Contents lists available at ScienceDirect

IRBM

journal homepage: www.elsevier.com/locate/irbm

Elsevier Masson France
EM|consulte
www.em-consulte.com

Original Article

Prediction of Atrial Fibrillation from Sinus-Rhythm Electrocardiograms Based on Deep Neural Networks: Analysis of Time Intervals and Longitudinal Study

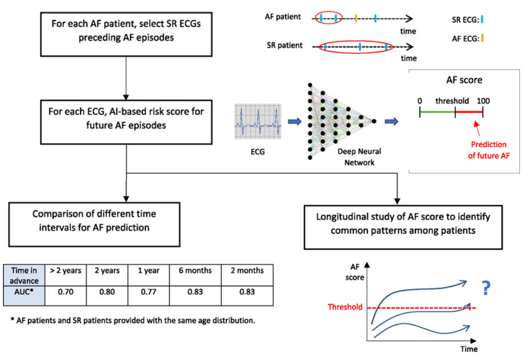
Pietro Melzi [a,*], Ruben Vera-Rodriguez [a], Ruben Tolosana [a], Ancor Sanz-Garcia [b], Alberto Cecconi [b], Guillermo J. Ortega [b,1], Luis Jesus Jimenez-Borreguero [b,c,1]

[a] *Biometric and Data Pattern Analytics Lab, Universidad Autonoma de Madrid, C/ Francisco Tomas y Valiente 11, Labs. C109, 28049 Madrid, Spain*
[b] *Instituto de Investigacion Sanitaria del Hospital Universitario de la Princesa, Calle de Diego de León, 62, 28006 Madrid, Spain*
[c] *CIBERCV, Centro de Investigacion Biomedica en Red Enfermedades Cardiovasculares, Av. Monforte de Lemos, 3-5. Pabellón 11. Planta 0, 28029 Madrid, Spain*

HIGHLIGHTS

- AI predicts future episodes of atrial fibrillation from sinus-rhythm 12-lead ECGs.
- The effect of patients' age and time windows on neural networks predicting AF.
- Longitudinal analysis reveals common patterns in the development of AF.

GRAPHICAL ABSTRACT


For each AF patient, select SR ECGs preceding AF episodes
AF patient
time
SR patient
time
SR ECG:
AF ECG:
For each ECG, AI-based risk score for future AF episodes
ECG
Deep Neural Network
AF score
0
threshold
100
Prediction of future AF
Comparison of different time intervals for AF prediction
Longitudinal study of AF score to identify common patterns among patients
AF score
Threshold
?
Time

Time in advance	> 2 years	2 years	1 year	6 months	2 months
AUC*	0.70	0.80	0.77	0.83	0.83

* AF patients and SR patients provided with the same age distribution.

ARTICLE INFO

Article history:
Received 12 January 2023
Received in revised form 20 October 2023
Accepted 22 October 2023
Available online 27 October 2023

Keywords:
Atrial fibrillation
Healthcare
Artificial intelligence
Deep learning
ECG

ABSTRACT

Objective: Artificial Intelligence (AI) in electrocardiogram (ECG) analysis helps to identify persons at risk of developing atrial fibrillation (AF) and reduces the risk for severe complications. Our aim is to investigate the performance of AI-based methods predicting future AF from sinus rhythm (SR) ECGs, according to different characteristics of patients, time intervals for prediction, and longitudinal measures.
Methods: We designed a retrospective, prognostic study to predict AF occurrence in patients from 12-lead SR ECGs. We classified patients in two groups, according to their ECGs: 3,761 developed AF and 22,896 presented only SR ECGs. We assessed the impact of age on the overall performance of deep neural network (DNN)-based systems, which consist in a variation of Residual Networks for time series. Then, we analysed how much in advance our system can predict AF from SR ECGs and the performance for different categories of patients with AUC and other metrics.
Results: After balancing the age distribution between the two groups of patients, our model achieves AUC of 0.79 (0.72-0.86) without additional constraints, 0.83 (0.76-0.89) for ECGs recorded in the last six months before AF, and 0.87 (0.81-0.93) for patients with stable AF risk measures over time, with sensitivity of 90.62% (80.70-96.48) and diagnostic odd ratio of 20.49 (8.56-49.09).

Abbreviations: AI, Artificial Intelligence; ECG, Electrocardiogram; AF, Atrial Fibrillation; SR, Sinus Rhythm; DNN, Deep Neural Network; AUC, Area Under the Curve; DOR, Diagnostic Odd Ratio; CI, Confidence Interval; IQR, Interquartile Range.

[*] Corresponding author at: Escuela Politecnica Superior, Calle Francisco Tomas y Valiente, 11, C-109-bis, 28049, Madrid, Spain.
E-mail address: pietro.melzi@uam.es (P. Melzi).

[1] These authors have contributed equally.

https://doi.org/10.1016/j.irbm.2023.100811

Conclusion: This study shows the ability of DNNs to predict new onsets of AF from SR ECGs, with the best performance achieved for patients with stable AF risk score over time. The introduction of this time-based score opens new possibilities for AF prediction, thanks to the analysis of long-span time intervals and score stability.

1. Introduction

Atrial fibrillation (AF) is one of the most common sustained arrhythmias, increasing the risk of strokes [1], heart failure [2], and other heart-related complications [3]. AF is often asymptomatic and can be undiagnosed until a first manifestation of stroke [4,5]. Identifying individuals at risk of developing AF and providing appropriate treatment can reduce mortality and strokes, as well as cut healthcare costs [6].

Electrocardiograms (ECGs) hold meaningful information that can significantly aid in estimating AF risk [7]. Over the years, many techniques for predicting AF have emerged, based on ECG analysis. Traditional techniques involve the processing of discrete ECG features with machine learning classifiers [8,9]. These methods only rely on partial ECG information and require a time-consuming phase, also prone to errors, to handcraft features [10].

In contrast, deep neural networks (DNNs) autonomously learn representative and robust features from ECG signals, to replace or complement traditional handcrafted features. DNNs have demonstrated their ability to detect subtle abnormalities derived from structural derangements in 12-lead ECGs [11]. Despite DNN-based ECG analysis has been explored in the literature to assess the risk of future AF, providing important results [12,13] and outperforming traditional techniques, DNNs remain mostly black-box tools. Their integration into medical diagnosis, planning, and control requires a higher level of trust in the machine's capabilities [14].

In this study, we propose a new DNN-based system designed to process 12-lead raw ECGs recorded during normal sinus rhythm (SR) with the aim of predicting patients at risk of developing AF. Unlike previous works, we consider a variety of conditions for patients and ECGs, analysing our system capabilities as described in the following.

Firstly, we observe that AF prevalence in adults is closely linked to age, with numerous studies confirming the increase in incidence among elderly individuals [15–17]. In a previous study, we identified significant variations of performance when evaluating our system with ECGs from patients in diverse age groups [18]. In this study, we focus on disparities in age distribution between patients who developed AF and patients who did not. Specifically, we determine if AF prediction is feasible when age is not a distinguishing factor between the two groups of patients.

For patients who developed AF, we also hypothesise that AF prediction is more accurate when we consider ECGs collected in proximity to the day when AF was initially detected. ECGs recorded too early might not exhibit any signs of AF, while ECGs collected in the days immediately before AF may pertain to hospitalised patients who also suffered other diseases or underwent surgery. In the case of hospitalised patients, AF may be induced by causes beyond the scope of this work, for instance coronary artery bypass grafting surgery [19].

Finally, for every patient who developed AF, we conduct a longitudinal study based on their sequences of ECGs recorded over time. By considering the AF prediction computed for each ECG, we categorise these patients into five distinct groups, suggesting different diagnostic, prognostic, or therapeutic implications for each category [20]. This approach may reduce the occurrence of false predictions, whose consequences include the prescription of unnecessary treatments, with possible side effects.

In summary, this study analyses the ability of our system to predict AF according to various factors, such as the age distribution of patients and the time span between the analysed ECG and their respective onsets of AF, up to eight years in advance. Furthermore, we conduct a longitudinal study for patients who developed AF, to identify common patterns in AF development and recommend tailored treatment approaches. To the best of our knowledge, this study represents the first analysis of the temporal evolution of AF, quantified with a numerical score provided by a DNN-based model. Valuable findings in this area are expected to lead to the implementation of new strategies for AF screening in population.

2. Related works

The limited availability of suitable public databases hinders the development of DNN systems for predicting AF from ECGs recorded during SR. Public ECG databases have a small patient dataset [21], or focus on ECG classification into rhythm classes [22,23].

A review of publications from the past decade has been published in 2022 [24] focusing on AF episode prediction, detection, and classification using wavelets and artificial intelligence (AI). We observe a scarcity of studies on AF prediction, with only one utilising DNNs [25]. However, such study was conducted with a small set of 139 ECG samples, from which 30-second segments were extracted and randomly divided for training and testing.

In the last years, valuable studies made significant advancements in the field employing large private ECG databases to train DNNs. These studies demonstrated that using AI it is possible to diagnose the signs of AF at an early stage and predict the onset of AF attacks [12,13,26].

As stated by authors, the first DNN-based study to predict AF from ECGs recorded during SR employed a database composed of 454,789 ECGs collected at the Mayo Clinic ECG laboratory (USA) between 1993 and 2017. Their DNN model was trained to predict new onsets of AF within 31 days and provided AUC of 0.87 during evaluation [12]. Subsequently, other DNN-based systems were trained with a vast database of 1.6 M 12-lead ECGs from 430,000 patients to predict new onsets of AF within one year, providing AUC of 0.85 [13]. However, compared to our study, these two works predicted AF in a period closer to its occurrence, and did not provide any insights into the features learnt by the network. This could be a drawback in clinical practice, where explainability is essential for healthcare professionals. Finally, these studies did not address the age distribution differences between healthy and unhealthy patients in their test sets.

Recently, a random forest was trained to predict the risk of developing AF within five years using a database composed of 1.1 M 12-lead ECGs from 415,389 patients, achieving the highest AUC of 0.91 [26]. However, it is important to note that the two groups of healthy and unhealthy patients used to develop the system exhibited a significant difference in age distribution. As demonstrated in our experiments, the age distribution disparity can significantly influence the overall model performance.

The works described are the most relevant in the literature for comparison to our study. In another study, a DNN was used to infer 5-year incident AF risk using 12-lead ECGs from patients under longitudinal primary care at Massachusetts General Hospital [27].

Rather than binary classification, the model proposed in [27] used a loss function accounting for the time distance from an AF event. The study makes a valuable contribution to existing research by introducing a deep learning model that explicitly considers time to AF outcome. However, there are significant differences in the design of the model that preclude direct comparison with the previous systems described above, as acknowledged by the authors. Additionally, in [27] it is shown that the combination of both clinical risk factors and AI-based analysis of ECGs can increase the predictive accuracy. The obtained results, with AUC values between 0.71 and 0.82 for different test sets, provide new evidence that ECG-derived risk estimates are generalizable, with predictive value maintained up to 5 years after an ECG is performed.

Two previous works utilised the same database as our current study. The first one predicted AF development using a Multivariate Logistic Regression system with 33 variables, achieving AUC of 0.80 [7]. This study also demonstrated that age and other variables related to P-wave are correlated with AF risk. In the second study, DNN-based systems were evaluated across various demographic patient groups, also providing graphical representations of the automatically learned features [18]. Our proposed study stands out from previous research in several ways. We deploy DNN-based systems for AF prediction with groups of patients well-balanced in terms of age. We also carry out a comprehensive analysis of time windows and establish strict criteria for selecting ECG data for evaluation. We consider ECGs recorded up to eight years before AF onset, excluding data that may introduce misleading effects, such as ECGs with artifacts or those from hospitalised patients.

Finally, we introduce a novel approach by conducting a longitudinal study that examines the progression of AF scores generated by DNN models. The absence of this longitudinal study is considered a limitation in [26]. Existing longitudinal studies in the literature focus on the cumulative incidence of AF over years, typically categorising patients into various AF risk groups [28,29].

3. Methods

3.1. Study design and participants

This retrospective study utilises an ECG database collected from a large cohort of patients at La Princesa University Hospital (Madrid, Spain) between May 5, 2010 and February 4, 2019. The Clinical Ethics Committee from Hospital La Princesa approved this study with a waiver of obtaining informed consent from patients (Protocol number EC1835). The initial database contains 296,022 12-lead ECGs from 122,394 patients. Each ECG has a sampling frequency of 500 Hz and lasts approximately ten seconds. We follow the study design proposed in [18], summarised below.

We exclude from the study patients with missing age information and patients with only one recorded ECG, not useful for our analysis. The remaining patients are 50,448 (25,558 female), with median age of 69 years (IQRs 53-80) and median number of ECGs equal to 3 (IQRs 2-5). We divide these patients into two groups: *i*) those who developed AF at least once along their clinical history (11,707, *AF patients*), and *ii*) those who have exclusively presented SR ECGs along their clinical history (23,302, *SR patients*). Patients not belonging to any of these two categories (15,439) are discarded.

We use the automatic interpretations of ECG rhythm provided by the recording machine [30] to detect the presence of AF. Acknowledging that automatic AF interpretations may be prone to errors [31], we employ a DNN-based ECG rhythm classifier to validate the AF interpretations made by the recording machine, avoiding the manual inspection of thousands of ECGs. DNN-based classifiers have demonstrated a high level of reliability, achieving AUC of 0.97 in the classification of 12 different ECG rhythms [32]. Inspired by [32], we deployed in-house a DNN-based classifier of ECG rhythms, to distinguish between ECGs presenting SR and AF. To train and evaluate our rhythm classifier, we use a set of ECGs with AF and SR rhythm, manually labelled by cardiologists at La Princesa University Hospital. We evaluate our DNN classifier with a test set of 190 ECG samples, that equally represents the two rhythms. The size of the test set is determined by the availability of ECG samples labelled by cardiologists with AF rhythm, and results from the traditional split of 70%, 15%, 15% between training, validation, and testing data. Our DNN classifier achieves AUC of 0.9986, with 95% CI between 0.9931 and 1.0000, providing strong evidence of the evaluation's robustness.

For each patient who developed AF, we define the AF index date as the date in which the first episode of AF occurred. We analyse 11,707 ECGs at AF index dates (i.e., one for each AF patient), and discard 3,066 patients (26%) whose ECG rhythm is not classified as AF by our DNN-based classifier. This approach ensures that we do not incorporate ECGs with potentially incorrect labels in our subsequent experiments, preventing misleading effects in our analysis. On the other side, we trust the automatic SR interpretations provided by the recording machine, and do not double-check the interpretation of SR data. In fact, a manual check performed at La Princesa University Hospital revealed that only two of 800 (0.25%) ECGs with automatic SR interpretation were inaccurately labelled [7].

Hereinafter, for both AF and SR patients, we only consider ECGs that present an automatic interpretation of "sinus rhythm", according to our goal of predicting AF from ECGs that exhibit no prior evidence of AF. Among the 58,949 ECGs belonging to AF patients whose ECG at AF index was positively checked (i.e., 8,641 patients), only 15,387 ECGs present an interpretation of "sinus rhythm" and can be used in the experiments.

For each patient in the SR group, we define the SR index date as the date of their last recorded ECG. In both groups we apply the following exclusion criteria: ECGs with age < 18 years, with extrasystoles, with atrial to ventricular ratio > 2 or $< 1/2$, with an average number of P waves per QRS complex $\neq 1$, and with number of QRS complexes in the rhythm group higher than the average number of P waves per QRS complex. These rules are set to avoid atrial oversensing secondary to artifacts and to increase the specificity of SR diagnosis, according to a previous work [7]. As a result, we obtain: *i*) the *AF group* composed of 12,198 SR ECGs from 3,761 AF patients, and *ii*) the *SR group* composed of 64,829 SR ECGs from 22,896 SR patients. A concise summary of the performed operations is reported in Fig. 1.

3.2. Time windows

Time windows play a crucial role in our study, defining for each patient the temporal interval containing the ECGs of interest, according to their AF or SR index date. We consider multiple consecutive time windows to evaluate our model. Their union results in a larger time window that includes, for each AF patient, all the SR ECGs recorded at least one week before the AF index date. This constraint is inspired by a previous work [7], and diverges from other studies [12,13,26]. Thisensures that we do not incorporate ECGs from hospitalised patients, whose AF might be induced by factors outside the scope of this work. ECGs recorded in the last week before AF index are only included in training sets to increase their size, but they are not used in any evaluation datasets. In total, we consider five intervals for AF prediction, evaluating the performance of our model up to "more than two years" in advance, with ECGs recorded until 8 years before AF.

For each patient in SR group, we set a time window comprising ECGs recorded at least two years before the SR index date. This constraint increases the confidence that SR data employed in the

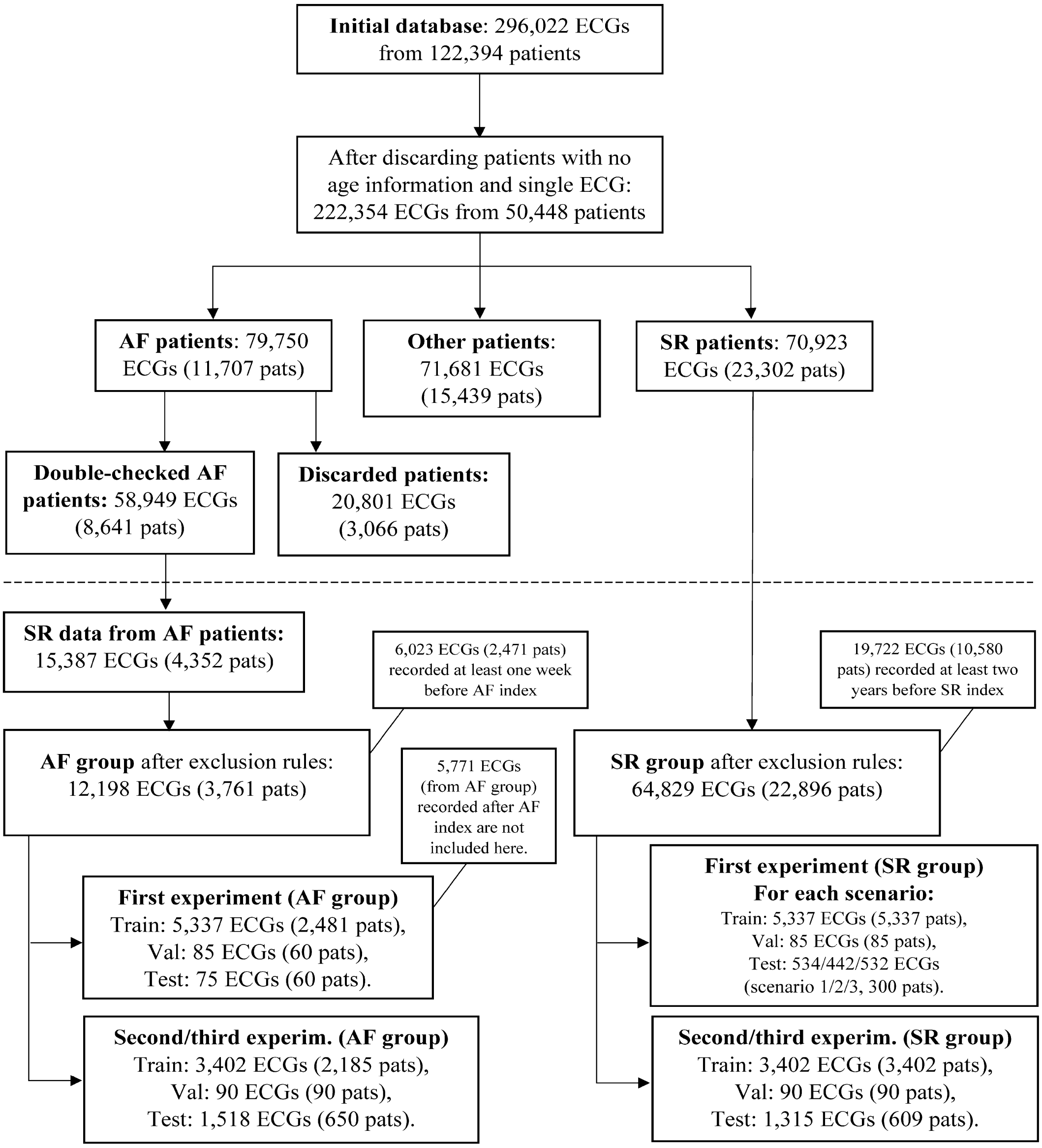

Initial database: 296,022 ECGs from 122,394 patients
After discarding patients with no age information and single ECG: 222,354 ECGs from 50,448 patients
AF patients: 79,750 ECGs (11,707 pats)
Other patients: 71,681 ECGs (15,439 pats)
SR patients: 70,923 ECGs (23,302 pats)
Double-checked AF patients: 58,949 ECGs (8,641 pats)
Discarded patients: 20,801 ECGs (3,066 pats)
SR data from AF patients: 15,387 ECGs (4,352 pats)
6,023 ECGs (2,471 pats) recorded at least one week before AF index
19,722 ECGs (10,580 pats) recorded at least two years before SR index
AF group after exclusion rules: 12,198 ECGs (3,761 pats)
5,771 ECGs (from AF group) recorded after AF index are not included here.
SR group after exclusion rules: 64,829 ECGs (22,896 pats)
First experiment (AF group) Train: 5,337 ECGs (2,481 pats), Val: 85 ECGs (60 pats), Test: 75 ECGs (60 pats).
First experiment (SR group) For each scenario: Train: 5,337 ECGs (5,337 pats), Val: 85 ECGs (85 pats), Test: 534/442/532 ECGs (scenario 1/2/3, 300 pats).
Second/third experim. (AF group) Train: 3,402 ECGs (2,185 pats), Val: 90 ECGs (90 pats), Test: 1,518 ECGs (650 pats).
Second/third experim. (SR group) Train: 3,402 ECGs (3,402 pats), Val: 90 ECGs (90 pats), Test: 1,315 ECGs (609 pats).

Fig. 1. Diagram flow of patients and ECGs - Data cleaning and creation of experimental datasets. In the first experiment, the number of test patients for the SR group remains the same in the three considered scenarios. In the second experiment, the number of test patients in the AF group is higher compared to the first experiment because ECGs from multiple time intervals need to be tested. AF=Atrial Fibrillation, ECG=Electrocardiogram, SR=Sinus Rhythm.

experiments are isolated from episodes of AF, which may not be part of the initial database. In Fig. 2 we provide an example of time window for patients in AF and SR groups.

3.3. Overview of the AI model

In our experiments we consider a DNN-based model and investigate its ability to predict AF from the different features automatically learned from 12-lead raw ECGs. We have made the model architecture available at https://github.com/eHealthUAM/ECGpredictAF, along with the final weights used in our experiments. The model considered in this study is the one that provides the best results in our previous study [18]. The model processes two-dimensional input signals: one-dimensional 10-second signals from 12 different leads. The bottom layers of our model are obtained from a Residual Network for time series [33] (details in Supplemental Methods) and a fully connected layer that generates four aggregated features. They are concatenated with two numerical values, representing the age and the sex of the patient, and processed by another fully connected layer that outputs the final AF score, constrained in the $[0, 1]$ interval. Our DNN model is trained to provide AF scores near 1 for ECGs belonging to patients who developed AF, and scores near 0 for ECGs from patients with exclusively ECGs recorded during SR.

3.4. Experiments

In the first experiment, we assess the model performance by training and evaluating it using datasets with distinct age distributions. The goal of this experiment is to analyse the effect of different age distributions on AF prediction. We predict AF in the smallest time window of the study, that contains ECGs recorded between two months and one week before the first AF episode, and consider three different scenarios, in which we keep the same data for the AF group and constantly change data for the SR group. In the first scenario we randomly sample SR data, in the second scenario we sample SR data to match the age distributions of AF patients, and in the third scenario we consider SR patients that are in average younger than all the previously considered sets of patients. We verify that the age distributions of AF and SR pa-

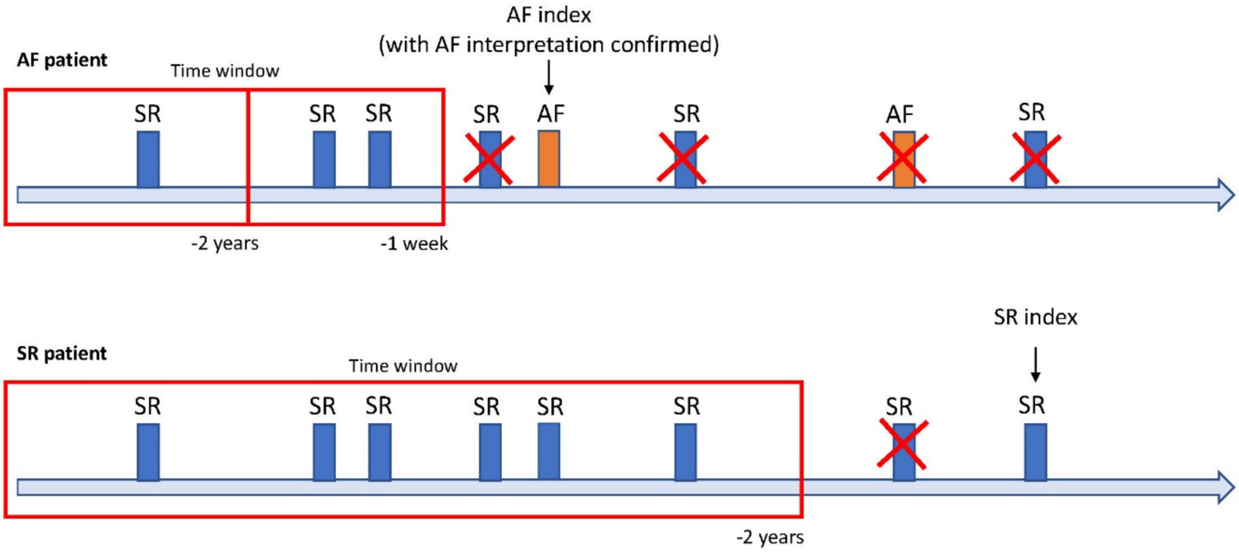


Fig. 2. Time windows for AF patients and SR patients – Representation of ECG included and excluded from time window for patients in AF group (above) and SR group (below).

Table 1
Details on the test sets for Experiment 2. Number of patients, ECGs, and median age for different time windows for AF data. For each time window, the same SR group is considered. AF=Atrial Fibrillation, SR=Sinus Rhythm.

Test subsets	Patients	ECGs	Age (yr)
AF, more than 2 years	504	939	79 (71-84)
AF, 1 year to 2 years	208	294	82 (75-87)
AF, 6 months to 1 year	88	117	83 (76-88)
AF, 2 months to 6 months	64	85	82 (73.75-87)
AF, 1 week to 2 months	58	83	82.5 (74-86.75)
SR group	609	1315	77 (76-81)

Table 2
Details on the test sets for Experiment 3. Number of patients, ECGs, and median age for different time windows for AF data. For each time window, the same SR group is considered (patients with a single ECG recorded at least two years before SR index). ECGs in AF groups are recorded in the last two years before AF episodes. AF=Atrial Fibrillation, SR=Sinus Rhythm.

Test subsets	Patients	ECGs	Age (yr)
AF, Single ECG	83	83	82 (71.5-87.5)
Distant ECGs, stable AF score	64	96	82.75 (75.75-88)
Distant ECGs, unstable AF score	43	65	82 (75.5-86.75)
Close ECGs, stable AF score	38	101	83.5 (75.8-87)
Close ECGs, unstable AF score	65	234	82.25 (74.5-87)
SR group	312	312	77 (76-82)

tients considerably affect the model performance. In subsequent experiments, we exclusively work with datasets characterized by balanced age distributions between the two patient groups, to explore factors other than patient age [7].

The second experiment assesses the model performance in predicting AF risk using ECGs from five consecutive time windows, covering a period up to one week before the occurrence of AF. The goal is to analyse the prediction performance from more than 2 years up to one week before AF. We train our model with an age-balanced dataset that equally represents the different time windows. Each time window is evaluated with a specific AF group and the same SR group, to obtain a fair comparison of results. The considered AF and SR groups present identical age distributions. The number of patients and ECGs considered for each group in the second experiment are reported in Table 1. AF patients may appear in the test sets of multiple time windows, leading to a cumulative number of AF patients higher than the number reported in Fig. 1 (i.e., 650 patients).

In the last experiment, we focus on the evolution of AF history in patients. ECGs are a convenient, cost-effective, non-invasive, and accessible diagnostic tool, often included in routine patient examinations. Consequently, patients typically undergo multiple ECGs over time. Our model provides a measure of AF risk (AF score) for each ECG sample, enabling us to perform a longitudinal study.

We define five categories of patients presenting the following characteristics: i) only one ECG, ii) at least three months between consecutive ECGs (“distant ECGs”) and stable AF score, iii) at least three months between consecutive ECGs and unstable AF score, iv) less than three months between consecutive ECGs (“close ECGs”) and stable AF score, and v) less than three months between consecutive ECGs and unstable AF score. The number of patients, ECGs, and the median age of patients for each AF and SR category are reported in Table 2 (only ECGs recorded in the last two years before AF are considered here). We define the AF score stable for a specific patient when the gap between their maximum and minimum AF scores is 0.15 or less. We select this threshold as suitable to represent the concept of stability. Our patient categories aim at providing clinical value by identifying patients that may be already under monitoring or treatment (because of the short time between consecutive ECGs) and patients at various stages of the disease, based on the evolution of their AF scores. We evaluate and compare the model performance (already trained during the second experiment) across the five categories of patients.

We compare different test sets of AF group with the subset of SR patients with a single ECG in the test set (i.e., a single ECG recorded at least two years before SR index), which is expected for the average healthy population.

Further details regarding the experimental protocol of the three experiments can be found in Supplemental Material.

3.5. Statistical analysis

To train our model we consider categorical cross entropy as loss function, with Adam optimizer and initial learning rate of 0.001. At each epoch we evaluate the loss function on the validation set. We halve the learning rate if the function does not decrease for two consecutive epochs. We stop the training if the function does not decrease for six consecutive epochs.

For each evaluation we compute AUC, sensitivity, specificity, and diagnostic odd ratio (DOR) with confidence intervals (CIs) of

Table 3
Experiment 1: median ages and AUC achieved in the three different scenarios considered. We provide the median age of ECGs in the training and validation sets, and the median age of patients in the test set. Between brackets, interquartile ranges for median ages and 95% confidence intervals for AUC are reported. In "AF data" row, the median ages of the datasets combined in each scenario with SR data are reported. The results reported in the "AUC" column show how the model performance changes according to the set of SR data considered in the different scenarios. Sensitivity and specificity are computed with a threshold of 0.55. AF=Atrial Fibrillation, AUC=Area Under the Curve, SR=Sinus Rhythm.

Scenarios	Age in training set (yr)	Age in validation set (yr)	Age in test set (yr)	AUC	Sensitivity	Specificity
1. Random sampling (SR group)	63 (50-75)	66 (54-77)	64 (50-75)	0.89 (0.83-0.95)	82% (70-90)	79% (74-83)
2. Balancing distributions (SR group)	75 (67-80)	77 (72-87)	82.5 (73-88)	0.79 (0.72-0.86)	78% (66-88)	68% (62-73)
3. Young SR group	53 (43-60)	53 (45-59)	53 (43.38-60)	0.98 (0.95-1.00)	95% (86-99)	95% (92-97)
AF group (same in 1., 2., 3.)	79 (70-84)	77 (72-87)	82.5 (73-88)	-	-	-

95%. We calculated 2-sided P values from Z score to evaluate if the difference between AUCs is statistically significant ($P < 0.05$). The size of test sets in the different experiments is limited by the amount of data dedicated to training. This has been established empirically after many trials in which it was found beneficial to increase the training set as much as possible to obtain better model performance.

4. Results

Patients in AF group present a median age of 80 years (IQRs 70-86) at their AF index date and 49.61% of them are females. Patients in SR group present a median age of 64 years (IQRs 50-77) at their SR index date and 52.96% of them are females. We report age statistics that refer to index dates, to consider unique ages for patients with multiple ECGs recorded through years. However, we consider in the experiments exact ages of patients at the date of ECG recordings.

If we only consider the SR ECGs in AF group that are recorded at least one week before the first event of AF, 6,023 ECGs in total, the median distance from the AF index date 719 days (IQRs 305-1,370.5). If we only consider the SR ECGs in SR group that are recorded at least two years before their SR index dates, 19,722 in total, the median distance from the SR index date is 1,453 days (IQRs 1,036-1,987).

When computing AUC in the experiments, for each patient in the test set we average the AF scores related to their ECGs and provide a unique score representing their own risk of developing AF. While our model has been trained with the same number of ECGs belonging to AF and SR groups, in all the experiments we can observe that sensitivity is generally higher than specificity. Hence, this means it is easier to recognise patients who will develop AF (i.e., to identify those SR ECGs that anticipate AF development) compared to patients who will not develop it. This is favourable to our goal of predicting AF, although this behaviour may change according to the threshold set or training settings.

4.1. Experiment 1: effect of age in the AF and SR groups

In the first experiment, we predict AF in the smallest time window of the study, that contains ECGs recorded between two months and one week before the first AF episode. Given the limited number of ECGs recorded in this time window (412 in total), we select two small sets for validation and testing (Fig. 1), each one representing approximately 20% of ECGs in the time window. We include in the training set the remaining 252 ECGs (around 60%) recorded from two months to one week before AF, to represent as best as possible this specific time interval during training. Also, in experimental trials we have found beneficial to increase the size of the training dataset with ECGs recorded in other time windows that precede the first AF episode.

Median ages and performance achieved by our model are provided in Table 3 for the three different scenarios considered. The results achieved in the first scenario are consistent with those of previous studies [12,13,26]. However, unlike them we prevent the evaluation of performance with data recorded in the last week before AF, as the causes of AF may be out of the scope of this work [34].

We observe that if we increase the difference between the age distributions of AF and SR patients, we also increase the model performance. Even if additional ageing-related information can be extracted from ECGs and improve the performance of the model, with this experiment we simply aim to emphasise the relationship between age distributions and model performance. In general, related works do not report the difference in the distributions of patients age [12,13] or do not mitigate this aspect during data selection [26]. This aspect should be always addressed because, as we show in the experiment, it is very easy to boost the model performance by selecting opportune sets of patients. To assess if other-than-age information can be exploited to predict AF, we perform the next experiments with age-balanced datasets. To support our findings, we calculate the P value of the AUC difference between the common scenario of random sampling (scenario 1) and the proposed scenario of age-balancing (scenario 2). For the AUC difference of 0.1 between the two scenarios, we obtain a statistically significant P value = 0.03, with standard error = 0.05.

4.2. Experiment 2: effect of time windows

The results obtained in the analysis of the different time windows are reported in Table 4, where the same model was tested with different age-balanced datasets. Hence, these results are hardly comparable with other studies that do not consider age balance. The goal is to show the different performances achieved with different time windows for the AF group. Our model provides similar performances when considering ECGs recorded in the last two years before AF, with increment of AUC when time windows refer to a period closer to the AF event. Also, it is very interesting to see that even in the furthest time window the model can predict future AF with an AUC of 0.70 (0.67-0.73). Sensitivity and DOR provide an equal trend, as the values of specificity are constant across the different time windows.

4.3. Experiment 3: categories of patients based on longitudinal measures

During evaluation we do not consider ECGs recorded more than two years before AF: as we observed in Experiment 2, they pro-

Table 4
Experiment 2: performance obtained for the different time windows considered. 95% confidence intervals are reported in brackets for all the metrics. The results reported in the "AUC" column show how the model performance decreases with the increase of the distance from the day of AF event. For each time window we consider the same set of ECGs for the SR group, always obtaining the same specificity of 66.83% (62.93-70.56). AUC=Area Under the Curve, DOR=Diagnostic Odd Ratio.

Time windows	AUC	Sensitivity	DOR
More than 2 years	0.70 (0.67-0.73)	62.50% (58.11-66.74)	3.36 (2.62-4.30)
1 year to 2 years	0.80 (0.76-0.84)	79.81% (73.70-85.04)	7.96 (5.46-11.62)
6 months to 1 year	0.77 (0.71-0.83)	68.18% (57.39-77.71)	4.32 (2.67-6.97)
2 months to 6 months	0.83 (0.76-0.89)	84.38% (73.14-92.24)	10.88 (5.43-21.81)
1 week to 2 months	0.83 (0.76-0.90)	82.76% (70.57-91.41)	9.67 (4.79-19.51)

Table 5
Experiment 3: performances obtained for the different categories of patients described in detail in Section 4.3. 95% confidence intervals are reported in brackets for all the metrics. The results reported in the "AUC" column show how the model performance is higher when predicting AF for patients who provide a stable AF score over time. For each category we consider the same set of ECGs for the SR group, always obtaining the same specificity of 67.95% (62.46-73.09). AF=Atrial Fibrillation, AUC=Area Under the Curve, DOR=Diagnostic Odd Ratio, ECG=Electrocardiogram.

Categories	AUC	Sensitivity	DOR
Single ECG	0.79 (0.73-0.85)	78.31% (67.91-86.61)	7.66 (4.31-13.59)
Distant ECGs, stable AF score	0.87 (0.81-0.93)	90.62% (80.70-96.48)	20.49 (8.56-49.09)
Distant ECGs, unstable AF score	0.77 (0.68-0.85)	74.42% (58.83-86.48)	6.17 (2.99-12.73)
Close ECGs, stable AF score	0.85 (0.77-0.93)	84.21% (68.75-93.98)	11.31 (4.58-27.92)
Close ECGs, unstable AF score	0.78 (0.71-0.85)	72.31% (59.81-82.69)	5.54 (3.06-10.02)

vide less accurate information compared to ECGs closer to AF onset (note that the sum of ECGs from AF patients in Table 1, recorded in the last two years between AF, is equal to the sum of ECGs from AF patients in Table 2). The results obtained in the final test of our model are reported in Table 5. The best performance is obtained in the categories of patients with stable AF score, with sensitivity of 90.62% (80.70%-96.48%) and DOR of 20.49 (8.56-49.09) for patients with distant ECGs.

In Supplemental Appendix, we show the evolution of AF score for some patients with at least four ECGs and belonging to the two categories of patients with distant ECGs, which are more likely patients under regular monitoring (stable AF score in Supplemental Figures 3–6 and unstable AF score in Supplemental Figures 7–10). We observe the sequences of AF scores in the last two years before AF and provide summary statistics in Supplemental Results.

In the category of patients with stable AF score, the minimum scores of each patient are in average higher and settled in high values. Hence, we can increase the threshold that identifies AF to reduce the number of misclassified patients. In the category of patients with unstable AF score, the maximum scores of each patient are in average lower and AF score tends to increase over time. For these patients it is more difficult to detect AF and the increase of AF score is a signal to watch out. For the first time, our results showed the relevance of assessing an AF score over time as stable high-risk patients have a higher risk of developing AF, according to the AF score.

5. Discussion

The study underscores the significance of age distributions in influencing model performance, noting that predicting AF is significantly easier when the healthy patients are younger than the unhealthy ones. We balance the age distributions to force our model to rely on features other than age to predict AF. Despite this setting leads to a decrease in performance, we obtain high sensitivity values for specific subsets of patients.

We confirm the effectiveness of DNN-based AF prediction and demonstrate the utility of complementing the monitoring of patients over time with a trusted implementation of AF score. Our analysis of time windows reveals that AF score performs better within the two years preceding the AF event, with further improvement in the last six months. In addition, AF becomes observable in some patients only after a certain period. The growing trend of AF score may facilitate AF prediction while the score remains below a certain threshold.

AF is associated to several conditions which may produce secondary ECG changes. These changes may be transient or persistent depending on the underlying clinical situation. Specifically, transient ECG changes might be related to electrolytes disorders, acute diseases, or intercurrent drugs use whereas persistent ECG changes might be related to structural heart diseases, conduction disorders, or chronic drug use. Therefore, patients with a stable high-risk score over time might have a persistent clinical disorder which present a higher risk of AF development. Further studies are needed to clarify this hypothesis.

AI-based models also remove the clinicians' subjective, error-prone interpretations and provide an objective prediction within seconds [35].

Our system is promising and easy to implement, although it is necessary to overcome some limitations before employing it in a real scenario. First, a better understanding of the features that DNNs automatically learn in the task may encourage the acceptance of such decision-making systems in practice. In this sense, the graphical representation of features provided in [18] provides a foundation for exploration and further investigation. Secondly, the integration of additional clinical data may improve the prognostic value of our proposed method. For instance, without precise clinical information, we can only define heuristic rules for data cleaning and time windows. Finally, we consider the limitations related to the creation of AF and SR groups, common to similar studies. Paroxysmal AF is the initial form of AF and in up to one-third of patients it may be silent. This means that some patients included in the AF group may have experienced asymptomatic AF episodes before the diagnostic ECG recording. Data contained in the SR group are chosen to represent the condition of ordinary population, even if they may include data from patients with undetected or asymptomatic episodes of paroxysmal AF. Hopefully, the increasing interest for wearables with the capability of long-term ECG monitoring may overcome this limitation, however, their use in general population is still reduced.

For future work, multiple age-specific models can be trained for specific age intervals when a larger database is available. The prediction of AF would be computed by an age-specific model that exploits features learned for the appropriate age interval of interest. Also, the proposed score should be prospectively tested in a clinical trial before implementing it into the clinical practice.

6. Conclusion

This study demonstrates the ability of DNNs in predicting future AF from SR ECGs, with higher predictive capability in the two years preceding AF occurrence, and for patients with stable AF scores over time. Our proposed approach offers an affordable and accessible technique to assess the risk of developing AF in the general population. It may also help cardiologists to consider specific treatments for different categories of patients. ECGs can be easily integrated into routine check-ups, making the proposed technique suitable for a screening strategy, and potentially opening new possibilities for AF early detection.

Human and animal rights

The authors declare that the work described has not involved experimentation on humans or animals.

Funding

This work has been supported by: Carlos III Institute of Health under the health Strategy action 2020-2022 with reference PI20/00792 and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860813 – TReSPAsS-ETNTRESPASS-ETN (H2020-MSCAITN-2019-860813).

Author contributions

All authors attest that they meet the current International Committee of Medical Journal Editors (ICMJE) criteria for Authorship.

CRediT authorship contribution statement

PM, RV-R, RT, AS-G, AC, GJO, and LJJ-B designed the study. PM, RV-R, and RT performed the experiments. AS-G, AC, GJO, and LJJ-B provided guidance and expertise in the medical field. PM, AS-G, and GJO verified the underlying data. PM, RV-R, and RT wrote the manuscript. PM, RV-R, RT, AS-G, AC, GJO, and LJJ-B critically reviewed the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial or personal relationships that could be viewed as influencing the work reported in this paper.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Use of generative AI and AI-assisted technologies

Nothing to disclose.

Acknowledgement

This work has been supported by: Carlos III Institute of Health under the health Strategy action 2020-2022 with reference PI20/00792 and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860813 – TReSPAsS-ETNTRESPASS-ETN (H2020-MSCAITN-2019-860813). JO, AS-G, LJJ-B received a research grant from the Carlos III Institute of Health under the health Strategy action 2020-2022 with reference PI20/00792. This study is also supported partially by projects TRESPASS-ETN (H2020-MSCAITN-2019-860813), PRIMA (H2020-MSCA-ITN-2019-860315), IDEA-FAST (IMI2-2018-15-853981), INTER-ACTION (PID2021-126521OB-I00 MICINN/FEDER). Funding sources were not involved in any phase of the study.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.irbm.2023.100811.

References

[1] Jørgensen HS, Nakayama H, Reith J, Raaschou HO, Olsen TS. Acute stroke with atrial fibrillation. Stroke 1996;27:1765–9. https://doi.org/10.1161/01.STR.27.10.1765.

[2] Daubert JC. Introduction to atrial fibrillation and heart failure: a mutually noxious association. EP Europace 2003;5:S1–4. https://doi.org/10.1016/j.eupc.2004.07.002.

[3] Alpert JS, Petersen P, Godtfredsen J. Atrial fibrillation: natural history, complications, and management. Annu Rev Med 1988;39:41–52. https://doi.org/10.1146/annurev.me.39.020188.000353.

[4] Turakhia MP, Shafrin J, Bognar K, Trocio J, Abdulsattar Y, Wiederkehr D, et al. Estimated prevalence of undiagnosed atrial fibrillation in the United States. PLoS ONE 2018;13:1–11. https://doi.org/10.1371/journal.pone.0195088.

[5] Sandhu RK, Healey JS. Screening for undiagnosed atrial fibrillation. Expert Rev Cardiovasc Ther 2018;16:591–8. https://doi.org/10.1080/14779072.2018.1496018.

[6] Wolf P, Mitchell J, Baker C, Kannel W, D'Agostino R. Impact of atrial fibrillation on mortality, stroke, and medical costs. Arch Intern Med 1998;158:229–34. https://doi.org/10.1001/archinte.158.3.229.

[7] Sanz-García A, Cecconi A, Vera A, Moreno JMC, Alfonso F, Ortega GJ, et al. Electrocardiographic biomarkers to predict atrial fibrillation in sinus rhythm ECGs. Heart BMJ J 2021. Published online first.

[8] Couceiro R, Carvalho P, Henriques J, Antunes M, Harris M, Habetha J. Detection of atrial fibrillation using model-based ECG analysis. In: 2008 19th int. conf. pattern recognit.; 2008. p. 1–5.

[9] Zabihi M, Rad AB, Katsaggelos AK, Kiranyaz S, Narkilahti S, Gabbouj M. Detection of atrial fibrillation in ECG hand-held devices using a random forest classifier. In: 2017 comput. cardiol. CinC; 2017. p. 1–4.

[10] German DM, Kabir MM, Dewland TA, Henrikson CA, Tereshchenko LG. Atrial fibrillation predictors: importance of the electrocardiogram. Ann Noninvasive Electrocardiol 2016;21:20–9. https://doi.org/10.1111/anec.12321.

[11] Attia Z, Kapa S, Lopez-Jimenez F, McKie P, Ladewig D, Satam G, et al. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nat Med 2019;25:70–4. https://doi.org/10.1038/s41591-018-0240-2.

[12] Attia Z, Noseworthy P, Lopez-Jimenez F, Asirvatham S, Deshmukh A, Gersh B, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. Lancet 2019;394:861–7. https://doi.org/10.1016/S0140-6736(19)31721-0.

[13] Raghunath S, Pfeifer JM, Ulloa-Cerna AE, Nemani A, Carbonati T, Jing L, et al. Deep neural networks can predict new-onset atrial fibrillation from the 12-lead ECG and help identify those at risk of atrial fibrillation-related stroke. Circulation 2021;143:1287–98. https://doi.org/10.1161/CIRCULATIONAHA.120.047829.

[14] Chakraborty S, Tomsett R, Raghavendra R, Harborne D, Alzantot M, Cerutti F, et al. Interpretability of deep learning models: a survey of results. In: 2017 IEEE SmartWorld ubiquitous intell. comput. adv. trust. comput. scalable comput. commun. cloud big data comput. Internet people smart city innov. SmartWorldSCALCOMUICATCCBDComIOPSCI; 2017. p. 1–6.

[15] Go A, Hylek E, Phillips K, Chang Y, Henault L, Selby J, et al. Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the Anticoagulation and Risk Factors in Atrial Fibrillation (ATRIA) study. JAMA 2001;285:2370–5. https://doi.org/10.1001/jama.285.18.2370.

[16] Kannel WB, Wolf PA, Benjamin EJ, Levy D. Prevalence, incidence, prognosis, and predisposing conditions for atrial fibrillation: population-based estimates. Am J Cardiol 1998;82:2N–9N. https://doi.org/10.1016/S0002-9149(98)00583-9.

[17] Heeringa J, van der Kuip DAM, Hofman A, Kors JA, van Herpen G, Stricker BHCh, et al. Prevalence, incidence and lifetime risk of atrial fibrillation: the Rotterdam study. Eur Heart J 2006;27:949–53. https://doi.org/10.1093/eurheartj/ehi825.

[18] Melzi P, Tolosana R, Cecconi A, Sanz-Garcia A, Ortega GJ, Jimenez-Borreguero LJ, et al. Analyzing artificial intelligence systems for the prediction of atrial fibrillation from sinus-rhythm ECGs including demographics and feature visualization. Sci Rep 2021;11:22786. https://doi.org/10.1038/s41598-021-02179-1.

[19] Sovilj S, Oosterom AV, Rajsman G, Magjarevic R. ECG-based prediction of atrial fibrillation development following coronary artery bypass grafting. Physiol Meas 2010;31:663–77. https://doi.org/10.1088/0967-3334/31/5/005.

[20] Cámara A. Characterization of diagnostic tests. Neurol Barc Spain 2004;19:31–8.

[21] Matias I, Garcia N, Pirbhulal S, Felizardo V, Pombo N, Zacarias H, et al. Prediction of atrial fibrillation using artificial intelligence on electrocardiograms: a systematic review. Comput Sci Rev 2021;39:100334. https://doi.org/10.1016/j.cosrev.2020.100334.

[22] Clifford GD, Liu C, Moody B, Lehman LH, Silva I, Li Q, et al. AF classification from a short single lead ECG recording: the PhysioNet/computing in cardiology challenge 2017. In: 2017 comput. cardiol. CinC; 2017. p. 1–4.

[23] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet. Circulation 2000;101:e215–20. https://doi.org/10.1161/01.CIR.101.23.e215.

[24] Serhal H, Abdallah N, Marion J-M, Chauvet P, Oueidat M, Humeau-Heurtier A. Overview on prediction, detection, and classification of atrial fibrillation using wavelets and AI on ECG. Comput Biol Med 2022;142:105168. https://doi.org/10.1016/j.compbiomed.2021.105168.

[25] Erdenebayar U, Kim H, Park J-U, Kang D, Lee K-J. Automatic prediction of atrial fibrillation based on convolutional neural network using a short-term normal electrocardiogram signal. J Korean Med Sci 2019:34. https://doi.org/10.3346/jkms.2019.34.e64.

[26] Biton S, Gendelman S, Ribeiro AH, Miana G, Moreira C, Ribeiro ALP, et al. Atrial fibrillation risk prediction from the 12-lead electrocardiogram using digital biomarkers and deep representation learning. Eur Heart J, Digit Health 2021:ztab071. https://doi.org/10.1093/ehjdh/ztab071.

[27] Khurshid S, Friedman S, Reeder C, Achille PD, Diamant N, Singh P, et al. ECG-based deep learning and clinical risk factors to predict atrial fibrillation. Circulation 2022. https://doi.org/10.1161/CIRCULATIONAHA.121.057480.

[28] Christopoulos G, Graff-Radford J, Lopez CL, Yao X, Attia ZI, Rabinstein AA, et al. Artificial intelligence–electrocardiography to predict incident atrial fibrillation. Circ Arrhythm Electrophysiol 2020;13:e009355. https://doi.org/10.1161/CIRCEP.120.009355.

[29] Staerk L, Wang B, Preis SR, Larson MG, Lubitz SA, Ellinor PT, et al. Lifetime risk of atrial fibrillation according to optimal, borderline, or elevated levels of risk factors: cohort study based on longitudinal data from the Framingham Heart Study. BMJ 2018:k1453. https://doi.org/10.1136/bmj.k1453.

[30] The Philips 12-Lead Algorithm Physician's Guide n.d. http://incenter.medical.philips.com/doclib/enc/fetch/577817/577818/12-Lead_Algorithm_Physician_s_Guide_for_Algorithm_Verion_PH080A%2C_(ENG).pdf%3Fnodeid%3D3325283%26vernum%3D-2. [Accessed 12 February 2021].

[31] Bae MH, Lee JH, Yang DH, Park HS, Cho Y, Chae SC, et al. Erroneous computer electrocardiogram interpretation of atrial fibrillation and its clinical consequences. Clin Cardiol 2012;35:348–53. https://doi.org/10.1002/clc.22000.

[32] Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019;25:65–9. https://doi.org/10.1038/s41591-018-0268-3.

[33] Wang Z, Yan W, Oates T. Time series classification from scratch with deep neural networks: a strong baseline. In: 2017 int. jt. conf. neural netw. IJCNN; 2017. p. 1578–85.

[34] Sovilj S, Rajsman G, Magjarević PR. ECG based prediction of atrial fibrillation using support vector classifier. Automatika 2011;52:58–67. https://doi.org/10.1080/00051144.2011.11828404.

[35] Kashou A, Noseworthy P. Predicting incident atrial fibrillation in sinus rhythm: more than just trusting the 'black box'. Heart 2021;107:1770–1. https://doi.org/10.1136/heartjnl-2021-319385.