data, both had a similar accuracy of 83% while the average prediction horizon was 32.1min for the ECG and 30.8min for the RRI data. This slight reduction is compensated by the fact that RRI data can easily be obtained continuously from simple wearable devices, such as a smartwatch. While critical patients could still use Holter or patches devices to measure ECG and take advantage of their better performance, this is not reasonable for other patients and the general population. Overall, on a standard computer (2GHz Dual-Core Intel Core I5, 8GB RAM), the total computational time for each window was around 100ms. This is considerably less than the sampling time of 15s, making it feasible to implement AFiRMo in smartphones to process the streamed data in real-time from wearable devices[55]. Moreover, the deep learning model used in this paper, EfficientNetV2, can be adapted for mobile devices through the TensorFlow Lite framework. Finally, noise played a strong role in false predictions. However, the use of smartwatches and smart bands can reduce noise by being worn tight on the wrist.

There are several opportunities for improving performance. This paper used the data from 280 patients to train AFiRMo. However, individual patients have specific traits that may not be picked by an "average" model across all patients. Hence, continuously monitoring a single patient could lead to a personalized AFiRMo that regularly retrains itself to maximize performance for that individual. Under this modality, a physician could eventually prescribe AFiRMo for continuous, personalized patient follow-up. We expect that in the near future, algorithms like AFiRMo will be embedded in smart devices to monitor our health 24/7, to warn us against imminent dangers, and to seek immediate medical attention. AFiRMo may also

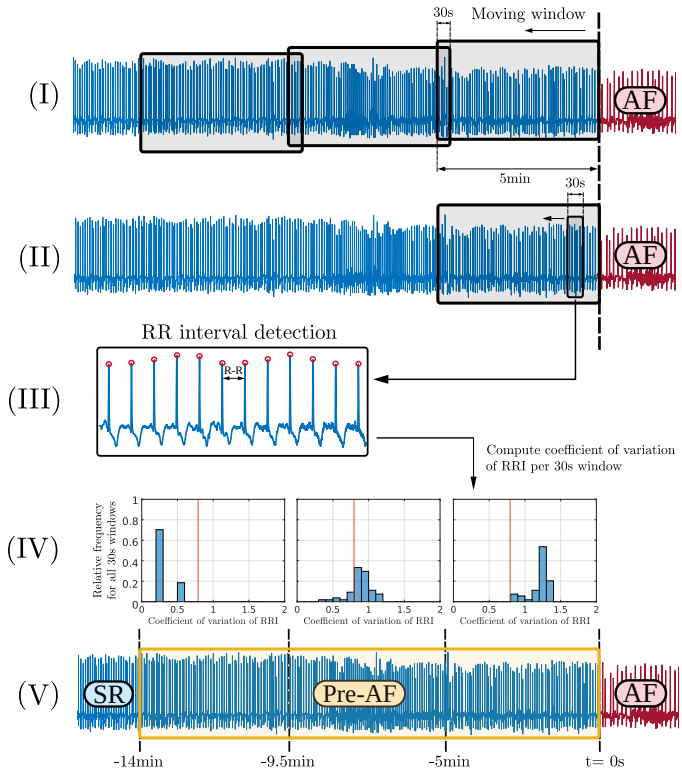

Figure 5. Pre-AF labeling process for a representative patient. (I) Starting from the AF onset and travelling back in time, a sliding window is generated to extract ECG samples of 5min with 30s overlapping. (II) For each 5min window, a second sliding window is generated, to extract samples of 30s every 5s. (III) R waves are detected for each 30 s window and the RR intervals are calculated. (IV) The coefficient of variation of the RR intervals of all 30s windows within a 5min window is calculated. The process is then repeated for each 5min window to generate a histogram of coefficient of variation of *RRI*. (V) When the histogram of σ_{RRI} has a median less than 0.7, the Pre-AF section is segmented from the beginning of this last window until the onset of AF. In this example, Pre-AF lasts 14min before the onset of AF.

be improved by running a cross-validation to tune the two hyperparameters on data from different centers and countries to capture patient heterogeneity.

Methods

Pre-AF ECG segments

The labeling of the Pre-AF ECG segments consists of five steps, illustrated in Fig. 5 for a representative patient. First, starting from the AF onset (labeled by clinicians), we generate a sliding window to extract ECG samples of 5min with 30s overlapping; the sliding window moves backward in time. Second, within each 5min window, we generate a second sliding window to extract smaller samples of 30s every 5s. Third, we use the Pan–Tompkins algorithm[56,57] to detect R waves for each 30s window and calculate the RRI. Fourth, we compute the coefficient of variation of the RRI for each 30s window and generate the corresponding histogram for each 5min window. Fifth, we analyze the evolution of the distribution of frequencies until their median is less than 0.7. The threshold of 0.7 is selected as the interception point between the distributions of frequencies of the coefficient of variation of the AF and SR regimes for all patients from the training set (Fig. S1). Below this threshold, the heart dynamics have low variability and can be associated with SR[58]. At this fifth and last stage, the Pre-AF section is segmented, including the beginning of this last window until the onset of AF.

Data segmentation by an AI model

In the first stage of AFiRMo's algorithm, the ECG Holter recordings are segmented into three classes, as illustrated in Fig. 2a: SR, Pre-AF, and AF segments. As explained above, Pre-AF segments are labeled as the ECG data preceding the AF onset

characterized by high RRI variability, in contrast with SR segments that typically have lower variability[53]. Pre-AF segments vary in length from patient to patient[58] and may also vary within multiple onsets of AF for the same patient due to morphological and electrical changes in the heart over time[66].

The baseline wander and interference noise from the ECG are reduced using a band-pass filter with cut-off frequencies of 0.5 and 40 Hz. From a moving window of 30s of an ECG record (Fig. 2a), we detect the R peaks using the Pan-Tompkins algorithm, which has an average error rate of about 1%[56,57] (Fig. 2b). Then, the difference between heartbeats is computed (i.e., from R to R peaks) to generate the RRI signal (Fig. 2c), which is then converted into a 2-dimensional recurrence plot[61] (Fig. 2d). Recurrence plots provide a visual representation of RRI signals and can be used to identify hidden periodicities in physiological signals[11,67,68] and detect dynamical transitions in time-series data[59,69]. For instance, the SR and AF states are represented by spatiotemporal recurrent patterns of particular rhythmicity, which can be detected by recurrence plots[70] (see Fig. S7 for details). Thus, we hypothesize that an AI-enabled model can capture such patterns, including the transient Pre-AF state before AF onset. Finally, to take advantage of existing and efficient deep learning algorithms on image recognition[71], we convert each recurrence plot into a 2-dimensional image that is then fed to the deep CNN (Fig. 2e).

For building AFiRMo, we implemented EfficientNetV2, a deep CNN with 479 layers, developed by Google in 2021[72]. The EfficientNetV2 is a modified and optimized version of EfficientNet[73], winner of the ImageNet 2019 competition[74]. The input of EfficientNetV2 are images of size 224×224 pixels. The last fully connected layer was modified to perform the classification among the three classes. Hence, three probabilities are output by the network: $P(\text{SR})$, $P(\text{Pre-AF})$ and $P(\text{AF})$, which correspond to the probability of the input data belonging to each of the three regimes, and satisfying $P(\text{SR}) + P(\text{Pre-AF}) + P(\text{AF}) = 1$ (Fig. 2f).

Training of the deep convolutional neural network

AFiRMo is enabled with artificial intelligence (AI) for early warning detection of AF. It was trained and cross-validated on random samples from 280 patients. The network was trained using categorical cross-entropy as the loss function, ADAM as the optimizer[75], and stochastic gradient descent as the objective function optimizer[76]. To compensate for class imbalance, the loss function was weighted according to the proportion of the sample labels (SR, Pre-AF, and AF). The training ended after the validation-loss did not decrease over 8 consecutive epochs.

We investigated the optimal length of the sampling window to generate the recurrence plot from the RRI data, starting from 10s up to 5min. We computed the average accuracy to predict individual samples, from the 10-fold cross-validation of the EfficientNetV2 using different windows length (Table S2). The best performance was obtained using a window length of 30 seconds, which has been also reported in[39]. Changes in performance are associated with tradeoffs between the number of samples generated and the length of the window. The wider the window, the lower the amount of samples obtained for training, hence reducing the effectiveness of AFiRMo to properly generalize the data. On the other hand, a smaller window length may lead to information losses[77]. After fixing the sampling window length to 30s, we compared the performance of AFiRMo with two other network benchmarks: 1-D CNN and LSTM, commonly used on arrhythmia detection and prediction[78,79] (Table S3). The proposed AFiRMo model over-performed the benchmark networks, achieving an average validation accuracy of 0.74 and a good generalization of the data as represented by a small standard deviation of 0.03 across all 10 folds. Finally, the best model was selected for performance analysis on the test set of 70 patients.

Data availability

The data was provided by the Tongji Hospital from China and the Clínica y Maternidad Suizo Argentina, acquired between 2014 and 2022. To protect patients' privacy, the data was anonymized. The data collection teams from each center handled sample collection and anonymization. The algorithm development team received anonymized data containing only age and gender information for the subsequent algorithm development. The study design was evaluated and exempted from a full review by the Huazhong University of Science and Technology Institutional Review Board (approval number: TJ-IRB20220423), and approved by the Ethics Review Panel of the University of Luxembourg (approval number: ERP 22-057 RTMonitor). All data have been obtained according to the principles of the declaration of Helsinki. Training and validation sets from Tongji Hospital will be accessible pending approval by Xiaoyun Yang. Data use is limited to non-commercial research. Readers can request access to the training and validation sets to Xiaoyun Yang. The data used as external validation of this study are available from the open-access PAF Prediction Challenge Database `https://physionet.org/content/afpdb/1.0.0/`.

Code availability

Data preprocessing and segmentation were implemented using the MATLAB software. The neural network was implemented on the Keras Framework with Tensorflow backend on Python 3.7. Codes have been deposited on GitHub at `https://github.com/marino-gavidia/AFiRMo.git`.

References

1. S. S. Chugh, *et al.*, Worldwide epidemiology of atrial fibrillation: a global burden of disease 2010 study, *Circulation* **129**, 837–847 (2014).
2. E. J. Benjamin, *et al.*, Heart disease and stroke statistics 2019 update: a report from the american heart association, *Circulation* **139**, e56–e528 (2019).
3. C. T. January, *et al.*, 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society, *Journal of the American College of Cardiology* **64**, e1–e76 (2014).
4. S. B. Uittenbogaart, W. A. Lucassen, F. S. van Etten-Jamaludin, J. R. de Groot, H. C. van Weert, Burden of atrial high-rate episodes and risk of stroke: a systematic review, *EP Europace* **20**, 1420–1427 (2018).
5. G. L. Botto, G. Tortora, M. C. Casale, F. L. Canevese, F. A. M. Brasca, Impact of the pattern of atrial fibrillation on stroke risk and mortality, *Arrhythmia & Electrophysiology Review* **10**, 68 (2021).
6. A. AlTurki, M. Marafi, V. Russo, R. Proietti, V. Essebag, Subclinical atrial fibrillation and risk of stroke: past, present and future, *Medicina* **55**, 611 (2019).
7. E. N. Prystowsky, Management of atrial fibrillation: therapeutic options and clinical decisions, *The American Journal of Cardiology* **85**, 3–11 (2000).
8. R. E. Wilson, K. L. Rush, L. Hatt, R. C. Reid, C. G. Laberge, The symptom experience of patients with atrial fibrillation before their initial diagnosis, *Journal of Cardiovascular Nursing* **35**, 347–357 (2020).
9. A. Rizwan, *et al.*, A review on the state of the art in atrial fibrillation detection enabled by machine learning, *IEEE Reviews in Biomedical Engineering* **14**, 219–239 (2020).
10. P. Panindre, V. Gandhi, S. Kumar, Comparison of performance of artificial intelligence algorithms for real-time atrial fibrillation detection using instantaneous heart rate, *IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI* pp. 168–172 (2020).
11. H. Zhang, *et al.*, Recurrence plot-based approach for cardiac arrhythmia classification using inception-resnet-v2, *Frontiers in Physiology* **12**, 648950 (2021).
12. P. Kora, A. Annavarapu, P. Yadlapalli, K. S. R. Krishna, V. Somalaraju, Ecg based atrial fibrillation detection using sequency ordered complex hadamard transform and hybrid firefly algorithm, *Engineering Science and Technology, an International Journal* **20**, 1084–1091 (2017).
13. A. Annavarapu, P. Kora, Ecg-based atrial fibrillation detection using different orderings of conjugate symmetric–complex hadamard transform, *International Journal of the Cardiovascular Academy* **2**, 151–154 (2016).
14. R. J. Martis, U. R. Acharya, H. Prasad, C. K. Chua, C. M. Lim, Automated detection of atrial fibrillation using bayesian paradigm, *Knowledge-Based Systems* **54**, 269–275 (2013).
15. A. H. Ribeiro, *et al.*, Automatic diagnosis of the 12-lead ECG using a deep neural network, *Nature Communications* **11**, 1760 (2020).
16. H. Zhu, *et al.*, Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study, *The Lancet Digital Health* **2**, e348–e357 (2020).
17. M. Aljanabi, H. Qutqut, M. Hijjawi, Machine learning classification techniques for heart disease prediction: A review, *International Journal of Engineering & Technology* **7**, 5373–5379 (2018).
18. Z. I. Attia, *et al.*, An artificial intelligence-enabled ecg algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction, *The Lancet* **394**, 861–867 (2019).
19. H. El Moaqet, *et al.*,A new algorithm for short term prediction of persistent atrial fibrillation, *IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies*, 1–6 (2017).
20. S. Biton, *et al.*, Atrial fibrillation risk prediction from the 12-lead electrocardiogram using digital biomarkers and deep representation learning, *European Heart Journal-Digital Health* **2**, 576–585 (2021).
21. S. Raghunath, *et al.*, Deep neural networks can predict new-onset atrial fibrillation from the 12-lead ecg and help identify those at risk of atrial fibrillation–related stroke, *Circulation* **143**, 1287–1298 (2021).
22. P. Langley, *et al.*, Can paroxysmal atrial fibrillation be predicted?, *Computers in Cardiology* **28**, 121–124 (2001).

23. W. Zong, R. Mukkamala, R. Mark, A methodology for predicting paroxysmal atrial fibrillation based on ecg arrhythmia feature analysis, *Computers in Cardiology* **28**, 125–128 (2001).

24. T. Thong, J. McNames, M. Aboy, B. Goldstein, Prediction of paroxysmal atrial fibrillation by analysis of atrial premature complexes, *IEEE Transactions on Biomedical Engineering* **51**, 561–569 (2004).

25. A. Yang, H. Yin, Prediction of paroxysmal atrial fibrillation by footprint analysis, *Computers in Cardiology* **28**, 401–404 (2001).

26. K. Aytemir, S. Aksoyek, A. Yildirir, N. Ozer, A. Oto, Prediction of atrial fibrillation recurrence after cardioversion by p wave signal-averaged electrocardiography, *International Journal of Cardiology* **70**, 15–21 (1999).

27. L. Clavier, J.-M. Boucher, R. Lepage, J.-J. Blanc, J.-C. Cornily, Automatic p-wave analysis of patients prone to atrial fibrillation, *Medical and Biological Engineering and Computing* **40**, 63–71 (2002).

28. C. Blanche, N. Tran, F. Rigamonti, H. Burri, M. Zimmermann, Value of p-wave signal averaging to predict atrial fibrillation recurrences after pulmonary vein isolation, *EP Europace* **15**, 198–204 (2013).

29. K. H. Boon, M. Khalil-Hani, M. Malarvili, C. W. Sia, Paroxysmal atrial fibrillation prediction method with shorter hrv sequences, *Computer Methods and Programs in Biomedicine* **134**, 187–196 (2016).

30. K. H. Boon, M. Khalil-Hani, M. Malarvili, Paroxysmal atrial fibrillation prediction based on hrv analysis and non-dominated sorting genetic algorithm iii, *Computer Methods and Programs in Biomedicine* **153**, 171–184 (2018).

31. R. Alcaraz, A. Martínez, J. J. Rieta, Role of the p-wave high frequency energy and duration as noninvasive cardiovascular predictors of paroxysmal atrial fibrillation, *Computer Methods and Programs in Biomedicine* **119**, 110–119 (2015).

32. Y. V. Chesnokov, Complexity and spectral analysis of the heart rate variability dynamics for distant prediction of paroxysmal atrial fibrillation with artificial intelligence methods, *Artificial Intelligence in Medicine* **43**, 151–165 (2008).

33. K. Tateno, L. Glass, Automatic detection of atrial fibrillation using the coefficient of variation and density histograms of rr and δrr intervals, *Medical and Biological Engineering and Computing* **39**, 664–671 (2001).

34. S. Sarkar, D. Ritscher, R. Mehra, A detector for a chronic implantable atrial tachyarrhythmia monitor, *IEEE Transactions on Biomedical Engineering* **55**, 1219–1224 (2008).

35. S. Dash, K. Chon, S. Lu, E. Raeder, Automatic real time detection of atrial fibrillation, *Annals of Biomedical Engineering* **37**, 1701–1709 (2009).

36. J. Lee, Y. Nam, D. D. McManus, K. H. Chon, Time-varying coherence function for atrial fibrillation detection, *IEEE Transactions on Biomedical Engineering* **60**, 2783–2793 (2013).

37. M. Mohebbi, H. Ghassemian, Prediction of paroxysmal atrial fibrillation based on non-linear analysis and spectrum and bispectrum features of the heart rate variability signal, *Computer Methods and Programs in Biomedicine* **105**, 40–49 (2012).

38. Y. Guo, *et al.*, Photoplethysmography-based machine learning approaches for atrial fibrillation prediction: a report from the huawei heart study, *JACC: Asia* **1**, 399–408 (2021).

39. U. Erdenebayar, H. Kim, J.-U. Park, D. Kang, K.-J. Lee, Automatic prediction of atrial fibrillation based on convolutional neural network using a short-term normal electrocardiogram signal, *Journal of Korean Medical Science* **34** (2019).

40. J. Cho, Y. Kim, M. Lee, Prediction to atrial fibrillation using deep convolutional neural networks, *International Workshop on Predictive Intelligence in Medicine* pp. 164–171 (2018).

41. H.-A. Tzou, S.-F. Lin, P.-S. Chen, Paroxysmal atrial fibrillation prediction based on morphological variant p-wave analysis with wideband ecg and deep learning, *Computer Methods and Programs in Biomedicine* **211**, 106396 (2021).

42. A. Jalali, M. Lee, Atrial fibrillation prediction with residual network using sensitivity and orthogonality constraints, *IEEE Journal of Biomedical and Health Informatics* **24**, 407–413 (2019).

43. A. Y. Hannun, *et al.*, Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network, *Nature Medicine* **25**, 65–69 (2019).

44. Y.-Y. Jo, *et al.*, Explainable artificial intelligence to detect atrial fibrillation using electrocardiogram, *International Journal of Cardiology* **328**, 104–110 (2021).

45. T. Mahmud, S. A. Fattah, M. Saquib, Deeparrnet: An efficient deep cnn architecture for automatic arrhythmia detection and classification from denoised ecg beats, *IEEE Access* **8**, 104788–104800 (2020).

46. H. Costin, C. Rotariu, A. Păsărică, Atrial fibrillation onset prediction using variability of ecg signals, *8th International Symposium on Advanced Topics in Electrical Engineering* pp. 1–4 (2013).

47. J. Kim, O. Sangjun, Y. Kim, M. Lee, Convolutional neural network with biologically inspired retinal structure, *Procedia Computer Science* **88**, 145–154 (2016).

48. Y. Shen, *et al.*, *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE, 2016), pp. 2419–2422.

49. Z. Li, *et al.*, A novel atrial fibrillation prediction algorithm applicable to recordings from portable devices, *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* pp. 4034–4037 (2018).

50. E. Ebrahimzadeh, *et al.*, Prediction of paroxysmal atrial fibrillation: a machine learning based approach using combined feature vector and mixture of expert classification on hrv signal, *Computer Methods and Programs in Biomedicine* **165**, 53–67 (2018).

51. I. Matias, *et al.*, Prediction of atrial fibrillation using artificial intelligence on electrocardiograms: A systematic review, *Computer Science Review* **39**, 100334 (2021).

52. G. Moody, A. Goldberger, S. McClennen, S. Swiryn, Predicting the onset of paroxysmal atrial fibrillation: The computers in cardiology challenge 2001, *Computers in Cardiology* **28**, 113–116 (2001).

53. A. Narin, Y. Isler, M. Ozer, M. Perc, Early prediction of paroxysmal atrial fibrillation based on short-term heart rate variability, *Physica A: Statistical Mechanics and its Applications* **509**, 56–65 (2018).

54. J. C. L. Himmelreich, *et al.*, Frequent premature atrial contractions are associated with atrial fibrillation, brain ischaemia, and mortality: a systematic review and meta-analysis, *EP Europace* **21**, 698-707 (2018).

55. C. Luo, *et al.*, Comparison and benchmarking of ai models and frameworks on mobile devices, *arXiv:2005.05085* (2020).

56. J. Pan, W. J. Tompkins, A real-time qrs detection algorithm, *IEEE Transactions on Biomedical Engineering* pp. 230–236 (1985).

57. H. Sedghamiz, Matlab implementation of Pan Tompkins ECG QRS detector. Code available at 10.13140/RG.2.2.14202.59841 (2014).

58. B. L. Lan, Y. W. Liew, M. Toda, S. H. Kamsani, Flickering of cardiac state before the onset and termination of atrial fibrillation, *Chaos* **30**, 053137 (2020).

59. L. L. Portes, A. N. Montanari, D. C. Correa, M. Small, L. A. Aguirre, The reliability of recurrence network analysis is influenced by the observability properties of the recorded time series, *Chaos* **29**, 083101 (2019).

60. N. Marwan, M. Carmen Romano, M. Thiel, J. Kurths, Recurrence plots for the analysis of complex systems, *Physics Reports* **438**, 237–329 (2007).

61. J.-P. Eckmann, S. O. Kamphorst, D. Ruelle, *et al.*, Recurrence plots of dynamical systems, *World Scientific Series on Nonlinear Science Series A* **16**, 441–446 (1995).

62. F. Takens, *Dynamical Systems and Turbulence, Warwick 1980* (Springer, 1981), pp. 366–381.

63. T. Sauer, J. A. Yorke, M. Casdagli, Embedology, *Journal of Statistical Physics* **65**, 579–616 (1991).

64. B. Kliková, A. Raidl, Reconstruction of phase space of dynamical systems using method of time delay, *Proceedings of the 20th Annual Conference of Doctoral Students* pp. 83–87 (2011).

65. C. Rhodes, M. Morari, False-nearest-neighbors algorithm and noise-corrupted time series, *Physical Review E* **55**, 6162 (1997).

66. P. Pławiak, Novel methodology of cardiac health recognition based on ecg signals and evolutionary-neural system, *Expert Systems with Applications* **92**, 334–349 (2018).

67. R. Sun, Y. Wang, Predicting termination of atrial fibrillation based on the structure and quantification of the recurrence plot, *Medical Engineering & Physics* **30**, 1105–1111 (2008).

68. N. C. Carvalho, L. L. Portes, A. Beda, L. M. Tallarico, L. A. Aguirre, Recurrence plots for the assessment of patient-ventilator interactions quality during invasive mechanical ventilation, *Chaos* **28**, 085707 (2018).

69. M. H. Trauth, *et al.*, Classifying past climate change in the chew bahir basin, southern ethiopia, using recurrence quantification analysis, *Climate Dynamics* **53**, 2557–2572 (2019).

70. F. Censi, *et al.*, Recurrent patterns of atrial depolarization during atrial fibrillation assessed by recurrence plot quantification, *Annals of Biomedical Engineering* **28**, 61–70 (2000).

71. Y. Xia, N. Wulan, K. Wang, H. Zhang, Detecting atrial fibrillation by deep convolutional neural networks, *Computers in Biology and Medicine* **93**, 84–92 (2018).

72. M. Tan, Q. Le, Efficientnetv2: Smaller models and faster training, *arXiv:2104.00298* .

73. M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, *International Conference on Machine Learning* pp. 6105–6114 (2019).

74. J. Wu, Y. Yu, C. Huang, K. Yu, Deep multiple instance learning for image classification and auto-annotation pp. 3460–3469 (2015).

75. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, *arXiv:1412.6980* (2014).

76. L. Bottou, *Neural networks: Tricks of the trade* (Springer, 2012), pp. 421–436.

77. K. H. Kraemer, R. V. Donner, J. Heitzig, N. Marwan, Recurrence threshold selection for obtaining robust recurrence characteristics in different embedding dimensions, *Chaos* **28**, 085720 (2018).

78. S. L. Oh, E. Y. Ng, R. San Tan, U. R. Acharya, Automated diagnosis of arrhythmia using combination of cnn and lstm techniques with variable length heart beats, *Computers in Biology and Medicine* **102**, 278–287 (2018).

79. S. Somani, *et al.*, Deep learning and the electrocardiogram: review of the current state-of-the-art, *EP Europace* (2021).

Acknowledgements
The authors want to thank the support from the Luxembourg National Research Fund (FNR) through PRIDE15/10907093/CriTiCS, and the National Natural Science Foundation of China under grants 92167201 and 82100531.

Author contributions statement
J.G. conceptualized the research; H.Z., S.D., and M.C. collected, curated, and annotated the data; M.E.G., J.G., A.N.M., Y.J., S.Z., and B.T. developed the AI model; M.E.G. implemented the codes and validated the results; M.E.G., H.Z., J.G., Y.Y., A.N.M., J.F., C.C., and X.H. analysed the results; J.G., Y.Y., A.N.M., J.F., X.Y., G.W., and H.T.Z. supervised the work; and M.E.G., J.G., H.Z., A.N.M., J.F., Y.Y., and X.Y. wrote the original draft. All authors reviewed the manuscript.

Competing interests
Authors declare that they have no competing interests.