ESC European Society of Cardiology

European Heart Journal - Digital Health (2023) **4**, 216–224
https://doi.org/10.1093/ehjdh/ztad018

**ORIGINAL ARTICLE**

# Automatic screening of patients with atrial fibrillation from 24-h Holter recording using deep learning

**Peng Zhang [1,2,†], Fan Lin[3,†], Fei Ma[3], Yuting Chen[1,2], Siyi Fang[1,2], Haiyan Zheng[4], Zuwen Xiang[5], Xiaoyun Yang[3,*], and Qiang Li[1,2,*]**

[1]Britton Chance Center for Biomedical Photonics, Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, 1037 Luoyu Road, Wuhan, Hubei 430074, China; [2]MoE Key Laboratory for Biomedical Photonics, Collaborative Innovation Center for Biomedical Engineering, School of Engineering Sciences, Huazhong University of Science and Technology, 1037 Luoyu Road, Wuhan, Hubei 430034, China; [3]Division of Cardiology, Department of Internal Medicine, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, 1095 Jiefang Avenue, Wuhan, Hubei 430030, China; [4]Department of Cardiovascular Medicine, Zigui County People's Hospital, 10 Changning Avenue, Yichang, Hubei 443600, China; and [5]Department of Rehabilitation of Traditional Chinese Medicine, Zigui County People's Hospital, 10 Changning Avenue, Yichang, Hubei 443600, China

*Received 4 December 2022; revised 25 February 2023; online publish-ahead-of-print 1 March 2023*

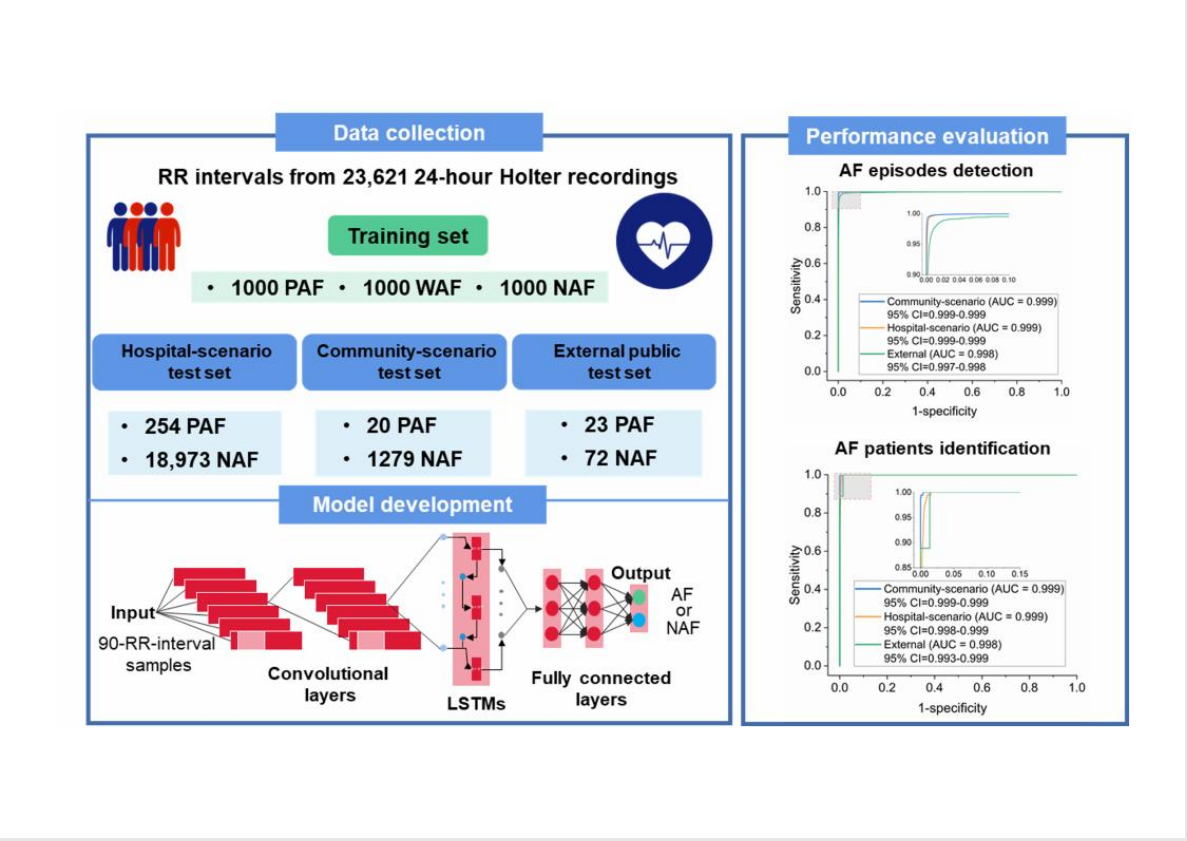
**Aims** As the demand for atrial fibrillation (AF) screening increases, clinicians spend a significant amount of time identifying AF signals from massive amounts of data obtained during long-term dynamic electrocardiogram (ECG) monitoring. The identification of AF signals is subjective and depends on the experience of clinicians. However, experienced cardiologists are scarce. This study aimed to apply a deep learning-based algorithm to fully automate primary screening of patients with AF using 24-h Holter monitoring.

**Methods and results** A deep learning model was developed to automatically detect AF episodes using RR intervals and was trained and evaluated on 23 621 (2297 AF and 21 324 non-AF) 24-h Holter recordings from 23 452 patients. Based on the AF episode detection results, patients with AF were automatically identified using the criterion of at least one AF episode lasting 6 min or longer. Performance was assessed on an independent real-world hospital-scenario test set (19 227 recordings) and a community-scenario test set (1299 recordings). For the two test sets, the model obtained high performance for the identification of patients with AF (sensitivity: 0.995 and 1.000; specificity: 0.985 and 0.997, respectively). Moreover, it obtained good and consistent performance (sensitivity: 1.000; specificity: 0.972) for an external public data set.

**Conclusion** Using the criterion of at least one AF episode of 6 min or longer, the deep learning model can fully automatically screen patients for AF with high accuracy from long-term Holter monitoring data. This method may serve as a powerful and cost-effective tool for primary screening for AF.

*Corresponding authors. Tel: +8615629037900, Fax: +027 83665460, Email: yangxiaoyun321@126.com (Xiaoyun Yang); Tel: +8618621108080, Fax: 027 87783003, Email: liqiang8@hust.edu.cn (Qiang Li)
†These authors shared first authorship.

**Graphical Abstract**




A deep learning algorithm uses RR interval data to automatically detect AF episodes and identify patients with AF. AF, atrial fibrillation; PAF, paroxysmal AF; WAF, whole-course AF; NAF, non-AF; AUC, area under the ROC curve; CI, confidence interval.

**Keywords** Deep learning • Atrial fibrillation • Electrocardiogram • Holter monitoring • Real-world clinical data

# Introduction

Atrial fibrillation (AF) is the most common tachyarrhythmia, with a lifetime risk of one in three, and it can lead to dangerous complications and increased cardiovascular mortality risk.[1,2] Paroxysmal AF (PAF) is associated with occasional or intermittent episodes, and long-term dynamic electrocardiogram (ECG) monitoring (Holter monitoring) is required to detect PAF in clinical practice.[3] Long-term Holter monitoring produces a large amount of ECG data, which clinicians must review for diagnosis. Moreover, the popularity of wearable devices has made the acquisition of heartbeat interval data increasingly convenient, making it possible to carry out AF screening in a large number of people.[4,5] However, the identification of AF signals depends heavily on the experience of clinicians, and existing clinician resources can hardly meet the requirements of screening for AF from these massive data. Therefore, it is important to develop automatic AF detection methods to improve the efficiency of AF screening.

Automatic AF detection has two main application scenarios. One is to assist clinicians in improving the accuracy and efficiency of their detection of AF in patient screening, and the other is to fully automate screening for AF without the immediate participation of clinicians. The former is applicable to the 'diagnosis of AF' scenario in which clinician resources are available and diagnostic accuracy requirements are high, whereas the latter is applicable to the 'primary screening for AF' scenario with insufficient clinician resources and relatively low diagnostic accuracy requirements. This study focused on the latter scenario, developing a fully automatic screening method for AF in a large patient population.

In recent years, many deep learning-based methods have been proposed for automatic AF detection and have achieved good performance on benchmark data sets.[6–15] However, most previous studies have mainly tested short ECG recordings from public data sets that include only a small number of patients.[16] Verification of automatic AF detection in large data sets from real-world environments (including various arrhythmias) is scarce.[17]

To address these challenges, we developed an RR interval-based deep learning method to fully automatically screen for AF using 24-h Holter recording data and evaluated its effectiveness using two large real-world clinical data sets. Distinguishing between AF and other arrhythmias with irregular RR intervals may be difficult using an RR interval-based method.[5,18–20] Therefore, we quantitatively assessed the performance of the method in distinguishing AF from seven other arrhythmias with irregular RR intervals.

# Methods

## Study design

Our primary goal was to develop a fully automatic method to screen for AF based on the RR intervals in long-term Holter monitoring. This method uses

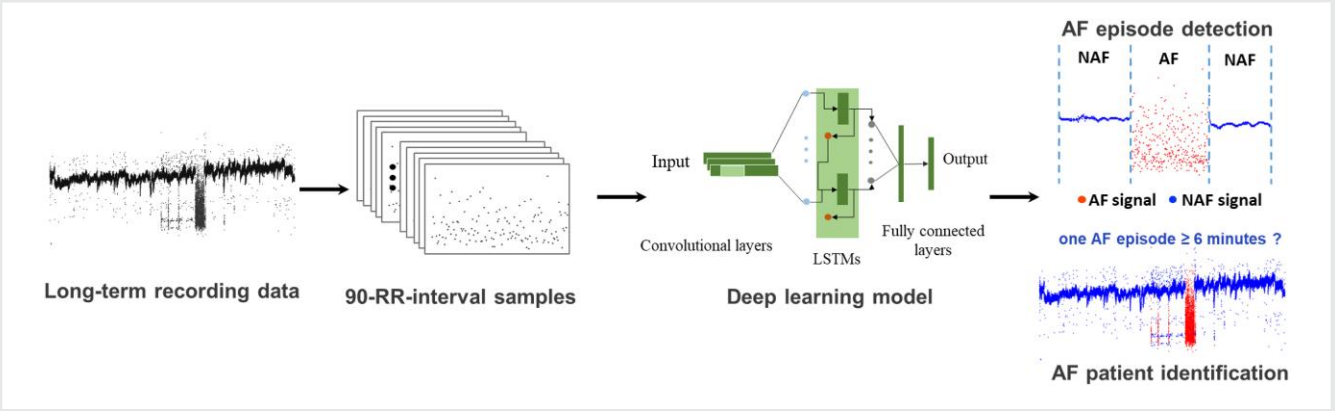



**Figure 1** Diagram for fully automatic atrial fibrillation screening. Each dot represents an RR interval. The dots of AF and NAF have been marked in the figure. In the atrial fibrillation patient identification, a 24-h recording (RR interval data) was diagnosed as an atrial fibrillation patient when it contained an atrial fibrillation episode of longer than 6 min. AF, atrial fibrillation; NAF, non-AF.

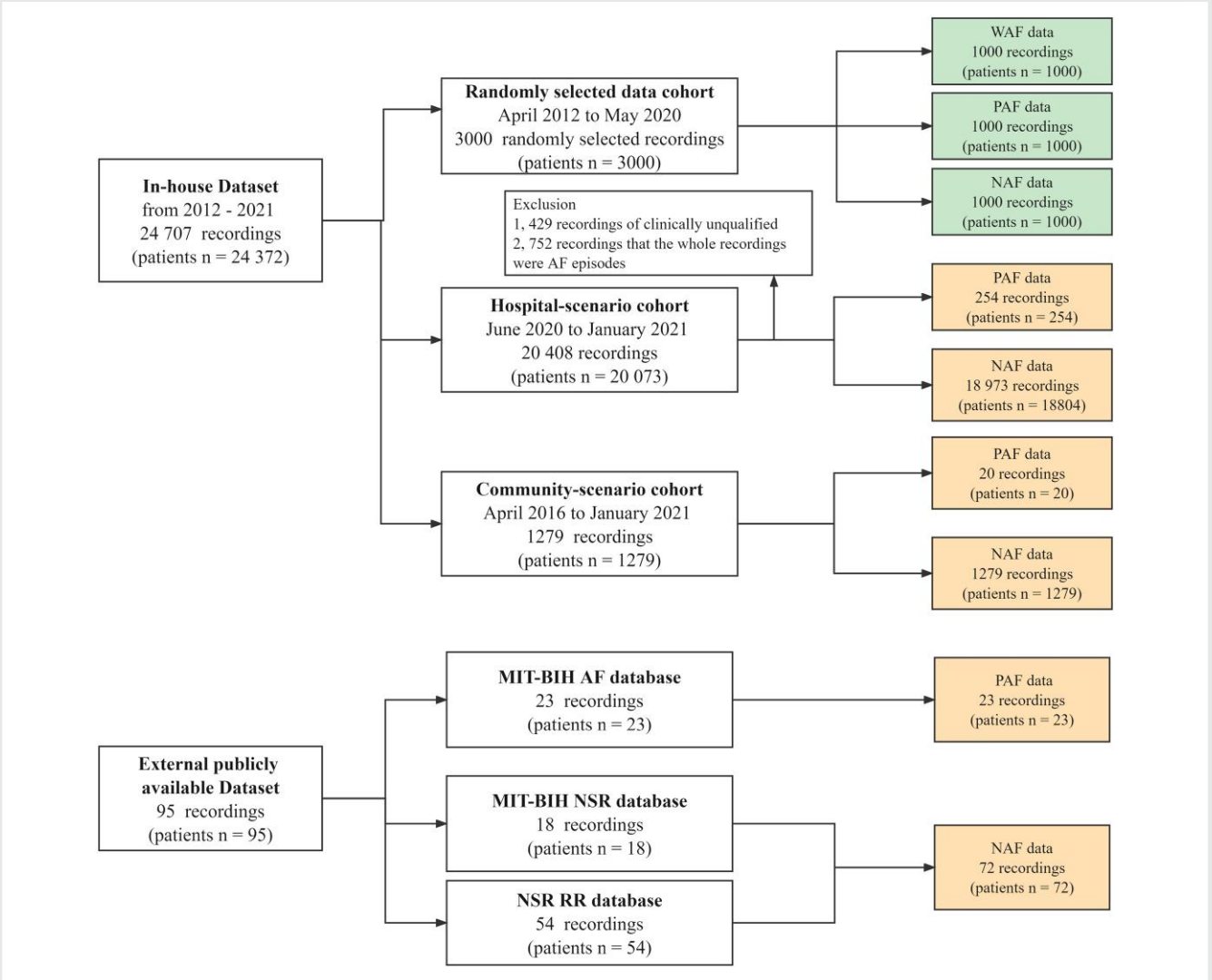



**Figure 2** Profile of the data sets. Data in the randomly selected data cohort were used for training the deep learning model, and data in the remaining cohorts were used for testing the model performance. PAF, paroxysmal atrial fibrillation (AF) patients; WAF, whole-course AF patients; NAF, non-AF patients.

**Table 1 Patient characteristics**

| | Randomly selected data ($n = 3000$) | | | Hospital-scenario ($n = 19\,227$) | | Community-scenario ($n = 1299$) | |
|---|---|---|---|---|---|---|---|
| | WAF ($n = 1000$) | PAF ($n = 1000$) | NAF ($n = 1000$) | PAF ($n = 254$) | NAF ($n = 18\,973$) | PAF ($n = 20$) | NAF ($n = 1279$) |
| **Age Group, *n* (%)** | | | | | | | |
| 18–25 | 0 (0) | 1 (0.1) | 38 (3.8) | 0 (0) | 383 (2.0) | 0 (0) | 51 (4.0) |
| 26–40 | 25 (2.5) | 19 (1.9) | 121 (12.1) | 7 (2.8) | 2235 (11.8) | 1 (5) | 393 (30.7) |
| 41–60 | 236 (23.6) | 274 (27.4) | 428 (42.8) | 61 (24.0) | 7596 (40.0) | 9 (45) | 595 (46.5) |
| 61–80 | 622 (62.2) | 623 (62.3) | 382 (38.2) | 158 (62.2) | 8102 (42.7) | 8 (40) | 239 (18.7) |
| ≥81 | 117 (11.7) | 83 (8.3) | 31 (3.1) | 28 (11.0) | 657 (3.5) | 2 (10) | 1 (0.1) |
| **Sex Group, *n* (%)** | | | | | | | |
| Male | 635 (63.5) | 653 (65.3) | 531 (53.1) | 181 (71.3) | 10 159(53.5) | 17 (85.0) | 671 (52.5) |
| Female | 365 (36.5) | 347 (34.7) | 469 (46.9) | 73 (28.7) | 8814 (46.5) | 3 (15.0) | 608 (47.5) |
| **Pacemaker, *n* (%)** | | | | | | | |
| | 1 (0.1) | 2 (0.2) | 1 (0.1) | 8 (3.1) | 188 (1.0) | 0 (0) | 0 (0) |
| **Frequent premature atrial contraction > 3000, *n* (%)** | | | | | | | |
| | 6 (0.6) | 359 (35.9) | 40 (4.0) | 85 (33.5) | 877 (4.6) | 5 (25) | 8 (0.6) |
| **Frequent ventricular premature contraction > 3000, *n* (%)** | | | | | | | |
| | 94 (9.4) | 46 (4.6) | 57 (5.7) | 11 (4.3) | 961 (5.1) | 0 (0) | 38 (3.0) |
| **First-degree AVB (AVB1), *n* (%)** | | | | | | | |
| | 0 (0) | 9 (0.9) | 43 (4.3) | 2 (0.1) | 577 (3.0) | 0 (0) | 0 (0) |
| **Second-degree AVB (AVB2), *n* (%)** | | | | | | | |
| | 0 (0) | 3 (0.3) | 9 (0.9) | 0 (0) | 345 (1.8) | 0 (0) | 0 (0) |
| **Third-degree AVB or complete heart block, *n* (%)** | | | | | | | |
| | 3 (0.3) | 0 (0) | 2 (0.2) | 0 (0) | 26 (0.1) | 0 (0) | 0 (0) |

Frequent premature atrial contraction (ventricular premature contraction) > 3000 means that there were more than 3000 atrial (ventricular) premature heartbeats during the 24-h Holter monitoring. AF, atrial fibrillation; PAF, paroxysmal AF; WAF, whole-course AF; NAF, non-AF; AVB, atrioventricular block.

deep learning technology without immediate intervention by clinicians. Our method consists of multiple steps, as shown in *Figure 1*. In this study, samples containing 90 RR intervals were extracted from 24-h Holter recordings. The deep neural network (DNN) then classified each 90-RR-interval sample as AF or non-AF (NAF). Finally, based on the results of AF episode detection, patients with AF were automatically identified using our proposed criterion of at least one continuous AF episode of 6 min or longer.

## Data sources

This study was approved by the ethical committee of Tongji Medical College, Huazhong University of Science and Technology (Institutional Review Board approval number 2022-S021). We constructed an in-house data set and an external publicly available data set to train and evaluate the deep learning model, as shown in *Figure 2*. The in-house data set consisted of 24 707 recordings collected from 24 372 adult patients (age > 18 years) who had a 24-h dynamic 12-lead ECG recording with a sampling rate of 512 Hz captured by a Holter machine (DMS Holter Company, Stateline, NV, USA) at Tongji Hospital (Huazhong University of Science and Technology, Wuhan, China). In-hospital and ambulatory patients were pooled together in our data set.

The in-house data set included three types of patient data: whole-course AF (WAF), PAF, and NAF. For WAF, the entire recording included only AF signals. Non-AF recording data had no AF signals but included normal sinus rhythm, sinus arrhythmia, atrial arrhythmia, ventricular arrhythmia, and atrioventricular block. The data of the patients with PAF included both AF and NAF signals. Because the WAF recordings were easily detected, the WAF data were used only for training the deep learning algorithm (model) and were not included in the data to test the performance of the model.

The profile of the in-house data set is shown in *Figure 2* and the characteristics of the patients are listed in *Table 1*. The in-house data set contained the following three cohorts:

(1) A randomly selected data cohort was created to include 3000 recordings (1000 WAF, 1000 PAF, and 1000 NAF) from 3000 adult patients randomly selected from the patient pool enrolled between April 2012 and May 2020. This cohort was employed only for training the deep learning model; therefore, strict inclusion and exclusion criteria were used, as shown in the *Supplementary Methods*.

(2) The hospital-scenario cohort included 20 408 recordings from all consecutive adult patients ($n = 20\,073$) who received 24-h dynamic 12-lead ECG monitoring at Tongji Hospital between June 2020 and January 2021. After excluding 429 recordings shorter than 16 h or clinically unqualified and 752 WAF recordings, 19 227 recordings were used as the test set. More details are provided in the *Supplementary Methods*. It is worth mentioning two special groups in the 254 PAF recordings. The first group included 54 PAF recordings in which the AF episodes were <6 min. These recordings could be tested at the sample level, but not at the patient level (see Algorithm Evaluation Section), because the patients with AF were identified based on a 6-min duration of AF episode. The second group included 53 patients with PAF whose recordings contained AF and NAF signals that were too ambiguous to be clearly distinguished by clinicians. These recordings could be tested at the patient level, but not at the sample level. This was a real-world clinical cohort that was employed only to test the deep learning model. Moreover, it represents the clinical application scenario of screening for AF in inpatient and outpatient departments.

(3) The community-scenario cohort included 20 recordings from 20 patients with PAF and 1279 recordings from 1279 individuals who did not report heart disease and were recruited to receive 24-h Holter monitoring at Tongji Hospital. More details are provided in the *Supplementary Methods*. This cohort was employed only to test the deep learning model and represented the application scenario of screening individuals for AF in a normal population.

In addition, three publicly available databases were used as the external test set: the MIT-BIH AF database,[21] MIT-BIH NSR database,[22] and NSR

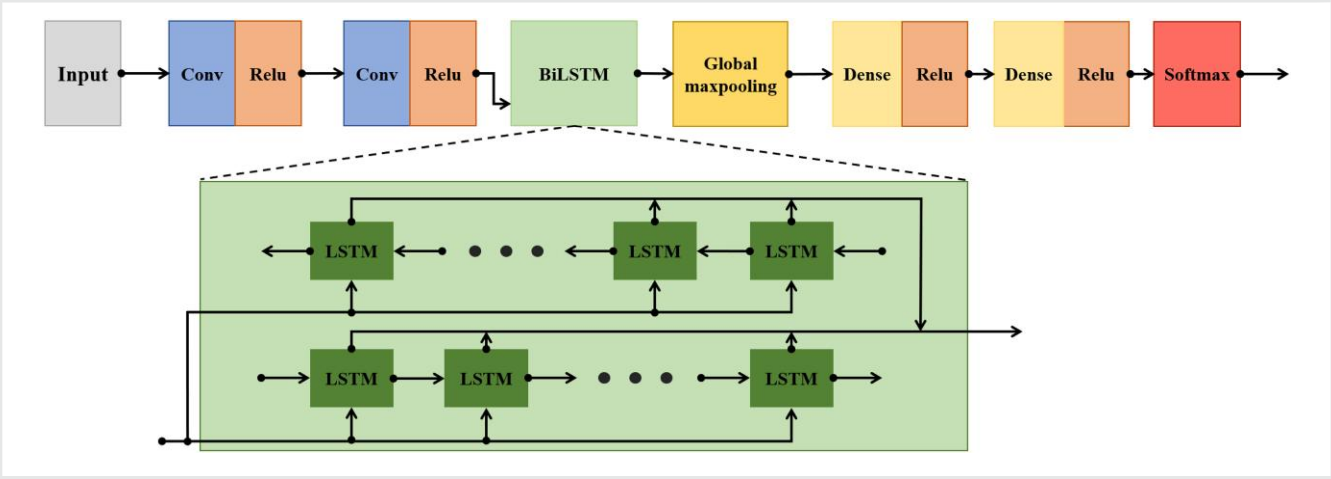



**Figure 3** Architecture of the deep learning model.

RR Interval database.[22] The MIT-BIH AF database includes 23 PAF recordings of 10 h from 23 patients. The MIT-BIH NSR and NSR RR interval databases include 18 and 54 NAF recordings of 24 h, respectively. The recordings in this test set were recorded using devices that were different from those used in the community- and hospital-scenario test sets and were used to test the generalization ability of the proposed method.

## Data pre-processing

Raw ECG signals were pre-processed using manufacturer-specific commercial software for the DMS CardioSca12 Satellite System (DMS Holter Company, Stateline, NV, USA) to detect noise and obtain RR interval data. Then, the RR interval data of a recording were divided into segments (samples) of 90 RR intervals for testing; samples with detected noise were removed. On average, 137 samples (12.4%) were removed from each recording. The deep learning model used in this study was trained and evaluated based on the remaining RR interval samples.

Each 90-RR-interval sample was labelled as either an AF (positive) or an NAF (negative) sample according to the labelled start and end times of the AF episodes. The samples extracted from patients designated WAF were all AF, and those from patients designated NAF were all NAF. Some testing samples extracted from patients with PAF contained both AF and NAF RR intervals and were named mixed samples. According to previous studies,[10,23] a mixed sample was labelled as AF when the percentage of annotated AF beats in the sample was equal to or >50%; otherwise, it was labelled NAF.

## Annotation procedures

All 24-h ECG data in the in-house data set underwent additional annotation. They were initially interpreted by primary cardiologists, and the randomly selected data cohort was further reviewed by three senior board-certified cardiologists. The other two cohorts were reviewed by one of the three senior board-certified cardiologists to ensure the correctness of the base diagnostic labels. Each AF episode included accurately labelled start and end times for patients with PAF. The start and end times of each AF episode were the time corresponding to the first atrial wave with an atrial rate >350 beats/min and the time corresponding to the first *P*-wave with sinus rhythm after the termination of AF.

Moreover, each interval of premature beat or tachycardia was marked 'A' (atrium event) or 'V' (ventricle event) for further labelling six types of arrhythmias with irregular RR intervals: premature atrial contraction (PAC), frequent premature atrial contraction (FPAC), ventricular premature contraction (VPC), frequent ventricular premature contraction (FVPC), atrial tachycardia (AT), and ventricular tachycardia (VT). The long RR interval caused by QRS wave dropping was marked 'B' to further label second-degree atrioventricular block (AVB2). The detailed standards are presented in Supplementary material online, *Table S1*. Because a 90-RR-interval sample might contain 'A', 'B', and 'V' at the same time, the sample could have multiple labels and was evaluated independently for each label during testing. The labels of the data were consistent with the diagnostic results of the clinical cardiologists.

## Deep learning model

In this study, we constructed a convolutional, long short-term memory, and fully connected DNN (CLDNN) to automatically detect AF episodes. The architecture of the model, illustrated in *Figure 3*, is composed of three modules: convolutional neural networks (CNNs), long short-term memory networks (LSTMs), and DNNs. Specifically, the CNNs were composed of two convolutional layers with 64 kernels of size 5 and 32 kernels of size 3, respectively. After the convolutional layers, we used the bidirectional LSTM to utilize the forward and backward information of the input. We took the output of all time steps as the output of the bidirectional LSTM and used them as the input of the global max pooling layer. The DNN module consisted of two fully connected layers. Before each fully connected layer, we applied dropout with a probability of 0.2 to prevent overfitting and improve the generalization ability. The final fully connected softmax layer produced, as the output of our model, a distribution over AF and NAF. The ReLU activation function was applied after each layer except the output layer. For each input of a 90-RR-interval sample, the model outputs a predicted label of AF or NAF. Additional technical details are provided in the *Supplementary Methods*.

## Algorithm evaluation

In this study, each 24-h recording was divided into many 90-RR-interval samples (segments), as described in the data pre-processing section. Our deep learning algorithm detected each 90-RR-interval sample as AF or NAF. Atrial fibrillation episode detection was assessed at the level of each 90-RR-interval sample, referred to as sample-level results. At this level, the detection results were compared with the reference standard for each 90-RR-interval sample. Based on the sample-level results, we further identified whether a patient could be diagnosed to have AF, referred to as patient-level results. At the patient level, a patient was identified as having AF if his/her recording included at least one continuous AF episode of 6 min or longer.

## Statistical analysis

We used different performance metrics to evaluate the performance of our model, including area under the receiver operating characteristic (ROC) curve (AUC), sensitivity, specificity, and accuracy. We used a two-sided 95% confidence interval (CI) to evaluate data variability for each metric.[24] The CI for the AUC was estimated using the DeLong method,[25] whereas

**Table 2 Atrial fibrillation detection performance at the 'sample-level' for all patients**

| Type of cohort | AUC (95% CI) | Sensitivity (95% CI) | Specificity (95% CI) | Accuracy (95% CI) |
|---|---|---|---|---|
| Community-scenario | 0.999 (0.999–0.999) | 0.993 (0.991–0.995) | 0.999 (0.999–0.999) | 0.999 (0.999–0.999) |
| Hospital-scenario | 0.999 (0.999–0.999) | 0.992 (0.991–0.993) | 0.997 (0.997–0.997) | 0.997 (0.997–0.997) |
| External | 0.998 (0.997–0.998) | 0.966 (0.962–0.973) | 0.994 (0.993–0.995) | 0.992 (0.991–0.993) |

AF, atrial fibrillation; AUC, area under the curve; CI, confidence interval.

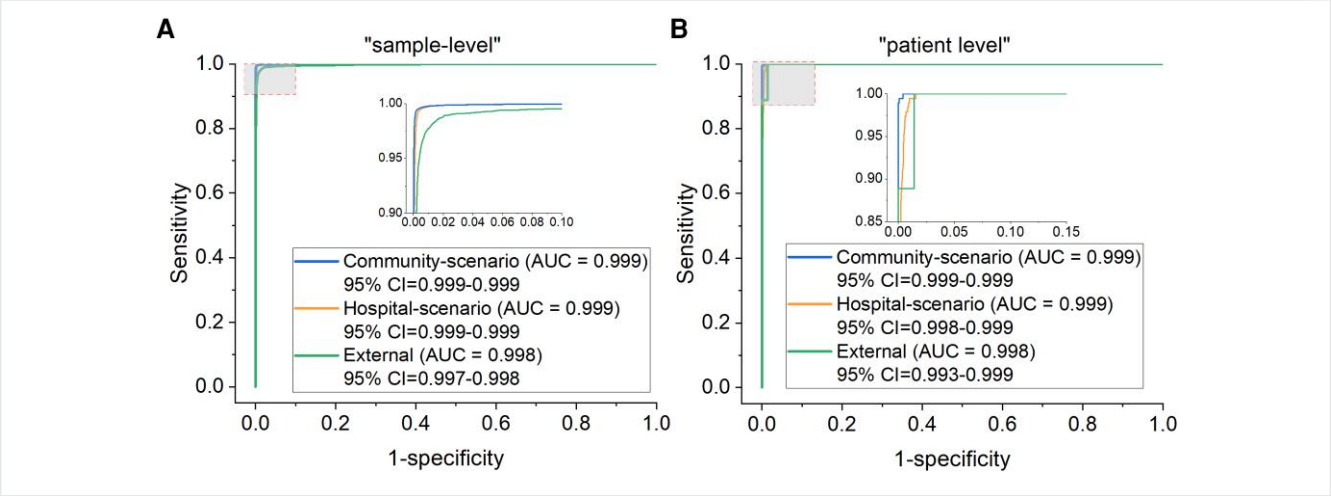



**Figure 4** The receiver operating characteristic curves for the 'sample-level' and the 'patient level' analyses in the three test sets. A, The results at the 'sample-level'. B, The results at the 'patient level'.

those for the other metrics were obtained using the bootstrap method with 2000 replications. More details are provided in the *Supplementary Methods*.

# Results

## Performance of the deep learning model at the sample level

None of the test sets had been included in model training. The performance of the model was first evaluated at the sample level for the three test sets, and the results are shown in *Table 2*. Specifically, the model achieved a sensitivity of 0.993 and specificity of 0.999 for the community-scenario test set, a sensitivity of 0.992 and specificity of 0.997 for the hospital-scenario test set, and a sensitivity of 0.966 and specificity of 0.994 for the external test set.

*Figure 4A* shows the ROC curves for the sample-level analyses of all testing recordings in the three test sets. For the ROC curves, our model achieved an AUC of 0.999, 0.999, and 0.997 for the community-scenario, hospital-scenario, and external test sets, respectively. In general, our model achieved very good performance for the community-scenario and hospital-scenario test sets, and it also obtained good and consistent performance for the external test set, indicating good generalization ability.

## Performance of the deep learning model at the patient level

Based on the results at the sample level, we tested the performance of our model at the patient level with different identification criteria of at least one AF episode of 3 min, 6 min, or 9 min in the hospital-scenario and community-scenario test sets, and the results are shown in *Table 3*. For the hospital-scenario test set, all patients with PAF were successfully identified using the criterion of 3 min; however, the false-positive rate was 0.032. Using the criterion of 6 min, the sensitivity decreased by 0.005, and the false positive rate decreased by 0.017. Using the criterion of 9 min, performance did not improve. The performance changes resulting from the criterion changes were consistent for the community-scenario test set. The method based on the criterion of shorter time will have better practicability, and after weighing the sensitivity and specificity, the criterion of 6 min was selected.

By the use of the criterion of at least one AF episode lasting 6 min for identifying patients with AF, the results for the three test sets at the patient level are shown in *Table 4* and *Figure 4B*. Our model achieved consistently higher performance for the community-scenario than the hospital-scenario test set. The performance on the external test set was slightly weaker but was still sufficiently good (AUC: 0.998). This indicates that our method obtained very strong and robust performance in the automatic screening for patients with AF.

## Distinction between AF and other arrhythmias with irregular RR intervals

Based on the results of sample-level analyses, we further quantitatively evaluated the performance of our model in distinguishing AF from other arrhythmias with irregular RR intervals in the NAF recordings of the community-scenario and hospital-scenario test sets. Seven types of arrhythmias with irregular RR intervals were included, and the detection results at the sample level are shown in *Table 5*. Our model achieved an average accuracy of 0.999 and 0.986 for specifically

**Table 3 Atrial fibrillation detection performance at the 'patient level' using different thresholds**

| Type of cohort | Threshold (min) | Sensitivity (95% CI) | Specificity (95% CI) | Accuracy (95% CI) |
|---|---|---|---|---|
| Community-scenario | 3 | 1.000 (1.000–1.000) | 0.988 (0.981–0.997) | 0.988 (0.981–0.997) |
| | 6 | 1.000 (1.000–1.000) | 0.997 (0.994–1.000) | 0.997 (0.994–1.000) |
| | 9 | 0.950 (0.800–1.000) | 1.000 (1.000–1.000) | 0.999 (0.997–1.000) |
| Hospital-scenario | 3 | 1.000 (1.000–1.000) | 0.968 (0.966–0.971) | 0.968 (0.966–0.972) |
| | 6 | 0.995 (0.9841.000) | 0.985 (0.984–0.988) | 0.985 (0.984–0.998) |
| | 9 | 0.990 (0.969–1.000) | 0.990 (0.989–0.992) | 0.990 (0.989–0.992) |

CI, confidence interval.

**Table 4 Atrial fibrillation detection performance at the 'patient level' using the criterion of 6 min on three test sets**

| Type of cohort | AUC (95% CI) | Sensitivity (95% CI) | Specificity (95% CI) | Accuracy (95% CI) |
|---|---|---|---|---|
| Community-scenario | 0.999 (0.999–0.999) | 1.000 (1.000–1.000) | 0.997 (0.994–1.000) | 0.997 (0.994–1.000) |
| Hospital-scenario | 0.999 (0.998–0.999) | 0.995 (0.984–1.000) | 0.985 (0.984–0.988) | 0.985 (0.984–0.988) |
| External | 0.998 (0.993–0.999) | 1.000 (1.000–1.000) | 0.972 (0.931–0.986) | 0.978 (0.944–0.989) |

CI, confidence interval.

detecting NAF samples in all seven types of arrhythmias in the community- and hospital-scenario test sets, respectively, with a relatively lower accuracy of 0.858 for AT in the hospital-scenario test set.

## Analysis of misclassified cases at the patient level

Finally, we analysed the misclassified cases at the patient level for the community- and hospital-scenario test sets. Only four cases in the community-scenario test set were misclassified; the misclassifications in the hospital-scenario test set are shown in *Table 6*. In the hospital-scenario test set, only one case of PAF was misclassified as NAF, and 278 NAF cases were misclassified as PAF, including atrial high-rate episodes (AHREs), FPAC, FVPC, and AVB2. Specifically, the misclassification rates in the AHREs, FPAC, FVPC, and AVB2 cases were 16.8%, 7.75%, 2.50%, and 2.32%, respectively. The misclassification rate of AHREs and FPAC was significantly higher than that in all NAF cases (1.50%). In general, our RR interval-based deep learning model effectively screened AF cases from NAF cases; however, patients with other arrhythmias with irregular RR intervals (especially AHREs and FPAC) were more easily misclassified as AF than other patients without AF.

# Discussion

This study aimed to develop a fully automatic screening method for AF in patients without immediate clinician intervention, for the 'primary screening for AF' scenario with insufficient clinician resources. At present, there is no well-recognized criterion for artificial intelligence (AI)-based automatic patient screening for AF using 24-h Holter monitoring; therefore, fully automatic AF screening cannot be achieved. In clinical practice, the 2020 ESC guidelines for the diagnosis and management of AF suggest that patients be diagnosed with AF when their data include at least one AF episode lasting 30 s or longer.[26] This is the criterion for clinical diagnosis of AF in patients; however, it cannot be applied in automatic AF screening because current automatic AF detection methods cannot meet this accuracy requirement. For the primary AF screening that is the aim of this study, our method identified AF in patients by using a relatively relaxed criterion of at least one AF episode of 6 min or longer. This criterion was used in the old version of the ESC guidelines, and this more relaxed condition significantly increased the risk of thromboembolism (stroke, transient ischaemic attack, or systemic embolism).[27,28] Our method achieved very high performance with this criterion, and we believe that it may be a general criterion for AI-based automatic identification of AF in patients in long-term Holter monitoring. However, the implications of our method for AF screening in clinical practice remain unclear. Compared with single-time-point or symptom-based AF screening, whether AF screening through long-term Holter monitoring using this criterion has similar clinical significance needs further investigation.

This study used RR intervals as the input to the deep learning model for two reasons: (i) the irregularity of RR intervals is a main feature of AF[18]; (ii) compared with raw ECG waves, RR interval data can be easily obtained from Holter devices with different leads and various wearable devices and require lower computational cost. In addition, a recent study by Han *et al.* [29] showed that common perturbations (adversarial attacks) to single-lead ECG data could lead to a 74% misdiagnosis rate for a raw-ECG-based deep learning model. In contrast, a method based on RR intervals appeared to be more robust to such adversarial attacks.[29] Our experiments further showed that our RR interval-based method could distinguish (with an average accuracy of more than 0.986 at the sample level) AF from seven types of other arrhythmias with irregular RR intervals, which are commonly considered to be difficult to recognize from RR intervals.[19,20] At the patient level, although the identification accuracy of AHREs (including atrial flutter) and FPAC cases was lower than that of other NAF cases, the overall identification accuracy at the patient level still exceeded 0.98. In general, this study has proven that the RR interval-based method is a good choice for automatic AF screening in long-term monitoring.

The data used in most AI-related clinical studies are selected with strict inclusion criteria, and marginal or uncertain data are excluded, which leads to a decline in model performance in real clinical application.[30] To minimize selection bias, our model was verified using real-

**Table 5 The results of distinguishing atrial fibrillation from other arrhythmias with irregular RR interval**

| | Community-scenario cohort (sensitivity:0.993) | | Hospital-scenario cohort (sensitivity:0.992) | |
|---|---|---|---|---|
| **Type** | **Number of samples** | **Specificity (95% CI)** | **Number of samples** | **Specificity (95% CI)** |
| FPAC | 3268 | 0.998 (0.996–0.999) | 392 875 | 0.973 (0.965–0.966) |
| PAC | 25 879 | 0.998 (0.998–0.999) | 1 099 797 | 0.995 (0.995–0.995) |
| FVPC | 19 455 | 0.998 (0.998–0.999) | 502 958 | 0.990 (0.992–0.992) |
| VPC | 36 251 | 0.999 (0.999–0.999) | 972 031 | 0.995 (0.995–0.995) |
| AT | 959 | 0.989 (0.983–0.996) | 109 723 | 0.858 (0.852–0.826) |
| VT | 71 | 1.000 (1.000–1.000) | 9855 | 0.942 (0.969–0.975) |
| AVB2 | 0 | N/A | 29 387 | 0.950 (0.899–0.906) |
| All seven types | 85 883 | 0.999 (0.999–0.999) | 3 116 626 | 0.986 (0.985–0.985) |
| Other NAF | 1 160 406 | 0.999 (0.999–0.999) | 15 156 223 | 0.999 (0.999–0.999) |
| All | 1 244 846 | 0.999 (0.999–0.999) | 18 048 239 | 0.997 (0.997–0.997) |

Seven types of arrhythmia with irregular RR interval were included: premature atrial contraction (PAC), frequent premature atrial contraction (FPAC), ventricular premature contraction (VPC), frequent ventricular premature contraction (FVPC), atrial tachycardia (AT), ventricular tachycardia (VT), and second-degree atrioventricular block (AVB2).

**Table 6 The analysis of misclassified cases at the 'patient level' for the hospital-scenario test set**

| Types of arrhythmia | False-positive patients ($n = 278$) |
|---|---|
| Second-degree atrioventricular block | 8 |
| Atrial high-rate episodes | 42 |
| Frequent premature atrial contraction | 68 |
| Frequent ventricular premature contraction | 24 |
| Others | 136 |

world data that included patients with PAF and patients with various other arrhythmias. This truly reflects the performance of the deep learning model in clinical applications and exposes shadow spots for some special data. The expert committee re-analysed the data misclassified by the model in this study and found that patients with AHREs were regularly misclassified as having AF. Atrial high-rate episodes (including atrial flutter) are classified as subclinical AF and convey a high risk of AF,[31] and such 'misclassifications' may have a positive impact on the prevention of AF. In general, the analysis of misclassifications in real-world clinical data helps us find shadow spots in our model for the detection of specific rare and special ECG signals. Accordingly, the application range can be set in advance, and the model can be further improved in later stages. Therefore, real-world data testing is an important step in the practical application of deep learning models. In addition, the potential methods for avoiding such misclassified cases may include (i) further improving the performance of the deep learning model to reduce misclassified cases and (ii) requiring clinicians to check the patients identified as having AF by the deep learning model to exclude the misclassified cases.

Our study had some limitations. First, the performance of the deep learning model may be improved in the future to further reduce the misclassification rate in AF identification and realize the ability to identify PAF in patients with AF episodes of <6 min. A second limitation is that the deep learning model achieved relatively lower accuracy in distinguishing AF from AT (0.858) compared with the average accuracy (0.986) over the seven arrhythmias with irregular RR intervals as a whole. More training data of AT may be required to improve its detection accuracy. In addition, annotation of the PAF recordings in this study was time-consuming. Minimizing this process through a semi-supervised learning-based automatic AF detection method could significantly reduce the workload of clinicians.[12] Finally, the effectiveness of the deep learning model for wearable devices with photoplethysmography requires further verification.

# Conclusions

We developed a deep learning model to fully automatically screen for AF in patients using long-term Holter monitoring data without the immediate participation of clinicians. Our method was evaluated on two large real-world clinical data sets and an external public data set, and consistently achieved good performance. In addition, we demonstrated that the RR interval-based method could effectively distinguish AF from other arrhythmias with irregular RR intervals. Our method has great potential for wide application in the primary screening for AF and will promote AF screening at a lower cost.

# Supplementary material

Supplementary material is available at *European Heart Journal – Digital Health*.

## Funding

This work was funded in part by the National Natural Science Foundation of China (62006087, 81500328), the National Key Research and Development Program of China (2022YFE0200600), the Science Fund for Creative Research Group of China (61721092), and the Director Fund of WNLO.

**Conflict of interest:** For the relationship with industry, P.Z., Y.C., and Q.L. are partially supported by the United Imaging Surgical Healthcare, Co., Ltd. All other authors have reported that they have no relationships relevant to the contents of this paper to disclose.

## Data availability

The publicly available data sets MIT-BIH AF database, MIT-BIH NSR database, and NSR RR Interval database are available at: https://physionet.org/content/afdb/1.0.0/, https://physionet.org/content/nsrdb/1.0.0/, and https://physionet.org/content/nsr2db/1.0.0/, respectively.

Restrictions apply to the availability of the in-house data, which were used with institutional permission through IRB approval for the current study, and are thus not publicly available. Please email all requests for academic use of raw and processed data to the corresponding author. The code for the deep learning model in this study is available at: https://codeocean.com/capsule/0 201 225/tree/v1; https://github.com/hustzp/Fully-automatic-AF-screening

## References


1. Rizas KD, Freyer L, Sappler N, von Stülpnagel L, Spielbichler P, Krasniqi A, *et al.* Smartphone-based screening for atrial fibrillation: a pragmatic randomized clinical trial. *Nat Med* 2022;**28**:1823–1830.
2. Kornej J, Benjamin EJ, Magnani JW. Atrial fibrillation: global burdens and global opportunities. *Heart* 2021;**107**:516–518.
3. Sanna T, Diener HC, Passman RS, Di Lazzaro V, Bernstein RA, Morillo CA, *et al.* Cryptogenic stroke and underlying atrial fibrillation. *N Engl J Med* 2014;**370**:2478–2486.
4. Perez MV, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, *et al.* Large-scale assessment of a smartwatch to identify atrial fibrillation. *N Engl J Med* 2019;**381**:1909–1917.
5. Pereira T, Tran N, Gadhoumi K, Pelter MM, Do DH, Lee RJ, *et al.* Photoplethysmography based atrial fibrillation detection: a review. *NPJ Digit Med* 2020;**3**:3.
6. Jo YY, Cho Y, Lee SY, Kwon JM, Kim KH, Jeon KH, *et al.* Explainable artificial intelligence to detect atrial fibrillation using electrocardiogram. *Int J Cardiol* 2021;**328**:104–110.
7. Xiong Z, Stiles MK, Zhao J. Robust ECG signal classification for detection of atrial fibrillation using a novel neural network. *Comput Cardiol* 2017;**2017**:1–4.
8. Andersen RS, Peimankar A, Puthusserypady S. A deep learning approach for real-time detection of atrial fibrillation. *Expert Syst Appl* 2019;**115**:465–473.
9. Pourbabaee B, Roshtkhari MJ, Khorasani K. Deep convolutional neural networks and learning ECG features for screening paroxysmal atrial fibrillation patients. *IEEE T Syst Man Cy-S* 2018;**48**:2095–2104.
10. Xia Y, Wulan N, Wang K, Zhang H. Detecting atrial fibrillation by deep convolutional neural networks. *Comput Biol Med* 2018;**93**:84–92.
11. Cai W, Chen Y, Guo J, Han B, Shi Y, Ji L, *et al.* Accurate detection of atrial fibrillation from 12-lead ECG using deep neural network. *Comput Biol Med* 2020;**116**:103378.
12. Zhang P, Chen Y, Lin F, Wu S, Yang X, Li Q. Semi-supervised learning for automatic atrial fibrillation detection in 24-hour Holter monitoring. *IEEE J Biomed Health Inform* 2022;**26**:3791–3801.
13. Liu S, Wang A, Deng X, Yang C. MGNN: a multiscale grouped convolutional neural network for efficient atrial fibrillation detection. *Comput Biol Med* 2022;**148**:105863.
14. Fan X, Yao Q, Cai Y, Miao F, Sun F, Li Y. Multiscaled fusion of deep convolutional neural networks for screening atrial fibrillation from single lead short ECG recordings. *IEEE J Biomed Health Inform* 2018;**22**:1744–1753.
15. Gao Y, Wang H, Liu Z. An end-to-end atrial fibrillation detection by a novel residual-based temporal attention convolutional neural network with exponential nonlinearity loss. *Knowl Based Syst* 2021;**212**:106589.
16. Olier I, Ortega-Martorell S, Pieroni M, Lip GY. How machine learning is impacting research in atrial fibrillation: implications for risk prediction and future management. *Cardiovasc Res* 2021;**117**:1700–1717.
17. Somani S, Russak AJ, Richter F, Zhao S, Vaid A, Chaudhry F, *et al.* Deep learning and the electrocardiogram: review of the current state-of-the-art. *Europace* 2021;**23**:1179–1191.
18. Faust O, Shenfield A, Kareem M, San TR, Fujita H, Acharya UR. Automated detection of atrial fibrillation using long short-term memory network with RR interval signals. *Comput Biol Med* 2018;**102**:327–335.
19. Rizwan A, Zoha A, Mabrouk IB, Sabbour HM, Al-Sumaiti AS, Alomainy A, *et al.* A review on the state of the art in atrial fibrillation detection on enabled by machine learning. *IEEE Rev Biomed Eng* 2021;**14**:219–239.
20. Petrutiu S, Ng J, Nijm GH, AlAngari H, Swiryn S, Sahakian AV. Atrial fibrillation and waveform characterization. *IEEE Eng Med Biol* 2006;**25**:24–30.
21. Moody GB, Mark RG. A new method for detecting atrial fibrillation using R-R intervals. *Comput Cardiol* 1983;**10**:227–230.
22. Goldberger A, Amaral L, Glass L, Hausdorff JM, Ivanov PC, Mark RG, *et al.* Physiobank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 2000;**101**:e215–e220.
23. Asgari S, Mehrnia A, Moussavi M. Automatic detection of atrial fibrillation using stationary wavelet transform and support vector machine. *Comput Biol Med* 2015;**60**:132–142.
24. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983;**148**:839–843.
25. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;**44**:837–845.
26. Hindricks G, Potpara T, Dagres N, Arbelo E, Bax JJ, Blomström-Lundqvist C, *et al.* 2020 ESC guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS): the task force for the diagnosis and management of atrial fibrillation of the European Society of Cardiology (ESC) developed with the special contribution of the European Heart Rhythm Association (EHRA) of the ESC. *Eur Heart J* 2021;**42**:373–498.
27. Uittenbogaart SB, Lucassen WAM, van Etten-Jamaludin FS, de Groot JR, van Weert HCPM. Burden of atrial high-rate episodes and risk of stroke: a systematic review. *Europace* 2018;**20**:1420–1427.
28. Wachter R, Stahrenberg R, Gröschel K. Subclinical atrial fibrillation and the risk of stroke. *N Engl J Med* 2012;**366**:1350–1351.
29. Han X, Hu Y, Foschini L, Chinitz L, Jankelson L, Ranganath R. Deep learning models for electrocardiograms are susceptible to adversarial attack. *Nat Med* 2020;**26**:360–363.
30. Lin D, Xiong J, Liu C, *et al.* Application of comprehensive artificial intelligence retinal expert (CARE) system: a national real-world evidence study. *Lancet Digit Health* 2021;**3**:e486-95.
31. Khan AA, Boriani G, Lip GYH. Are atrial high rate episodes (AHREs) a precursor to atrial fibrillation? *Clin Res Cardiol* 2020;**109**:409–416.