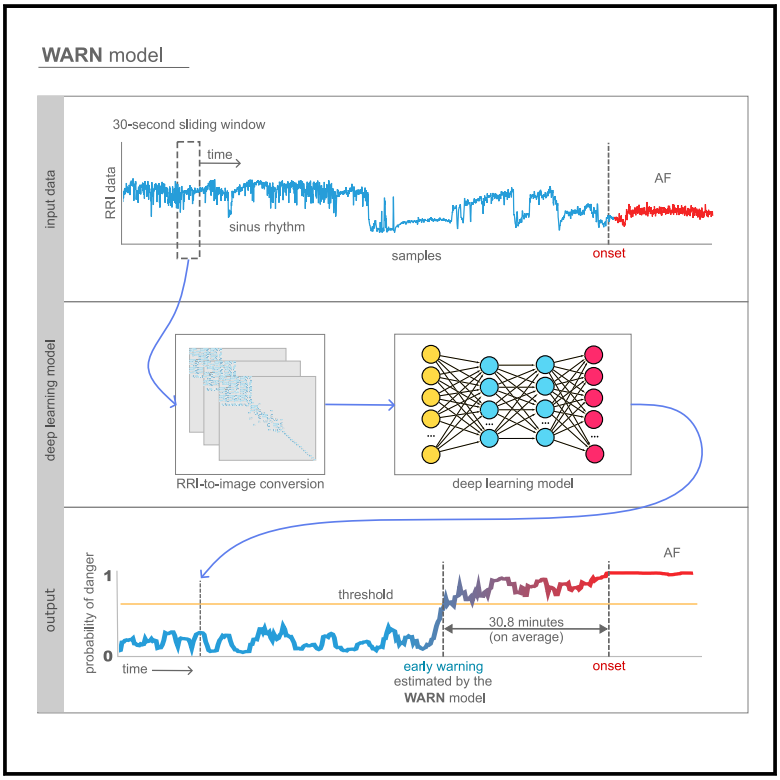
Article

Patterns

Early warning of atrial fibrillation using deep learning

Graphical abstract



Authors

Marino Gavidia, Hongling Zhu, Arthur N. Montanari, ..., Xin He, Xiaoyun Yang, Jorge Goncalves

Correspondence

yangxiaoyun321@126.com (X.Y.), jmg77@cam.ac.uk (J.G.)

In brief

Unlocking the potential of wearable technology for cardiac health, this paper presents a deep learning model that can predict atrial fibrillation onset on average over 30 min in advance with high accuracy. Leveraging everyday wearable data, this work paves the way for a new era in proactive heart rhythm monitoring and reducing emergency interventions, offering a glimpse into the future of personalized and preemptive healthcare strategies.

Highlights

- Predicts AF onset on average 30.8 min in advance on test data
- Achieves 83% accuracy and 85% F1 score on test data
- Uses R-to-R interval signals for monitoring, accessible via smartwatches
- Has the potential to lower interventions and costs by early AF detection

Gavidia et al., 2024, Patterns 5, 100970
June 14, 2024
https://doi.org/10.1016/j.patter.2024.100970

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Early warning of atrial fibrillation using deep learning

Marino Gavidia,[1,15] Hongling Zhu,[2,15] Arthur N. Montanari,[3] Jesús Fuentes,[1] Cheng Cheng,[4] Sergio Dubner,[5] Martin Chames,[6] Pierre Maison-Blanche,[7] Md Moklesur Rahman,[8] Roberto Sassi,[8] Fabio Badilini,[9] Yinuo Jiang,[4]

(Author list continued on next page)

[1]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 4367 Belvaux, Luxembourg
[2]Division of Cardiology, Department of Internal Medicine, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China
[3]Department of Physics and Astronomy, Northwestern University, Evanston, IL 60208, USA
[4]School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China
[5]Clinica y Maternidad Suizo Argentina, Buenos Aires 1461, Argentina
[6]Centro Integral Cardiovascular, Gualeguaychú, Entre Ríos, Argentina
[7]Department of Cardiology, Hôpital Bichat, 75018 Paris, France
[8]Computer Science Department, University of Milan, 20133 Milan, Italy
[9]Department of Physiologic Nursing, University of California, San Francisco, San Francisco, CA 94143, USA

(Affiliations continued on next page)

THE BIGGER PICTURE Atrial fibrillation (AF), the most prevalent heart rhythm disorder, affects millions globally, leading to significant increases in stroke risk, heart failure, and healthcare expenses. These challenges underscore the need for innovative monitoring solutions. Wearable technology, coupled with artificial intelligence, will eventually enable continuous, real-time tracking of heart health and warn users of imminent danger. This paper shows that such a future is not far. Our research introduces a model, WARN, that harnesses R-to-R intervals, the intervals between successive heartbeats, from readily available smartwatches to issue early warnings of AF onset. By leveraging extensive long-term data of individual patients, we expect that WARN can be personalized to significantly improve the prediction horizon, offering a future where many patients might manage AF proactively with as-needed medication rather than routine daily doses, thereby optimizing treatment regimens and improving quality of life.

SUMMARY

Atrial fibrillation (AF), the most prevalent cardiac rhythm disorder, significantly increases hospitalization and health risks. Reverting from AF to sinus rhythm (SR) often requires intensive interventions. This study presents a deep-learning model capable of predicting the transition from SR to AF on average 30.8 min before the onset appears, with an accuracy of 83% and an F1 score of 85% on the test data. This performance was obtained from R-to-R interval signals, which can be accessible from wearable technology. Our model, entitled Warning of Atrial Fibrillation (WARN), consists of a deep convolutional neural network trained and validated on 24-h Holter electrocardiogram data from 280 patients, with 70 additional patients used for testing and further evaluation on 33 patients from two external centers. The low computational cost of WARN makes it ideal for integration into wearable technology, allowing for continuous heart monitoring and early AF detection, which can potentially reduce emergency interventions and improve patient outcomes.

INTRODUCTION

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,[2] healthcare costs continue to increase mainly because of hospitalization and treatment costs.[3] AF episodes contribute to emergency department presentations due to high

Shengjun Zhang,[4] Hai-Tao Zhang,[4] Hao Du,[10] Basi Teng,[11] Ye Yuan,[4] Guohua Wan,[12] Zhouping Tang,[13] Xin He,[14] Xiaoyun Yang,[2,*] and Jorge Goncalves[1,11,16,*]

[10]Department of Ophthalmology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China
[11]Department of Plant Sciences, Cambridge University, CB2 3EA Cambridge, UK
[12]Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai 200052, China
[13]Department of Neurology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China
[14]School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China
[15]These authors contributed equally
[16]Lead contact
*Correspondence: yangxiaoyun321@126.com (X.Y.), jmg77@cam.ac.uk (J.G.)
https://doi.org/10.1016/j.patter.2024.100970

symptom burden and heart failure decompensation from tachycardiomyopathy. Maintaining sinus rhythm (SR) is a priority since AF events can increase the risk of other diseases,[4] such as stroke and dementia,[5,6] as well as lead to atrial remodeling, which may enhance susceptibility to future episodes.[7] The early prediction of AF episodes in patients with paroxysmal AF could prompt patients to take preventive measures to maintain SR (e.g., avoid alcohol consumption or take preventive antiarrhythmic and anticoagulation medication), possibly reducing emergency department presentations and associated healthcare costs. However, the identification of patients with a high likelihood of AF onset and its early warning prediction (on the timescale of minutes or hours) are challenging problems in the clinical setting.[8,9] To overcome these challenges, we develop a deep learning model that continuously monitors patients to provide early warnings of imminent AF onsets.

The automated *detection* of AF regimes from recorded electrocardiogram (ECG) data is a well-studied problem in the literature.[10,11] Recent approaches, based on machine learning and neural networks, have achieved over 99% accuracy in the classification task,[12–15] which led to functional applications on wearable devices of Apple, Fitbit, Samsung, and others using photoplethysmography.[16–19] On the other hand, the *prediction* of the onset of AF is still an open problem.[20] Numerous studies have developed models for long-term risk assessments of AF and other cardiovascular diseases, providing estimates typically on the order of weeks, months, or years.[21–25] Such machine learning models for AF detection and risk assessment are often trained on short-duration ECG samples obtained from sporadic cardiologist controls. Although these datasets are very extensive (thousands or even millions of recordings), they do not typically contain long-duration ECG recordings (on the order of hours)—a type of data required for the development of models for real-time monitoring and prediction. Long-duration recordings require the inconvenient use of Holter devices or patches and are thus often collected from patients with more severe AF conditions. These factors substantially reduce the amount of data available for model training in forecasting problems for cardiovascular diseases.

Despite these data challenges, recent advances based on machine learning and deep learning models have been proposed for short-term prediction of AF using models trained on features extracted from ECG leads,[26–31] R-to-R intervals (RRIs),[32–36] or a combination of both.[37–44] All of these methods have strong limitations. Most have limited data for model training—typically around 50 or fewer patients. More importantly, all methods use data up to, or very close to, the onset of AF to "predict" an AF event. Since little or no warning of AF onset is given in advance, these methods can be categorized as detection/classification tools rather than early warning models. As an example, one study[35] uses the entire window from 4 h to 0 h (i.e., at onset) to classify whether an AF event will follow or not; in practice, it does not provide an early warning. In contrast, our work departs from this approach to a more prospective prediction model. Utilizing a sliding window feature, our model is designed to identify precursors of AF that are far away from the onset, thus providing early warnings on a timescale of minutes. Figure 1A illustrates the distinct data windows used by different models in detection, prediction at onset, and early warning prediction.

This paper presents a retrospective study that develops a deep learning model for early warning of AF, entitled WARN (Warning of Atrial Fibrillation). The model is trained and tested on 350 individual 24-h Holter recordings. On the test data, WARN gives early warnings, on average, 30.8 min before onset of AF, with an accuracy and F1 score of 83% and 85%, respectively. Our model has a high performance using only RRI signals, which can be acquired from easy-to-wear and affordable pulse signal recorders, such as smartwatches or smart fitness bands. These devices can be used on a daily basis by patients, paving the way for real-time monitoring algorithms that learn and monitor long-term cardiac dynamics.

RESULTS

We used 24-h Holter recordings collected from 350 patients at Tongji Hospital (Wuhan, China) to develop and evaluate WARN for the early warning of AF. Recordings with short-duration AF episodes and/or significant noise artifacts were excluded from the original dataset. The cohort was divided in two groups for the training/cross-validation (280 patients) and testing (70 patients) of WARN, as summarized in Table 1 with variables expressed as mean or interquartile range (IQR). The ECG data had a sampling frequency of 128 Hz and a resolution of 10 bits. We used only the lead II recordings and converted them to RRI data, motivated by our ultimate goal to develop monitoring/prediction methods on easy-to-wear wearable devices. First, the ECG data were pre-processed with a 0.5–40-Hz band-pass filter to reduce noise, followed by the identification of R peaks using the Pan-Tompkins algorithm.[45,46] The data were labeled with three classes: AF, pre-AF, and SR. AF was labeled by cardiologists, while pre-AF was defined as the period just before AF onset with high RRI variability (Figure S2). The remaining segments were labeled as SR. Then, the RRI data were

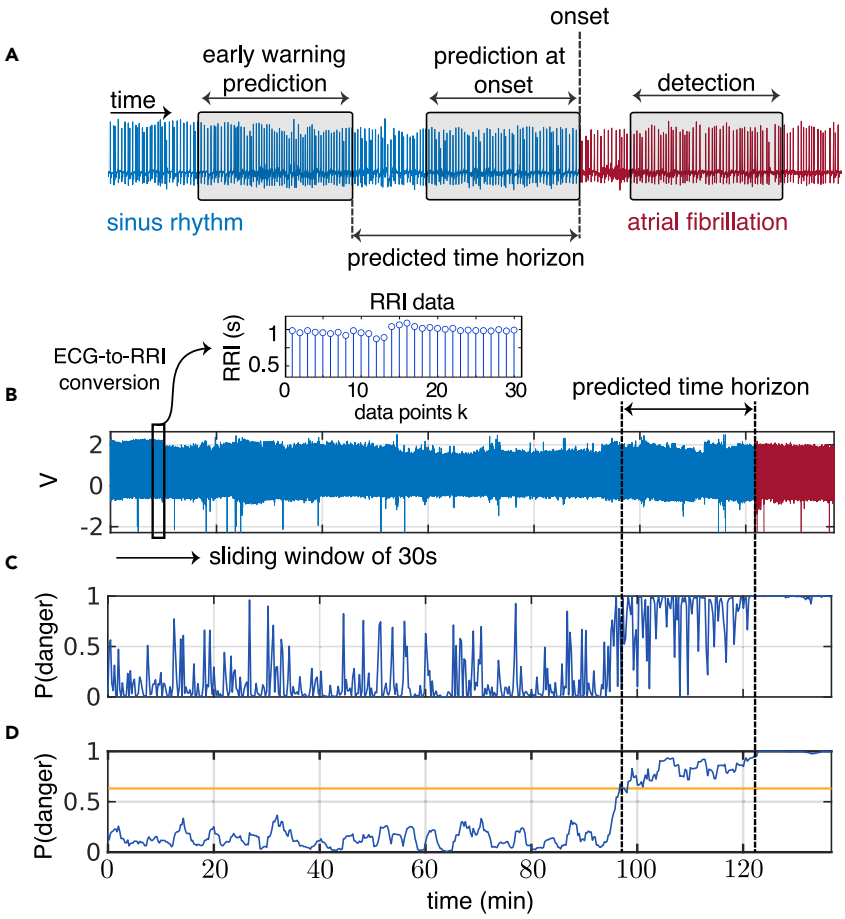


Figure 1. Detection versus prediction and early warnings

(A) Diagram of the types of data windows used for early warning AF prediction (left box), AF prediction at onset (middle), and AF detection (right). The predicted time horizon is the time between the early warning and AF onset.

(B) The model employs a sliding window, which sequentially samples the ECG data every 15 s. This window contains a duration of 30 s of ECG data, which are then converted to RRI data. Specifically, the total duration of the n RRI samples within the window is precisely 30 s.

(C) For each sliding window of RRI data, the model computes the "probability of danger" to transition to AF.

(D) Then, a non-anticipative moving average window (considering only past values) smooths this probability of danger. When it crosses a threshold (yellow line), it triggers a warning of an imminent AF onset. For this particular patient, an alert is triggered 24 min in advance of AF onset. The supplemental videos illustrate the probability of danger rising as the AF onset approaches for other patients.

used to train a deep convolutional neural network (CNN) with 479 layers that classifies between SR, pre-AF, and AF segments. Finally, for each sliding window of 30 s of RRI data, WARN outputs a "probability of danger" that the patient will have an imminent AF episode. This probability is defined as $P(\text{danger}) = P(\text{pre-AF}) + P(\text{AF})$, which represents the probability that a sliding window is either in the pre-AF or AF class. This is then repeated every 15 s (Figures 1B–1D). Further dataset descriptions and method developments are described in the supplemental information.

We initially evaluated WARN on a test dataset consisting of RRI data. For comparison purposes, WARN was also trained and tested on ECG data, achieving slightly better performance. Finally, to evaluate WARN's performance on out-of-sample data, we tested the model on 33 patients from external datasets in healthcare centers in Argentina and France. The performance results of WARN for the prediction of AF episodes are summarized in Table 2. Next, we analyze in more detail the performance of WARN in each of these datasets.

RRI data

Figure 2 presents the results for the RRI test dataset. The threshold of 0.57 was selected from the validation data as the optimal value that balances the tradeoff between accuracy and predicted time horizon (see supplemental information, Section S6). Figure 2 also includes results for two other thresholds to contrast tradeoffs between sensitivity and specificity. Figure 2B shows that, for the threshold of 0.57, AF onset is predicted at least 30 min in advance for around 60% of all patients in the test cohort while attaining relatively high performance metrics (Figure 2C). Overall, the performance of WARN is balanced between the "danger" and "normal" classes, attaining high area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) scores (Figure 2D). Smaller thresholds tend to increase the average predicted time horizon for an AF onset across patients at the expense of smaller accuracy and a larger number of false positives (Figure 2E). For patients at risk, the threshold could be smaller and more sensitive to reduce false negatives (Figure 2F). Finally, WARN seems to achieve similar performance for patients in different age groups; Figure 2A shows that AF in younger patients (less than 65 years old) can be predicted around 3 min earlier than for older patients.

We analyzed all false predictions with the 0.57 threshold, negative and positive, to gain insight into the algorithm. Table S3 lists the observations on the incorrectly classified patients by WARN. 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. Of the 22 false positives, 13 had premature atrial contractions (PACs), 5 had premature ventricular contractions, 6 had unstable baselines, 4 had sinus tachycardia, and one had atrial flutter. We noticed that the majority of these records (15 of 22) were very noisy, stressing the necessity of treating patients' skin with saline or disinfectant before wearing ECG devices to ensure the electrodes are well connected to the skin. Besides the noise influence, we speculate that some of these false positive events

Table 1. Characteristic of the patients

Characteristic	Tongji Hospital, China		External centers	
	Training cohort	Test cohort	France	Argentina
Total	280	70	25	8
Age < 65	115 (mean age 55)	31 (mean age 54)	8 (mean age 57)	2 (mean age 60)
Age ≥ 65	165 (mean age 73)	39 (mean age 73)	17 (mean age 73)	6 (mean age 80)
Male	163	26	15	5
Female	117	44	10	3
Age				
Mean	67	66	65	75
Median	67	67	64	77
Range	[20, 93]	[21, 92]	[49, 86]	[55, 91]
IQR	14	15	13	14

Table 2. Summary of key performance metrics for WARN

Dataset	Accuracy (%)	AUROC	AUPRC	Predicted time horizon	
				Mean	Median
Test (RRI)	82.7	0.90	0.88	30.8 min	38.0 min
Test (ECG)	82.4	0.95	0.96	32.5 min	43.4 min
External centers	75.0	0.80	0.73	31.8 min	41.3 min

correspond to moments where the heart was close to switching 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 toward the tipping point that leads to a dynamic transition from SR to AF. It is possible that, in some of the false positives, the heart was close to switching to AF, but it did not—especially for those patients with PACs (13 of 22), which are well-known precursors of AF and highly related or causal to the occurrence of AF.[47]

ECG data

We investigate whether there is a substantial gain in performance when the CNN is trained on the original ECG data instead. Figure S6 summarizes the performance of WARN on the test ECG data. Overall, with ECG data, there was a slight improvement of model performance compared to using RRI data; the AUROC and AUPRC increased, respectively, by 5.5% and 9.1%, albeit the accuracy and the mean predicted time horizons were relatively similar (Table 2). Given the loss of information present in the conversion of ECG to RRI data, it is not surprising that the model trained on ECG data has improved performance compared to the model trained on RRI data. What is surprising is that the improvement was relatively small, considering the richer, highly sampled, continuous-time nature of ECG data compared to the simpler, low-sampled, discrete-time RRI data. The achieved results show that prediction of AF onset can be efficiently performed using only RRI data, which opens possibilities for future development of real-time monitoring and early warnings from comfortable wearable devices.

RRI data from external centers

To further validate the performance of WARN on independent test datasets covering other demographics, we used ECG data collected from patients with AF from healthcare centers in Argentina (8 patients) and France (25 patients). The ECG data were first converted to RRI data. Then, using the same hyperparameters used in the WARN testing set, we obtained an accuracy of 75% and a mean (median) predicted time horizon until AF onset of 31.8 min (41.31 min). Figure S7 summarizes the model performance for this external-center dataset. The performance of WARN applied to this external dataset remains relatively high (close to the accuracy and mean predicted time obtained using the trained datasets), which demonstrates the potential of our method to generalize to "out-of-sample" data.

We also tested WARN's performance on the open-access data (Paroxysmal AF Prediction Challenge Database, AFPDB) from Physionet,[48,49] as summarized in Figure S8. The results are shown for a balanced set of 20 AF patients and 20 healthy patients. Note that WARN was not designed for healthy patients in general since the training/test data consisted only of recordings collected from patients already diagnosed with AF. This may explain the slightly worst results of WARN on the Physionet data (accuracy of 0.7, AUROC = 0.76 and AUPRC = 0.79). Furthermore, the Physionet dataset consists only of ECG records of 30 min, which led to shorter predicted time horizons for AF (mean of 12.9 min). Finally, it should be noted that previous models using Physionet were not always reproducible.[50]

DISCUSSION

This paper developed WARN, an automated prediction method for early warning of AF onset based on deep CNN and RRI signals. The method takes 30-s RRI samples every 15 s and computes the probability of danger of imminent AF onset. 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 the specified threshold. On the test data (70 patients) and two external validation sets (33 patients), WARN predicted AF onset on average 31 min and 33 min in advance with an accuracy of 83% and 73%, respectively.

Table 3 compares the performance of WARN to previous work on AF prediction. Among the listed references, WARN is the first method to provide an early warning of AF far from onset. The prediction horizon for all previous studies is near or at AF onset. Notably, the study[36] stands out as the only method that predicted AF prior to its onset, with a prediction time horizon of 30 s. It trained a CNN for binary classification from 5 min of RRI data with an accuracy of 66%. Among the other methods with predictions at AF onset, the highest accuracy of 98% was reported[39] by combining multilayer perceptron, K-nearest neighbor, and Support Vector Machine (SVM). Except for Guo et al.[35] all other studies relied on window lengths between 2 and 30 min. Table 3 also highlights that the dataset collected for our model training is the second larger dataset across all methods. See Section S7 in the supplemental information for further discussion.

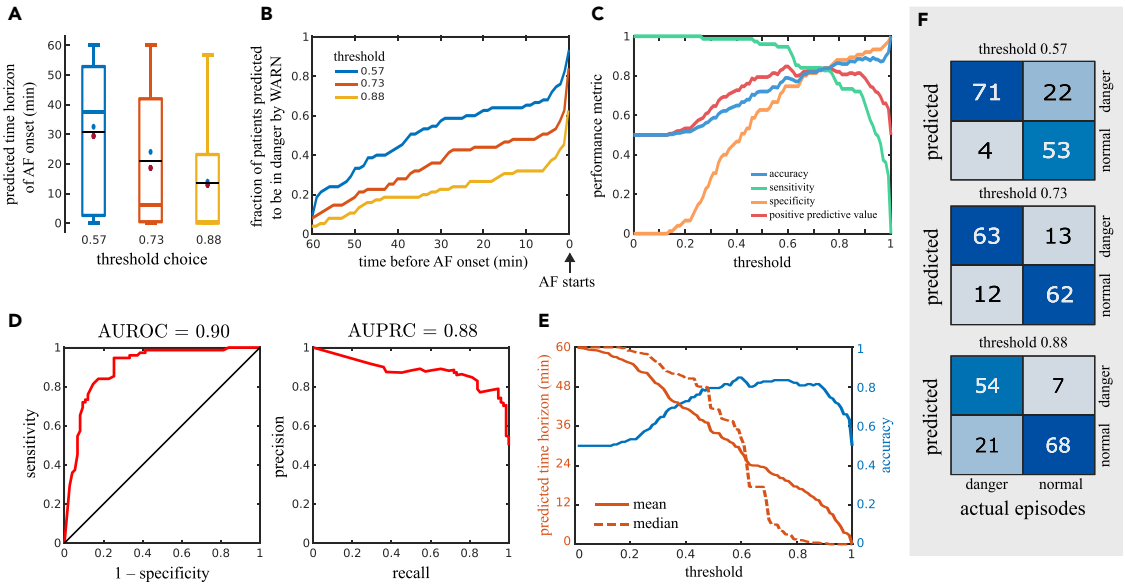


Figure 2. Performance of WARN on the RRI test dataset

We chose three probability thresholds (0.57, 0.73, and 0.88) to contrast tradeoffs between sensitivity and specificity.

(A) Boxplots of the predicted time horizon (first early warning until AF onset) for different probability thresholds across all patients. Median and mean values are marked by colored and black lines, respectively. Blue circles and red asterisks represent the means of patients younger and older than 65 years, respectively.

(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. The curves cross at a threshold of 0.74 (with a value of 83.6%).

(D) Receiver operating characteristic curve (ROC) (left) and precision-recall curve (PRC) (right).

(E) Trade-offs between the predicted time horizon and model accuracy as a function of the probability threshold.

(F) Confusion matrices for different thresholds computed on 75 episodes of AF of the 70 patients in the test set.

WARN 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 (Figure S5). Our choice in this paper was based on a simple tradeoff decision to keep the accuracy, F1 score, and prediction horizon all relatively high. Different situations may require a higher weight on one of these objectives. For example, smaller thresholds yield more sensitive models with reduced false negatives, which could be used for high-risk patients. On the other hand, higher thresholds lead to more specific models and reduced false positives,

Table 3. Performance comparison between WARN and previous work

Year	Study	Method	Patients (no.)	Window length	Prediction horizon	Accuracy	Sensitivity	Specificity
2012	Mohebbi et al.[33]	RRI, SVM	NR	30 min	onset	96	96	93
2013	Costin et al.[37]	ECG, QRS complexes	75	5 min	onset	90	89	89
2016	Boon et al.[28]	RRI, SVM	53	30 min	onset	80	81	79
2018	Li et al.[38]	ECG, Markov chain	5	2 min	onset	82	86	80
2018	Boon et al.[29]	RRI, SVM	53	5 min	onset	87	86	88
2018	Ebrahimzadeh et al.[39]	ECG, mixture of experts	53	5 min	onset	98	100	96
2021	Guo et al.[35]	RRI, XGBoost	554	1–4 h	onset	88	82	96
2021	Tzoul et al.[51]	ECG, CNN	8	5 min	onset	89	88	89
2022	Grégoire et al.[36]	RRI, CNN	140	300 RRI (~5 min)	30 beats (~30 s)	66	80	53
2023	WARN	RRI, CNN	350	30 s	30.8 min before onset	83	95	70
2023		ECG, CNN	350	30 s	32.5 min before onset	82	95	69

NR, not reported; ECG, electrocardiogram; RRI, R-to-R intervals; CNN, convolutional neural network; SVM, support vector machine.

which may be more suitable to monitor AF patients with a lower incidence.

Compared with ECG data, results using RRI data showed a slight reduction in performance. On the test data, both exhibited similar accuracy of approximately 83%, with average prediction horizons of 32.5 min for ECG and 30.8 min for RRI data. This slight decrease in performance is compensated by the ease of continuously obtaining RRI data from easily accessible and cost-effective wearable devices like smartwatches, making them ideal for long-term monitoring. Hence, ECG signals are not really needed, as the results would be similar to those acquired from smartwatches. On a standard laptop computer, the total computational time spent on each sliding window was around 100 ms. This is considerably less than 15 s, which is the time until the next window, making it feasible to implement WARN in smartphones to process the streamed data from a smartwatch in real time.[52] For instance, the deep learning model used in this paper, EfficientNetV2, can be adapted to mobile devices through the TensorFlow Lite framework (see Section S5 in the supplemental information). Since smartwatches can be worn for long-term monitoring and record RRI signals, the early warning provided by WARN could potentially provide sufficient time for patients to take oral antiarrhythmic drugs on demand to prevent the onset of AF or other targeted therapies or lifestyle interventions. Moreover, models could be retrained offline (e.g., once a day using high-performance computing) as new data become available.

Limitations of the study

WARN was trained on 24-h RRI data from 280 patients. Hence, it is an "average" algorithm among those patients. With much longer time horizons on single patients, WARN could be personalized to improve its performance and be converted into a real-time prediction algorithm that updates itself with newer incoming data. With improved performance and an even earlier warning of AF onset, some patients could benefit in the future from taking antiarrhythmic medication on demand (when patients receive warnings) instead of the current approach of taking medication daily. For example, patients with sporadic AF events could benefit from this new therapy. Hence, future work should focus on personalized models by including considerably more data for a single patient (compared to the 24-h datasets used in this paper) to achieve even earlier warnings of AF onset. Eventually, this approach could lead to new clinical trials and changes in therapies. Finally, WARN was trained on 100% Chinese patients. Although it was also tested on patients from France and Argentina with a good performance, the method can potentially be further improved if trained on specific demographics and commodities.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Requests for further information, resources, and reagents should be directed to and will be fulfilled by the lead contact, Jorge Goncalves. Training and validation datasets from Tongji Hospital will be accessible pending approval by Xiaoyun Yang.

Materials availability

This study did not generate new unique materials.

Data and code availability

The data were provided by Tongji Hospital from China, the Clínica y Maternidad Suizo Argentina, and the Groupe Hospitalier Privé Ambroise Paré-Hartmann (acquired between 2014 and 2022). To protect patients' privacy, the data were 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 were obtained according to the principles of the declaration of Helsinki.

The testing data are publicly available at Zenodo (https://doi.org/10.5281/zenodo.10815810).[53] The Physionet data used as part of the external validation of this study are available from the open-access paroxysmal AF Prediction Challenge Database[48,49] (https://physionet.org/content/afpdb/1.0.0/).

Data preprocessing and segmentation were implemented using MATLAB software. The neural network was implemented on the Keras Framework with the Tensorflow backend on Python 3.7. Codes have been deposited on Zenodo (https://doi.org/10.5281/zenodo.10815367).[54]

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2024.100970.

ACKNOWLEDGMENTS

The authors are thankful for support from the Luxembourg National Research Fund (grant PRIDE15/10907093/CriTiCS) and the National Natural Science Foundation of China (grants 92167201 and 82100531).

AUTHOR CONTRIBUTIONS

J.G. conceptualized the research. H.Z., S.D., P.M.-B., M.M.R., and M.C. collected, curated, and annotated the data. M.G., J.G., A.N.M., Y.J., S.Z., and B.T. developed the AI model. M.G. implemented the code and validated the results. M.G., H.Z., J.G., A.N.M., J.F., Y.Y., R.S., F.B., C.C., and X.H. analyzed the results. J.G., Y.Y., A.N.M., M.M.R., Z.T., H.D., Z.T., H.B., J.F., X.Y., G.W., and H.-T.Z. supervised the work. M.G., A.N.M., J.G., H.Z., J.F., Y.Y., and X.Y. wrote the original draft. All authors revised the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 14, 2023
Revised: February 21, 2024
Accepted: March 25, 2024
Published: April 18, 2024

REFERENCES

1. Chugh, S.S., Havmoeller, R., Narayanan, K., Singh, D., Rienstra, M., Benjamin, E.J., Gillum, R.F., Kim, Y.-H., McAnulty, J.H., Jr., Zheng, Z.-J., et al. (2014). Worldwide epidemiology of atrial fibrillation: a global burden of disease 2010 study. Circulation *129*, 837–847.
2. January, C.T., Wann, L.S., Alpert, J.S., Calkins, H., Cigarroa, J.E., Cleveland, J.C., Conti, J.B., Ellinor, P.T., Ezekowitz, M.D., Field, M.E., et al. (2014). 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. J. Am. Coll. Cardiol. *64*, e1–e76.
3. Benjamin, E.J., Muntner, P., Alonso, A., Bittencourt, M.S., Callaway, C.W., Carson, A.P., Chamberlain, A.M., Chang, A.R., Cheng, S., Das, S.R., et al. (2019). Heart disease and stroke statistics 2019 update: a report from the American Heart Association. Circulation *139*, e56–e528.

4. Kornej, J., Börschel, C.S., Benjamin, E.J., and Schnabel, R.B. (2020). Epidemiology of atrial fibrillation in the 21st century: novel methods and new insights. Circ. Res. *127*, 4–20.
5. de Leeuw, F.-E., de Groot, J.C., Oudkerk, M., Kors, J.A., Hofman, A., van Gijn, J., and Breteler, M.M. (2000). Atrial fibrillation and the risk of cerebral white matter lesions. Neurology *54*, 1795–1801.
6. Aldrugh, S., Sardana, M., Henninger, N., Saczynski, J.S., and McManus, D.D. (2017). Atrial fibrillation, cognition and dementia: A review. J. Cardiovasc. Electrophysiol. *28*, 958–965.
7. Prystowsky, E.N. (2000). Management of atrial fibrillation: therapeutic options and clinical decisions. Am. J. Cardiol. *85*, 3–11.
8. Wilson, R.E., Rush, K.L., Hatt, L., Reid, R.C., and Laberge, C.G. (2020). The symptom experience of patients with atrial fibrillation before their initial diagnosis. J. Cardiovasc. Nurs. *35*, 347–357.
9. Im, S., Kim, D., and Kim, B. (2017). P3624 clinical and electrocardiographic characteristics for prediction of new-onset atrial fibrillation in asymptomatic patients with atrial premature complexes. Eur. Heart J. *38*, ehx504.P3624.
10. Rizwan, A., Zoha, A., Mabrouk, I.B., Sabbour, H.M., Al-Sumaiti, A.S., Alomainy, A., Imran, M.A., and Abbasi, Q.H. (2021). A review on the state of the art in atrial fibrillation detection enabled by machine learning. IEEE Rev. Biomed. Eng. *14*, 219–239.
11. Panindre, P., Gandhi, V., and Kumar, S. (2020). Comparison of performance of artificial intelligence algorithms for real-time atrial fibrillation detection using instantaneous heart rate. In IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI, pp. 168–172.
12. Martis, R.J., Acharya, U., Prasad, H., Chua, C.K., and Lim, C.M. (2013). Automated detection of atrial fibrillation using bayesian paradigm. Knowl.-Based Syst. *54*, 269–275.
13. Annavarapu, A., and Kora, P. (2016). Ecg-based atrial fibrillation detection using different orderings of conjugate symmetric–complex Hadamard transform. Int. J. Cardiovasc. Acad. *2*, 151–154.
14. Ribeiro, A.H., Ribeiro, M.H., Paixão, G.M.M., Oliveira, D.M., Gomes, P.R., Canazart, J.A., Ferreira, M.P.S., Andersson, C.R., Macfarlane, P.W., Meira, W., Jr., et al. (2020). Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat. Commun. *11*, 1760.
15. Zhu, H., Cheng, C., Yin, H., Li, X., Zuo, P., Ding, J., Lin, F., Wang, J., Zhou, B., Li, Y., et al. (2020). Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study. Lancet. Digit. Health *2*, e348–e357.
16. Seshadri, D.R., Bittel, B., Browsky, D., Houghtaling, P., Drummond, C.K., Desai, M.Y., and Gillinov, A.M. (2020). Accuracy of apple watch for detection of atrial fibrillation. Circulation *141*, 702–703.
17. Lubitz, S.A., Faranesh, A.Z., Atlas, S.J., McManus, D.D., Singer, D.E., Pagoto, S., Pantelopoulos, A., and Foulkes, A.S. (2021). Rationale and design of a large population study to validate software for the assessment of atrial fibrillation from data acquired by a consumer tracker or smartwatch: The Fitbit heart study. Am. Heart J. *238*, 16–26.
18. Dörr, M., Nohturfft, V., Brasier, N., Bosshard, E., Djurdjevic, A., Gross, S., Raichle, C.J., Rhinisperger, M., Stöckli, R., and Eckstein, J. (2019). The WATCH AF trial: SmartWATCHes for detection of atrial fibrillation. JACC. Clin. Electrophysiol. *5*, 199–208.
19. Wang, Y.-C., Xu, X., Hajra, A., Apple, S., Kharawala, A., Duarte, G., Liaqat, W., Fu, Y., Li, W., Chen, Y., and Faillace, R.T. (2022). Current advancement in diagnosing atrial fibrillation by utilizing wearable devices and artificial intelligence: A review study. Diagnostics *12*, 689.
20. Aljanabi, M., Qutqut, H., and Hijjawi, M. (2018). Machine learning classification techniques for heart disease prediction: A review. Int. J. Eng. Technol. *7*, 5373–5379.
21. Attia, Z.I., Noseworthy, P.A., Lopez-Jimenez, F., Asirvatham, S.J., Deshmukh, A.J., Gersh, B.J., Carter, R.E., Yao, X., Rabinstein, A.A., Erickson, B.J., et al. (2019). An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: A retrospective analysis of outcome prediction. Lancet *394*, 861–867.
22. Biton, S., Gendelman, S., Ribeiro, A.H., Miana, G., Moreira, C., Ribeiro, A.L.P., and Behar, J.A. (2021). Atrial fibrillation risk prediction from the 12-lead electrocardiogram using digital biomarkers and deep representation learning. Eur. Heart J. Digit. Health *2*, 576–585.
23. Raghunath, S., Pfeifer, J.M., Ulloa-Cerna, A.E., Nemani, A., Carbonati, T., Jing, L., vanMaanen, D.P., Hartzel, D.N., Ruhl, J.A., Lagerman, B.F., et al. (2021). 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.
24. Diaz-Pinto, A., Ravikumar, N., Attar, R., Suinesiaputra, A., Zhao, Y., Levelt, E., Dall'Armellina, E., Lorenzi, M., Chen, Q., Keenan, T.D.L., et al. (2022). Predicting myocardial infarction through retinal scans and minimal personal information. Nat. Mach. Intell. *4*, 55–61.
25. Singh, J.P., Fontanarava, J., de Massé, G., Carbonati, T., Li, J., Henry, C., and Fiorina, L. (2022). Short-term prediction of atrial fibrillation from ambulatory monitoring ecg using a deep neural network. Eur. Heart J. Digit. Health *3*, 208–217.
26. Blanche, C., Tran, N., Rigamonti, F., Burri, H., and Zimmermann, M. (2013). Value of p-wave signal averaging to predict atrial fibrillation recurrences after pulmonary vein isolation. Europace *15*, 198–204.
27. Alcaraz, R., Martínez, A., and Rieta, J.J. (2015). Role of the p-wave high frequency energy and duration as noninvasive cardiovascular predictors of paroxysmal atrial fibrillation. Comput. Methods Programs Biomed. *119*, 110–119.
28. Boon, K.H., Khalil-Hani, M., Malarvili, M.B., and Sia, C.W. (2016). Paroxysmal atrial fibrillation prediction method with shorter HRV sequences. Comput. Methods Programs Biomed. *134*, 187–196.
29. Boon, K.H., Khalil-Hani, M., and Malarvili, M.B. (2018). Paroxysmal atrial fibrillation prediction based on HRV analysis and non-dominated sorting genetic algorithm III. Comput. Methods Programs Biomed. *153*, 171–184.
30. Khurshid, S., Friedman, S., Reeder, C., Di Achille, P., Diamant, N., Singh, P., Harrington, L.X., Wang, X., Al-Alusi, M.A., Sarma, G., et al. (2022). ECG-based deep learning and clinical risk factors to predict atrial fibrillation. Circulation *145*, 122–133.
31. Yuan, N., Duffy, G., Dhruva, S.S., Oesterle, A., Pellegrini, C.N., Theurer, J., Vali, M., Heidenreich, P.A., Keyhani, S., and Ouyang, D. (2023). Deep learning of electrocardiograms in sinus rhythm from US veterans to predict atrial fibrillation. JAMA Cardiol. *8*, 1131–1139.
32. Dash, S., Chon, K.H., Lu, S., and Raeder, E.A. (2009). Automatic real time detection of atrial fibrillation. Ann. Biomed. Eng. *37*, 1701–1709.
33. Mohebbi, M., and Ghassemian, H. (2012). Prediction of paroxysmal atrial fibrillation based on non-linear analysis and spectrum and bispectrum features of the heart rate variability signal. Comput. Methods Programs Biomed. *105*, 40–49.
34. Lee, J., Nam, Y., McManus, D.D., and Chon, K.H. (2013). Time-varying coherence function for atrial fibrillation detection. IEEE Trans. Biomed. Eng. *60*, 2783–2793.
35. Guo, Y., Wang, H., Zhang, H., Liu, T., Li, L., Liu, L., Chen, M., Chen, Y., and Lip, G.Y.H. (2021). Photoplethysmography-based machine learning approaches for atrial fibrillation prediction: a report from the Huawei Heart Study. JACC. Asia *1*, 399–408.
36. Grégoire, J.M., Gilon, C., Carlier, S., and Bersini, H. (2022). Role of the autonomic nervous system and premature atrial contractions in short-term paroxysmal atrial fibrillation forecasting: Insights from machine learning models. Arch. Cardiovasc. Dis. *115*, 377–387.
37. Costin, H., Rotariu, C., and Păsăricã, A. (2013). Atrial fibrillation onset prediction using variability of ECG signals. In 8th International Symposium on Advanced Topics in Electrical Engineering, pp. 1–4.
38. Li, Z., Derksen, H., Gryak, J., Ghanbari, H., Gunaratne, P., and Najarian, K. (2018). A novel atrial fibrillation prediction algorithm applicable to recordings from portable devices. In 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 4034–4037.

39. Ebrahimzadeh, E., Kalantari, M., Joulani, M., Shahraki, R.S., Fayaz, F., and Ahmadi, F. (2018). Prediction of paroxysmal atrial fibrillation: a machine learning based approach using combined feature vector and mixture of expert classification on hrv signal. Comput. Methods Programs Biomed. *165*, 53–67.

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

41. Jalali, A., and Lee, M. (2020). Atrial fibrillation prediction with residual network using sensitivity and orthogonality constraints. IEEE J. Biomed. Health Inform. *24*, 407–413.

42. Hannun, A.Y., Rajpurkar, P., Haghpanahi, M., Tison, G.H., Bourn, C., Turakhia, M.P., and Ng, A.Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat. Med. *25*, 65–69.

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

44. Jo, Y.-Y., Cho, Y., Lee, S.Y., Kwon, J.-m., Kim, K.-H., Jeon, K.-H., Cho, S., Park, J., and Oh, B.-H. (2021). Explainable artificial intelligence to detect atrial fibrillation using electrocardiogram. Int. J. Cardiol. *328*, 104–110.

45. Pan, J., and Tompkins, W.J. (1985). A real-time QRS detection algorithm. IEEE Transactions on Biomedical Engineering *BME-*. IEEE Trans. Biomed. Eng. *32*, 230–236.

46. Sedghamiz, H. (2014). MATLAB implementation of Pan Tompkins ECG QRS detector. MathWorks File Exchange. https://www.mathworks.com/matlabcentral/fileexchange/45840.

47. Himmelreich, J.C.L., Lucassen, W.A.M., Heugen, M., Bossuyt, P.M.M., Tan, H.L., Harskamp, R.E., van Etten-Jamaludin, F.S., and van Weert, H.C.P.M. (2019). Frequent premature atrial contractions are associated with atrial fibrillation, brain ischaemia, and mortality: a systematic review and meta-analysis. Europace *21*, 698–707.

48. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.-K., and Stanley, H.E. (2000). Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation *101*, e215–e220.

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

50. Gilon, C., Grégoire, J.-M., Hellinckx, J., Carlier, S., and Bersini, H. (2022). Reproducibility of machine learning models for paroxysmal atrial fibrillation onset prediction. Comput. Cardiol. *49*, 1–4.

51. Tzou, H.-A., Lin, S.-F., and Chen, P.-S. (2021). Paroxysmal atrial fibrillation prediction based on morphological variant P-wave analysis with wideband ECG and deep learning. Comput. Methods Programs Biomed. *211*, 106396.

52. Luo, C., He, X., Zhan, J., Wang, L., Gao, W., and Dai, J. (2020). Comparison and benchmarking of ai models and frameworks on mobile devices. Preprint at arXiv *2*. https://doi.org/10.48550/arXiv.2005.05085.

53. Zhu, H., Gavidia, M., Montanari, A.N., Fuentes, J., Goncalves, J., and Xiaoyun, Y. (2024). Early Warning of Atrial Fibrillation Using Deep Learning (Test Dataset). https://doi.org/10.5281/zenodo.10815810.

54. Gavidia, M., Zhu, H., Montanari, A.N., Fuentes, J., Xiaoyun, Y., and Goncalves, J. (2024). WARN: Initial Release. https://doi.org/10.5281/zenodo.10815367.

Patterns, Volume *5*

Supplemental information

Early warning of atrial fibrillation using deep learning

Marino Gavidia, Hongling Zhu, Arthur N. Montanari, Jesús Fuentes, Cheng Cheng, Sergio Dubner, Martin Chames, Pierre Maison-Blanche, Md Moklesur Rahman, Roberto Sassi, Fabio Badilini, Yinuo Jiang, Shengjun Zhang, Hai-Tao Zhang, Hao Du, Basi Teng, Ye Yuan, Guohua Wan, Zhouping Tang, Xin He, Xiaoyun Yang, and Jorge Goncalves

Supplemental Information Text

S1. 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 128Hz and a resolution of 10 bits. We excluded records that did not have both SR and AF episodes. The 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 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 WARN on independent test datasets, we considered additional ECG data collected from patients with AF from three healthcare centers in different countries: the Clínica y Maternidad Suizo Argentina (53 patients of 24h ECG), the Groupe Hospitalier Privé Ambroise Paré - Hartmann (250 patients of 24h ECG) in France, and the open-access data Atrial Fibrillation Prediction Database (AFPDB) from Physionet[1,2] (75 patients of 30min ECG). Applying the same exclusion criteria described above resulted in a total of 73 patients for external validation: 8 from Argentina, 25 from France, 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). In the Physionet database, 5 AF patients presented AF with duration shorter than 1min and 2 records from healthy controls presented AF and heavy distortion due to artifact noise, which were excluded. Overall, there were 20 records to predict AF plus an equal number of 20 randomly selected records from healthy subjects for control.

S2. Method overview. WARN is a method for early warning of the onset of AF episodes. Figure S1 illustrates the method pipeline, which is subdivided in four stages. We provide an overview of WARN as follows.

1. The ECG Holter recordings are segmented into three classes: SR, pre-AF (the instances just before AF onset), and AF segments. The AF segments were labeled by cardiologists. The pre-AF segments are labeled as the ECG data preceding the AF onset, which are characterized by high RRI variability. The SR segments comprise the remaining data, which typically have lower variability[3]. The segmentation of pre-AF intervals is described in Section S3.

2. A sliding window in the ECG lead II data extracts segments of 30s that are converted to RRI data. First, 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. For each 30s window of ECG lead II data, we detect the R peaks using the Pan-Tompkins algorithm, which has an average error rate of about 1%[4,5]. The interval between heartbeats (i.e., from R to R peaks) is calculated to generate the RRI signal.

3. Each 30s window of RRI data is converted to a recurrence plot. Recurrence plots are 2-dimensional representations of the recurrent states of a time-series signal, which can be used to assess the periodicity of a signal and detect dynamical transitions (such as SR to

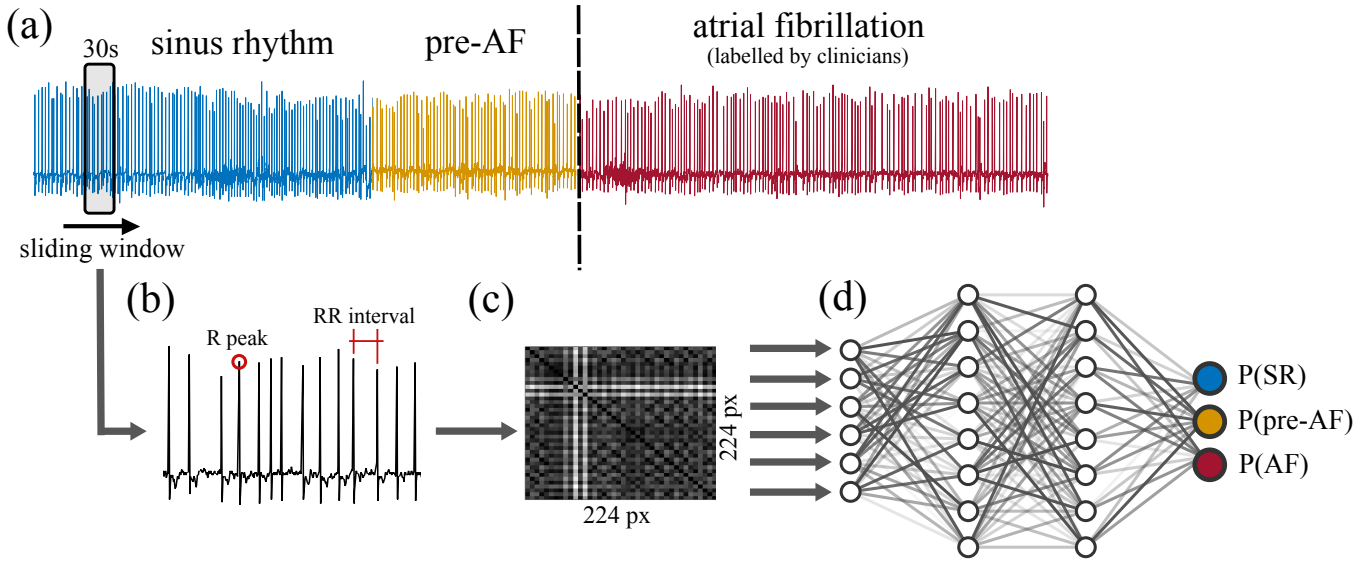


Figure S1: Pipeline of the first stage of WARN. **(a)** Each ECG record is split into three classes: SR, pre-AF, and AF. **(b)** The R peaks are automatically detected in a sliding window of ECG data, allowing the conversion of the ECG data to RRI data. **(c)** A recurrence plot is constructed using the RRI data window. **(d)** A deep CNN is trained using the recurrence plots as inputs. The model outputs are the probabilities that the sampled data belong to each of the three classes (SR, pre-AF, and AF).

AF) in a system[6–9]. The recurrence plots generate 2-dimensional images of size 224×224 pixels that are appropriate for CNN models used in image classification. The generation and interpretation of recurrence plots is described in Section S4.

4. A CNN model is trained to classify the inputs (recurrence plots) into the three classes segmented above: AF, pre-AF, and SR. We implemented the EfficientNetV2, a deep CNN with 479 layers, developed by Google in 2021[10]. The EfficientNetV2 is a modified and optimized version of EfficientNet[11], winner of the ImageNet 2019 competition[12]. 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 (satisfying $P(\text{SR}) + P(\text{pre-AF}) + P(\text{AF}) = 1$). The training of the deep CNN model is described in Section S5. Other machine-learning and deep-learning models were also trained and tested during the development of this project, but worst performance was obtained than the current pipeline (Section S5).

To provide an early warning of AF, WARN computes the probability of a patient switching to AF from the outputs of the trained CNN. We 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 either in the pre-AF or AF class. Given an entire recording of RRI data (possibly converted from ECG data), our method sequentially samples a sliding window of 30s every 15s. For each new window, the recorded 30s data are converted into a recurrence plot and fed to the CNN for classification. Figure 1c illustrates the probability of danger computed by WARN 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. 1d). A non-anticipative moving average calculates the average of a set of data points up to and including the current point and does not use any data points that occur after the current point. A binary early-warning indicator ("danger" or "normal") 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

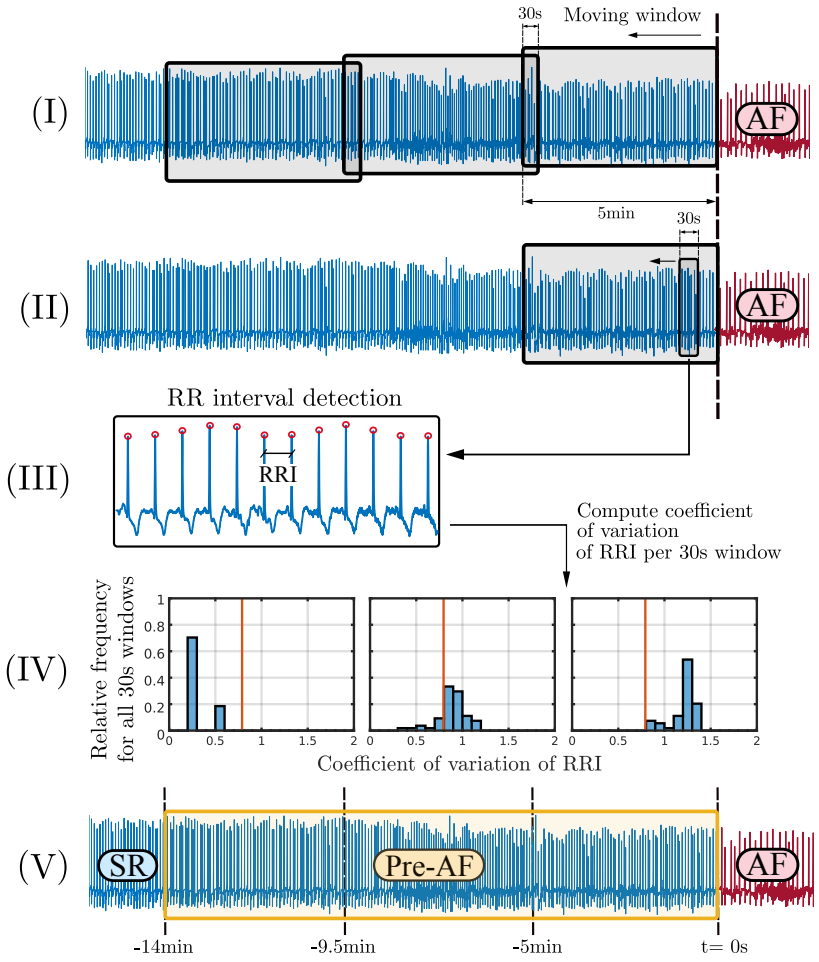


Figure S2: Pre-AF labeling process for a representative patient. **(I)** Starting from the AF onset and travelling back in time, a sliding window is generated to extract ECG samples of 5min with 30s overlapping. **(II)** For each 5min window, a second sliding window is generated, to extract samples of 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.

to maximize different performance metrics, depending on the needs of particular patients. The hyperparameter optimization is described in Section S6. 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).

S3. Pre-AF ECG segmentation. The labeling of the pre-AF ECG segments consists of five steps, illustrated in Fig. S2 for a representative patient. First, starting from the AF onset (labeled by clinicians), we select 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[4,5] to detect R waves from lead II 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. S4a). Below this threshold, the heart dynamics have low variability and can be associated with SR[13]. At this fifth and last stage, the pre-AF section is segmented, including the beginning of this last window until the onset of AF. The pre-AF segments vary in length from patient to patient[13] and may also vary within multiple onsets of AF for the same patient due to morphological and electrical changes in the heart over time[14].

S4. Recurrence plot generation. For a time-series window with N points, let the multidimensional data point $x(k) \in \mathbb{R}^n$ represent the "state" of the system at time $k = 1, \ldots, N$. The recurrence plot is a $N \times N$ matrix defined by the pairwise distance of all states along the

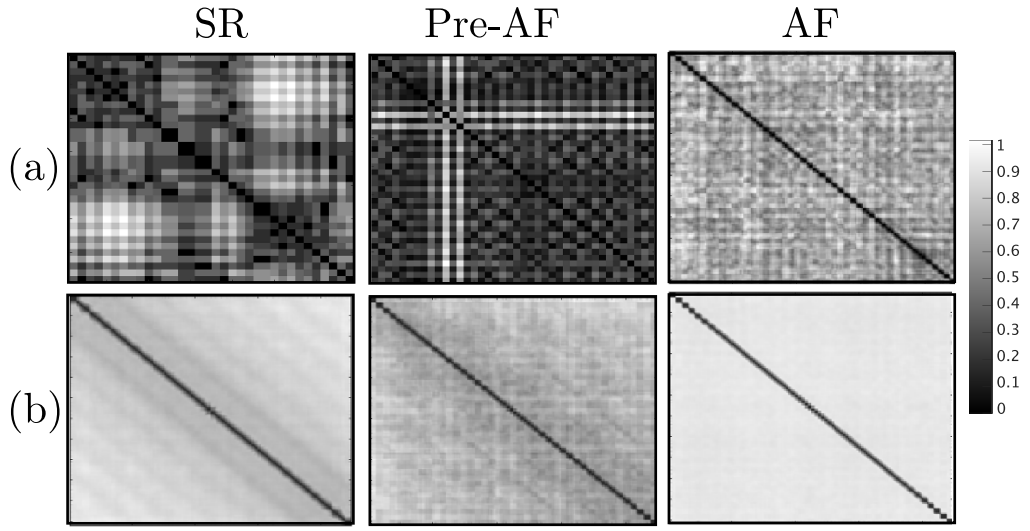


Figure S3: Recurrence plots, for each different segment (SR, pre-AF, and AF), generated from **(a)** a single 30 s sample and **(b)** an average of all recurrence plots generated for a representative patient.

recorded time-series window: $R_{ij} = \|x(i) - x(j)\|$, where $\|\cdot\|$ is the Euclidean norm[15]. For a given (i, j)-th cell in the recurrence plot, the darker the plot (i.e., smaller R_{ij}), the closer (recurrent) two states $x(i)$ and $x(j)$ are in the state space (see Fig. S3 for an example). For instance, darker parallel diagonal lines indicate periodicity in the state trajectory of a system, as R_{ij} decreases the closer two states are.

To generate the recurrence plots in this paper, for each time window of 30s, an RRI signal is computed from the recorded ECG time-series data. Let $y(k) \in \mathbb{R}^1$, for $k = 1, \ldots, N$, be the k-th data point of the RRI signal, where N is the number of data points. The recorded time series $y(k)$ is a 1-dimensional measurement of a (very likely) high-dimensional system. Following Taken's theorem, we can reconstruct an attractor that preserves the structure of the original (non-measurable) state space by employing a time-delay embedding of the recorded time series[16,17]. The embedded state vector is thus defined by $y_e(k) = [y(k)\ y(k+\tau)\ y(k+2\tau)\ \ldots\ y(k+(m-1)\tau)]^{\mathsf{T}}$, where m is the embedding dimension and τ is the parameter of delay in data points. For the RRI data the embedding parameters are $m = 2$ and $\tau = 3$, while for the ECG data we have $m = 5$ and $\tau = 6$. The selected parameters m and τ correspond, respectively, to the first local minimum of the mutual information function[18] and the smallest value such that the percentage of false nearest neighbors is below 10%[19]. Finally, the recurrence plot is computed as $R_{ij} = \|y_e(i) - y_e(j)\|$. The reader is referred to the article[6] for more details on recurrence plots and the choice of parameters.

Figure S3 shows recurrence plots corresponding to different regimes (SR, pre-AF, and AF). Comparing the SR and AF regimes, it is possible to observe how the number of states with high recurrence (small R_{ij}) decreases on average (Fig. S3b), implying that the periodicity of the system weakens. The pre-AF segments highlight a dynamical transition with an abundant number of "cross-shaped" regions of low recurrence, which are a consequence of intermittency in the RRI signal, that is, the alternation between periodic (SR) and non-periodic (AF) regimes. We hypothesize that WARN is capable of detecting subtle patterns across all three classes, enabling it to appropriately trigger an early warning for the onset of AF.

S5. Training of the deep CNN. WARN was trained and cross-validated on random samples from 280 patients. The CNN was trained using categorical cross-entropy as the loss function, using a batch size of 32 samples, ADAM as the optimizer[20], and stochastic gradient descent as the objective function optimizer[21]. To compensate for the class imbalance, the data was resampled and the loss function was weighted according to the ratio 3/1/2 for the SR, Pre-AF, and AF samples, respectively. From a total of 835,166 samples, this gave a final sample count of

417,583 (SR), 139,194 (pre-AF) and 278,388 (AF) segments. The learning rate is initialized at 10^{-4} and reduced by half after the validation loss did not improve over 5 consecutive epochs. The training is terminated after 8 consecutive epochs with no improvement.

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 for different windows length (Table S1). The best performance was obtained using a window length of 30 seconds, as also reported in another study[22]. 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 WARN to properly generalize the data. On the other hand, a smaller window length may lead to information loss[23].

Table S1: Optimal length of the sampling window.

Length (seconds)	Samples ($\times 10^6$)	Accuracy
10	2.3	0.70
30	0.8	0.74
60	0.4	0.72
120	0.2	0.69
300	0.1	0.66

After fixing the sampling window length to 30s, we compared the performance of WARN with two other network benchmarks: 1-D CNN and LSTM, commonly used on arrhythmia detection and prediction[24,25]. Using 2-D recurrence plot images as input yielded better performance than using 1-D RRI time series as input (Table S2). The proposed WARN 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. To assess performance during the second stage of WARN, the data was evenly divided between positive and negative classes, maintaining class balance and allowing fair comparisons within the confusion matrix.

Table S2: 10-fold cross-validation accuracy after training WARN and two benchmark networks.

Model	1	2	3	4	5	6	7	8	9	10	Average
EfficientNetV2	0.78	0.71	0.73	0.73	0.73	0.71	0.75	0.75	0.76	0.71	0.74
LSTM	0.64	0.61	0.63	0.61	0.61	0.58	0.60	0.63	0.57	0.60	0.60
1D-CNN	0.71	0.68	0.67	0.71	0.69	0.67	0.68	0.67	0.70	0.70	0.69

S6. Hyperparameters optimization. To optimize the hyperparameters (moving average window length and probability threshold), the performance of WARN was evaluated on a time series of 60min of sequential data before AF onset and also far from AF (that is, in SR), selected patient-wise. We selected this time-series length since more than 70% of all pre-AF segments have a duration shorter than one hour (Fig. S4). 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. S5a). For the validation

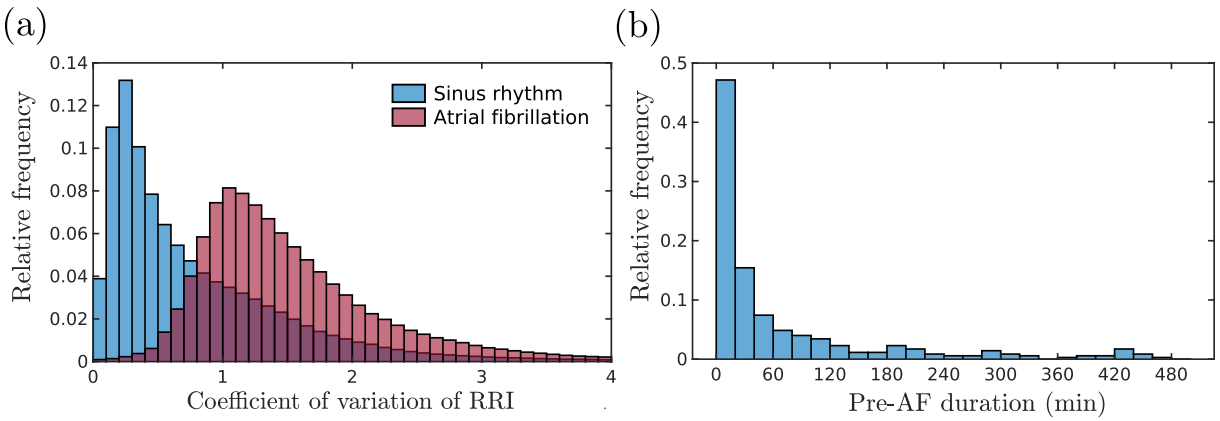


Figure S4: (a) Distributions of the coefficient of variation of the RRI for all patients from the training set, split by SR and AF regimes (as labelled by clinicians). **(b)** Distribution of Pre-AF length of for patients.

data, it is not possible to simultaneously maximize all performance measures as expected. For example, Figs. S5a-c show that the maximum predicted time horizon until AF onset (that is, the instant of the first early warning until AF onset) is achieved at low thresholds (Fig. S5b,c). However, the accuracy is very low for those values (bottom of Figs. S5a). 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 hyperparameters 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 hyperparameters satisfying this criterion. We selected the smallest moving average window, since this leads to lower computational and memory usage in smart devices, and, among those, the one that maximizes

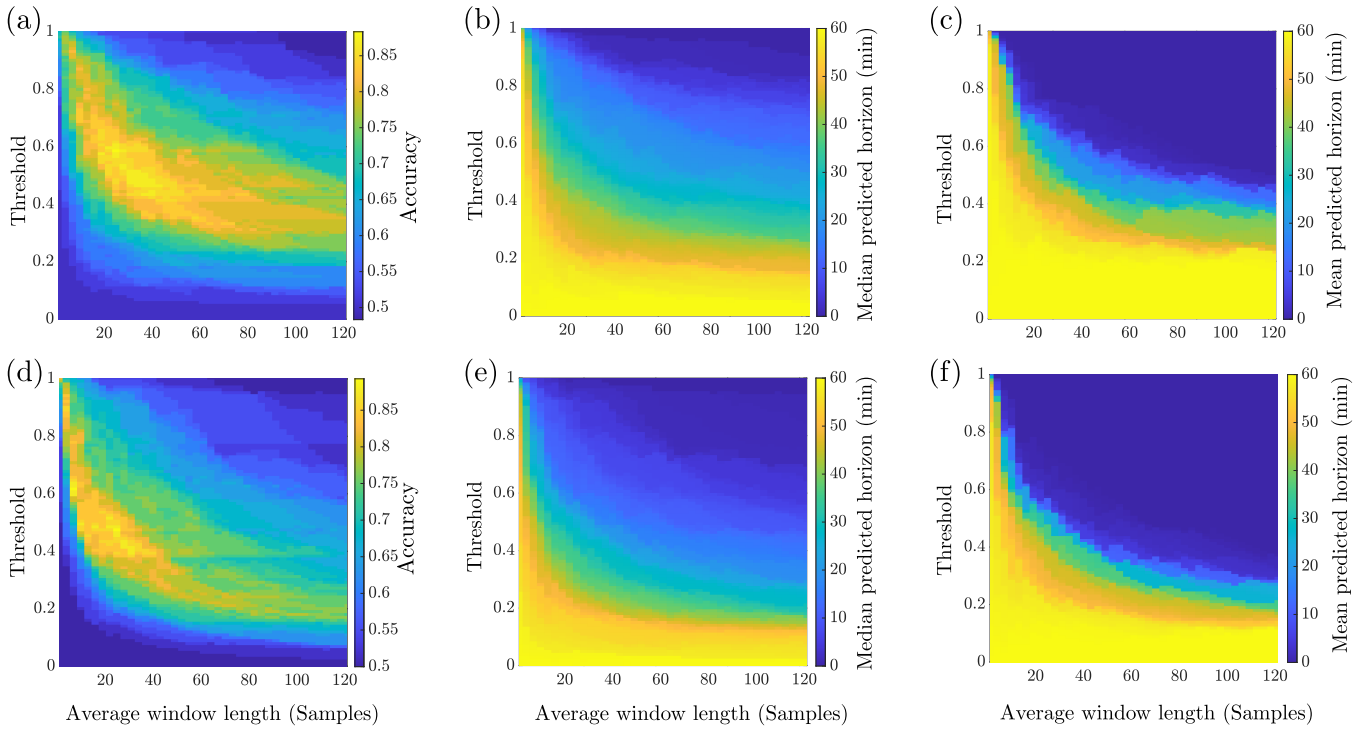


Figure S5: Tradeoffs on the validation data. **(a)** Model accuracy, **(b)** mean and **(c)** median of the predicted time horizon before AF onset on the validation set of the RRI data as a function of the probability threshold and the size of the moving average window that smooths the probability of danger. **(d, e, f)** Same plots computed for the ECG data.

accuracy, F1 score, and mean and median predicted time horizon. The resulting hyperparameters are a moving average window size of 7 samples and a threshold of 0.57.

For the case in which the CNN was trained on recurrence plots generated from ECG data (see section “Performance on ECG data”), we followed the same procedure to optimize hyperparameters. This led to 8 hyperparameter combinations that satisfy the criterion. The smallest moving average is 6 samples (corresponding to 1.25min windows) and the threshold is 0.48, maximizing accuracy, F1 score, and mean and median predicted time horizon. The mean predicted time horizon is 32.5min and the median 43.4min (Fig. S5d-f).

S7. Benchmarking. Table 3 presents the benchmarking outcomes of our method compared to earlier approaches for predicting atrial fibrillation (AF) both before and at its onset. Ebrahimzadeh et al.[26] reported the highest accuracy among all evaluated methods, reaching 98% by combining three classifiers (MLP, KNN, and SVM) to predict AF at its onset. Mohebbi et al.[27] also reported high performance, with an accuracy of 96%, using a combination of spectrum, bispectrum, and nonlinear features (e.g., sample entropy and Poincaré plot-extracted features) that were fed to a SVM for sample classification. However, their methodology required ECG segments of 30 minutes and was limited to predictions at the AF onset. Guo et al.[28] conducted a comprehensive analysis on a cohort of 554 patients, the largest reported in the table. Their results achieved an accuracy of 88% using an eXtreme Gradient Boosting (XGBoost) classifier on 17 engineered features from RRI data for AF prediction at its onset, requiring long data input segments of one to four hours. Although some of these models achieved high performance metrics, our proposed method, WARN, is the first that predicts AF substantially far from its onset.

S8. Supporting analysis. This section contains the supporting table, supporting figures, and captions of the supporting movies for the performance analysis of WARN.

Table S3: Observation on misclassification by WARN in the test dataset.

Sample	False Negatives
1	Very stable rhythm, only SR
2	Tachycardia, bradycardia, and noise
3	Tachycardia, unstable baseline and noise
4	Unstable baseline
Sample	**False Positives**
1	Atrial flutter
2, 6, 14, 19	Noise and unstable baseline
3, 8	Multiple PVCs and PACs
4, 15	PACs and noise
5	PACs and sinus tachycardia
7	Sinus tachycardia coupled with PACs and noise
9, 10, 16	Noise
11, 17, 22	PVCs, PACs and noise
12	Sinus tachycardia, multiple PACs and noise
13	Sinus tachycardia and unstable baseline
18	Multiple PACs and noise
20	PACs, unstable baseline and long RR intervals
21	PACs

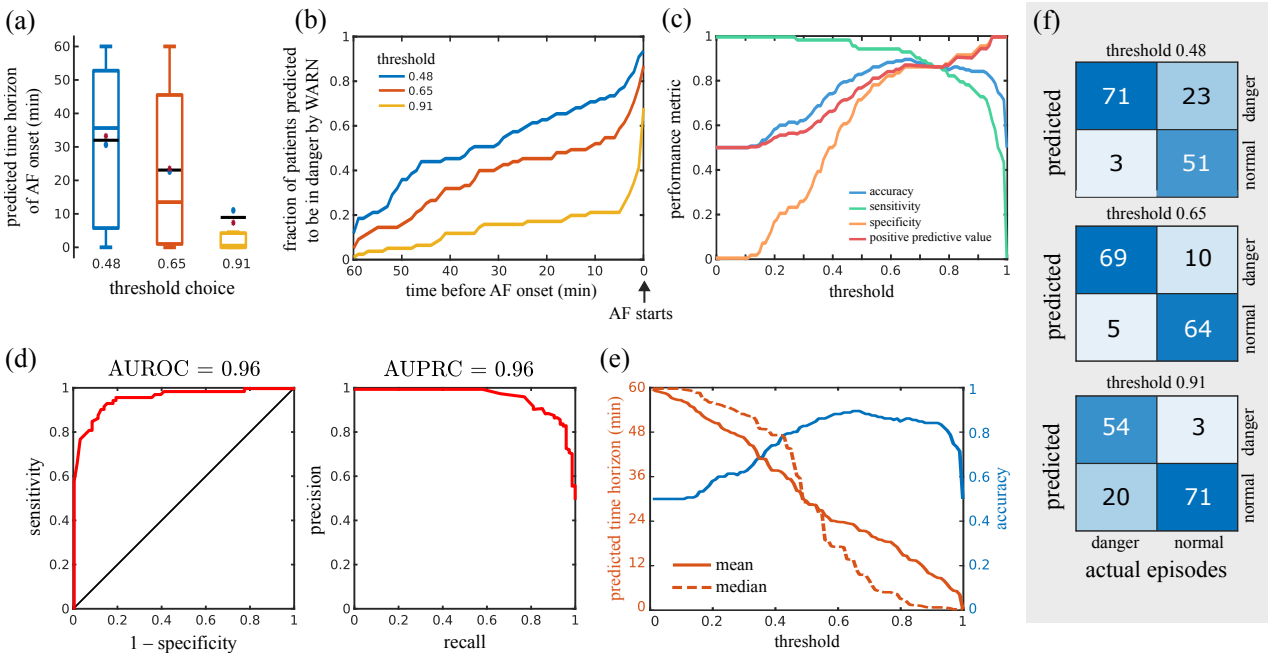


Figure S6: Performance of WARN on the test ECG dataset. **(a)** Boxplots of the predicted time horizon until AF onset for different probability thresholds across all patients. Median and mean values are marked by colored and black lines, respectively. **(b)** Fraction of patients predicted to be in danger as a function of time before the AF onset for different thresholds. **(c)** Performance metrics as a function of the probability threshold. **(d)** Receiver operator characteristic curve (left) and precision-recall curve (right). **(e)** Trade-off between the predicted time horizon and the model accuracy as a function of the probability threshold. **(f)** Confusion matrices for different thresholds.

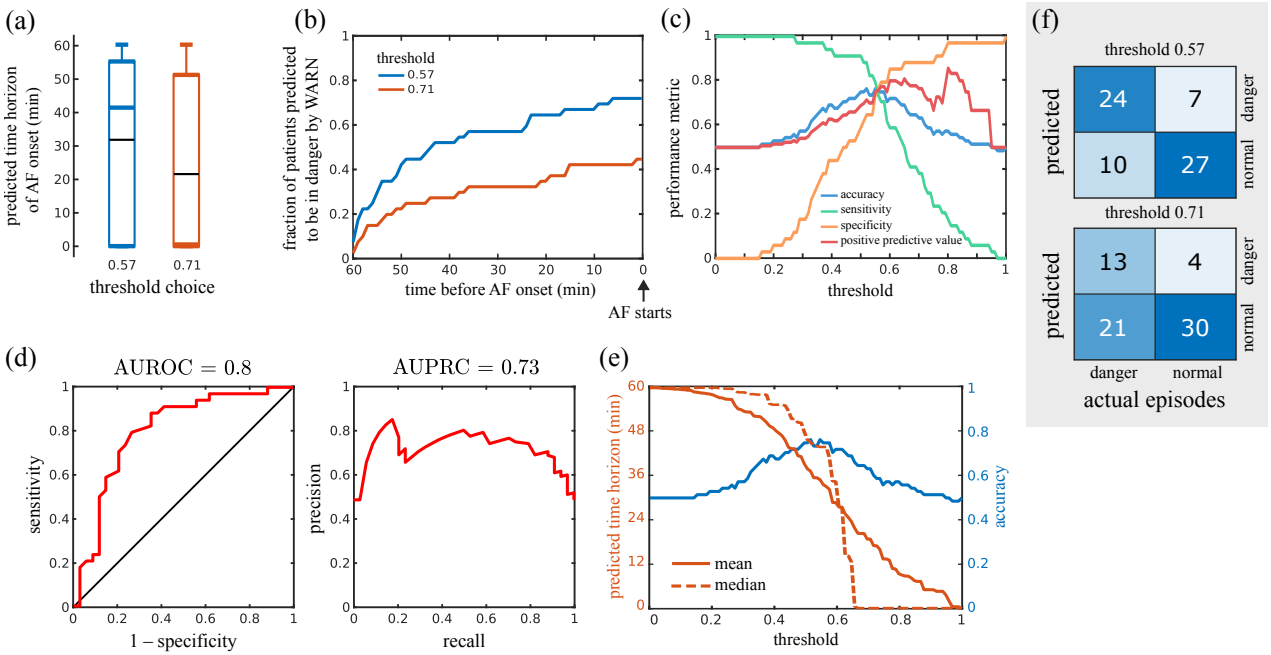


Figure S7: Performance of WARN on the external centers dataset. **(a)** Boxplots of the predicted time horizon until AF onset for different probability thresholds across all patients. Median and mean values are marked by colored and black lines, respectively.**(b)** Fraction of patients predicted to 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)** Receiver operator characteristic curve (left) and precision-recall curve (right). **(e)** Trade-off between the predicted time horizon and the model accuracy as a function of the probability threshold. **(f)** Confusion matrices for different thresholds.

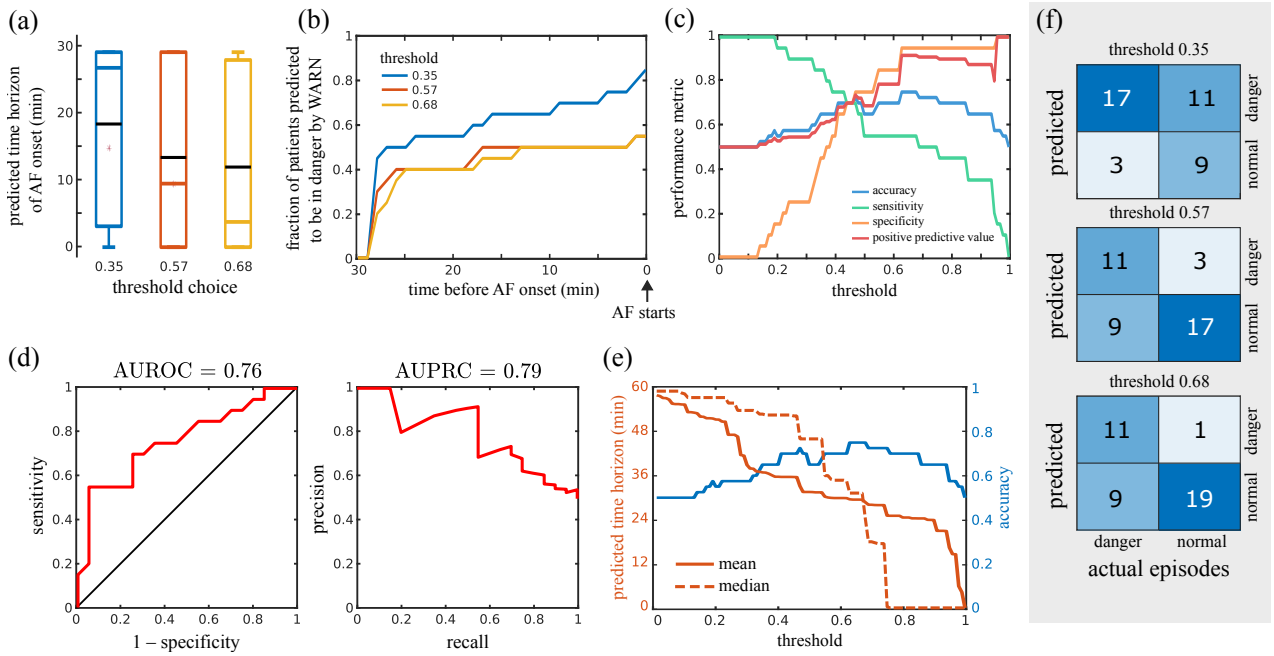


Figure S8: Performance of WARN on the Physionet challenge dataset. **(a)** Boxplots of the predicted time horizon until AF onset for different probability thresholds across all patients. Median and mean values are marked by colored and black lines, respectively.**(b)** Fraction of patients predicted 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)** Receiver operator characteristic curve (left) and precision-recall curve (right). **(e)** Trade-off between the predicted time horizon and the model accuracy as a function of the probability threshold. **(f)** Confusion matrices for different thresholds.

Supplemental References

1. Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation *101*, e215–e220.

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

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

4. Pan, J., and Tompkins, W. J. (1985). A real-time QRS detection algorithm. IEEE Transactions on Biomedical Engineering *BME-32*, 230–236.

5. Sedghamiz, H. (2014). MATLAB implementation of Pan Tompkins ECG QRS detector. MathWorks File Exchange. URL: `https://www.mathworks.com/matlabcentral/fileexchange/45840`.

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

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

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

9. Zhang, H., Liu, C., Zhang, Z., Xing, Y., Liu, X., Dong, R., He, Y., Xia, L., and Liu, F. (2021). Recurrence plot-based approach for cardiac arrhythmia classification using Inception-ResNet-v2. Frontiers in Physiology *12*, 648950.

10. Tan, M., and Le, Q. (2021). Efficientnetv2: Smaller models and faster training. arXiv:2104.00298.

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

12. Wu, J., Yu, Y., Huang, C., and Yu, K. (2015). Deep multiple instance learning for image classification and auto-annotation. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3460–3469.

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

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

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

16. Takens, F. Detecting strange attractors in turbulence. Springer (1981).

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

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

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

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

21. Bottou, L. Stochastic gradient descent tricks. Springer (2012).

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

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

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

25. Somani, S., Russak, A. J., Richter, F., Zhao, S., Vaid, A., Chaudhry, F., De Freitas, J. K., Naik, N., Miotto, R., Nadkarni, G. N. et al. (2021). Deep learning and the electrocardiogram: review of the current state-of-the-art. EP Europace *23*, 1179–1191.

26. Ebrahimzadeh, E., Kalantari, M., Joulani, M., Shahraki, R. S., Fayaz, F., and Ahmadi, F. (2018). 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.

27. Mohebbi, M., and Ghassemian, H. (2012). 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.

28. Guo, Y., Wang, H., Zhang, H., Liu, T., Li, L., Liu, L., Chen, M., Chen, Y., and Lip, G. Y. (2021). Photoplethysmography-based machine learning approaches for atrial fibrillation prediction: a report from the Huawei Heart Study. JACC: Asia *1*, 399–408.