European Heart Journal - Digital Health (2022) **3**, 208–217
https://doi.org/10.1093/ehjdh/ztac014

ORIGINAL ARTICLE

Short-term prediction of atrial fibrillation from ambulatory monitoring ECG using a deep neural network

Jagmeet P. Singh [1,]*, Julien Fontanarava[2], Grégoire de Massé[2], Tanner Carbonati[2], Jia Li[2], Christine Henry[2], and Laurent Fiorina[3]

[1]Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114, USA; [2]Cardiologs, 136 rue Saint Denis, 75002 Paris, France; and [3]Ramsay Santé, Institut Cardiovasculaire Paris Sud, Hôpital privé Jacques Cartier, 91300 Massy, France

Received 31 December 2021; revised 16 February 2022; online publish-ahead-of-print 6 April 2022

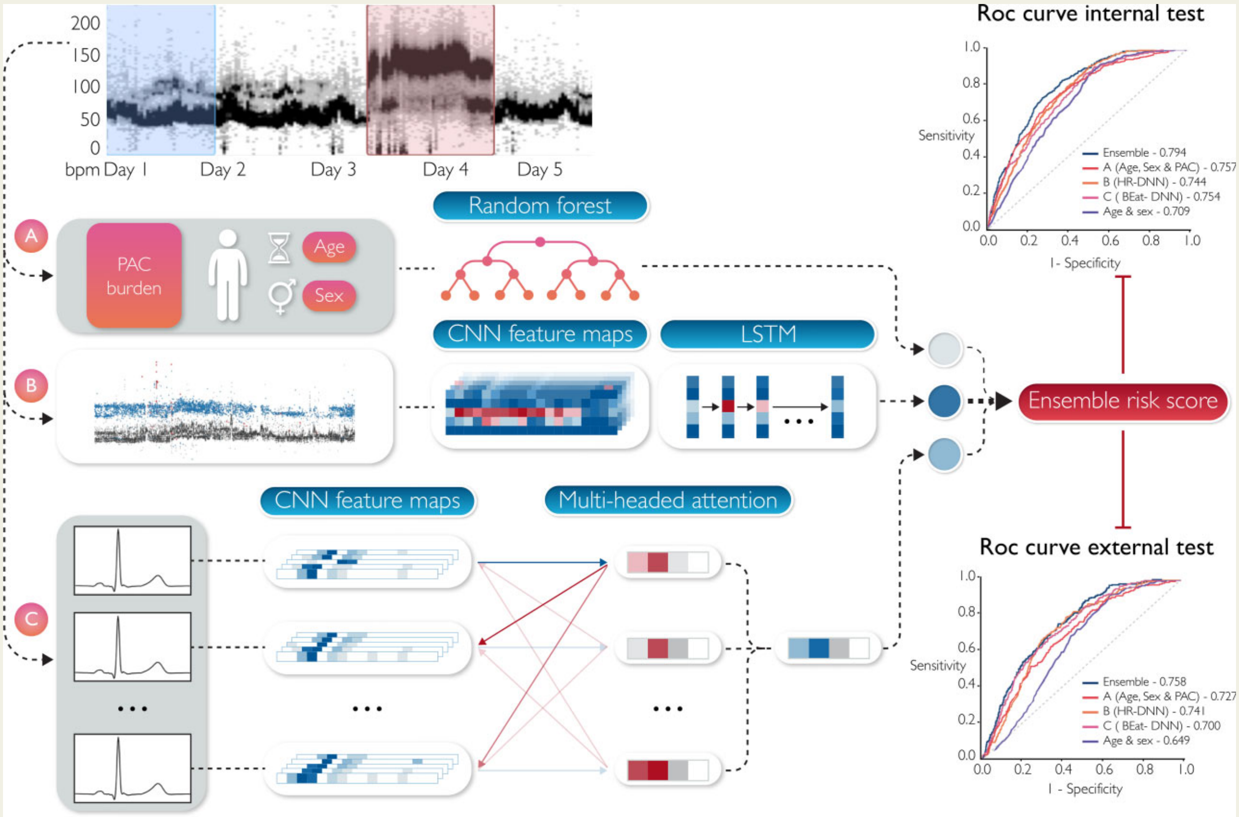
Aims Atrial fibrillation (AF) is associated with significant morbidity but remains underdiagnosed. A 24 h ambulatory electrocardiogram (ECG) is largely used as a tool to document AF but yield remains limited. We hypothesize that a deep learning model can identify patients at risk of AF in the 2 weeks following a 24 h ambulatory ECG with no documented AF.

Methods and results We identified a training set of Holter recordings of 7–15 days duration, in which no AF could be found in the first 24 h. We trained a neural network to predict the presence or absence of AF in the 15 following days, using only the first 24 h of the recording. We evaluated the neural network on a testing set and an external data set not used during algorithm development. In the testing data set, out of 9993 Holters with no AF on the first day, we found 361 (4%) recordings with AF within the 15 subsequent days of monitoring [5808, 218 (4%), respectively in the external data set]. The neural network could discriminate future AF with an area under the receiver operating curve, a sensitivity, and specificity of 79.4%, 76%, and 69%, respectively (75.8%, 78%, and 58% in the external data set), and outperformed ECG features previously shown to be predictive of AF.

Conclusion We show here the very first study of short-term AF prediction using 24 h Holter monitoring. This could help identify patients who would benefit the most from longer recordings and proactively initiate treatment and AF mitigation strategies in high-risk patients.

* Corresponding author. Email: jsingh@mgh.harvard.edu

Graphical Abstract



AF prediction ensemble model derived from 24-hour Holter data.

Keywords Atrial fibrillation • Risk prediction • Deep learning • Holter • Ambulatory monitoring

Introduction

Atrial fibrillation (AF) is the most common sustained cardiac arrhythmia that affects 46 million individuals worldwide and is associated with poor clinical outcomes due to increased risk of stroke, acute coronary events, and heart failure.[1,2] Early diagnosis of AF can facilitate early treatment and preventative strategies that in turn could mitigate downstream complications. Previous studies have shown that a longer duration of monitoring has a higher yield of AF detection.[3,4] Notably, short Holter recordings (24–48 h) although convenient and often used for AF screening have a low detection rate.[5]

Recent research has demonstrated the potential for artificial intelligence (AI) to predict incident AF from the 12-lead electrocardiogram (ECG) ranging over a period of months to years.[6,7] There is however no prior work examining the role of AI in predicting AF in the short-term (days to weeks) from a 24 h Holter (Holter-AI). Although there are challenges with using Holter recordings due to less standardized lead placement and a smaller number of leads, ambulatory ECG provides longer-duration signals that may offer additional inputs for prediction models.

The current study is the first of its kind, where we used a deep neural network (DNN) to predict short-term AF occurrence, within 15 days from a 24 h Holter. The study also seeks to explain the ECG features along with the heart rate (HR) trends and pre-mature atrial complexes that influence AF risk estimates which are critical in addressing potential bias and enhancing clinician confidence. Early short-term prediction of AF can either enable the early initiation of treatment or recommend the need for longer Holter recordings to detect incident AF.

Methods

Data sources and study population

We retrospectively collected and de-identified Holter recordings from six independent diagnostic testing facilities (IDTFs) in the USA, European Union, South Africa, India, and UK, from 1 January 2019 to 31 August 2021. These were gathered into an internal data set. Of note, a separate external data set was built from Holter recorded between 1 January 2018 and 31 August 2021, in two different IDTFs from the USA.

All recordings corresponded to adult patients. Recordings were pre-processed using the Cardiologs Holter Platform proprietary algorithm, re-sampled to 250 Hz, and analysed by an ECG technician. As all IDTFs did not use the same Holter devices and number of recording leads, only one lead of the Holter was used for the training and evaluation of each model, always the same for each IDTF.

The internal data set was divided into an internal development data set, consisting of Holter recordings from 1 January 2019 to 30 June 2021 and a testing set, consisting of Holters from 1 July 2021 to 31 August 2021. The inclusion diagram in *Figure 1* details the steps for sample selection. The internal development data set was divided into training and validation in the following way: Holters presenting AF in the first 24 h were all included in the training set so as to keep the characteristics of the validation set consistent with the evaluation setting (24 h of ECG signal with no AF); all other Holters were assigned to training and validation in an 80/20 proportion, based on date. The external data set was used solely for testing, to measure the generalizability of the proposed models. None of the Holters used for development were included in either testing set.

In particular, for the evaluation sets (validation, internal, and external testing), two main criteria were used for the inclusion of the recordings: (i) Holters lasting between 7 and 15 days to reliably assess whether or not AF was present in the extended recording and (ii) Holters free from AF event in the first 24 h. We further excluded Holters with no age and sex information from the internal and external testing sets.

Outcome

The primary objective of the study was to test the ability of using a DNN on 24 h Holters with no documented AF to identify patients likely to present with AF in the next 2 weeks using an extended recording. Episodes of AF were all annotated by ECG technicians. Atrial fibrillation and atrial flutter are considered indistinct in this study. This choice was motivated by the common coexistence of these conditions and the similarity in pathophysiology and treatment.[8]

Model development and evaluation

We developed different models for predicting the occurrence of AF up to 15 days. We used a simple model based on age and sex as a baseline. We then implemented three new models based on different inputs and finally combined those three models to form a final ensemble model (*Figure 2*).

Age and sex model

For reference, we implemented a random forest model using only the age and sex of a patient.

Age, sex, and pre-mature atrial contraction model

To improve upon the age and sex model, we proposed a stronger patient-feature model, as a random forest model using the patients age, sex, and pre-mature atrial contraction (PAC) count from the first 24 h of the recording [age, sex, and pre-mature atrial contraction (ASP) model]. This was motivated by the relationship that has been shown between PAC count and AF.[9] The Holter data allowed us to have an accurate estimate of the PAC count for each recording, based on the analysis by a technician.

Deep neural network for short-term atrial fibrillation prediction

Heart rate-deep neural network

While the duration of a Holter enables the use of aggregated information such as PAC count, we leveraged the greater granular information available in a 24-h ambulatory ECG. To that end, we implemented a neural network using the instantaneous HR as input, containing both long-term information like HR variations throughout the day and granular information related to the occurrence of PACs and pre-mature ventricular contractions (PVCs). The input to this model is an HR plot (*Figure 3*). We split the first 24 h of each recording into two 12 h windows where the final risk score was an average prediction from the two windows. The model's architecture was composed of two modules: (i) a convolutional neural network (CNN) that outputs a feature map and (ii) a long short-

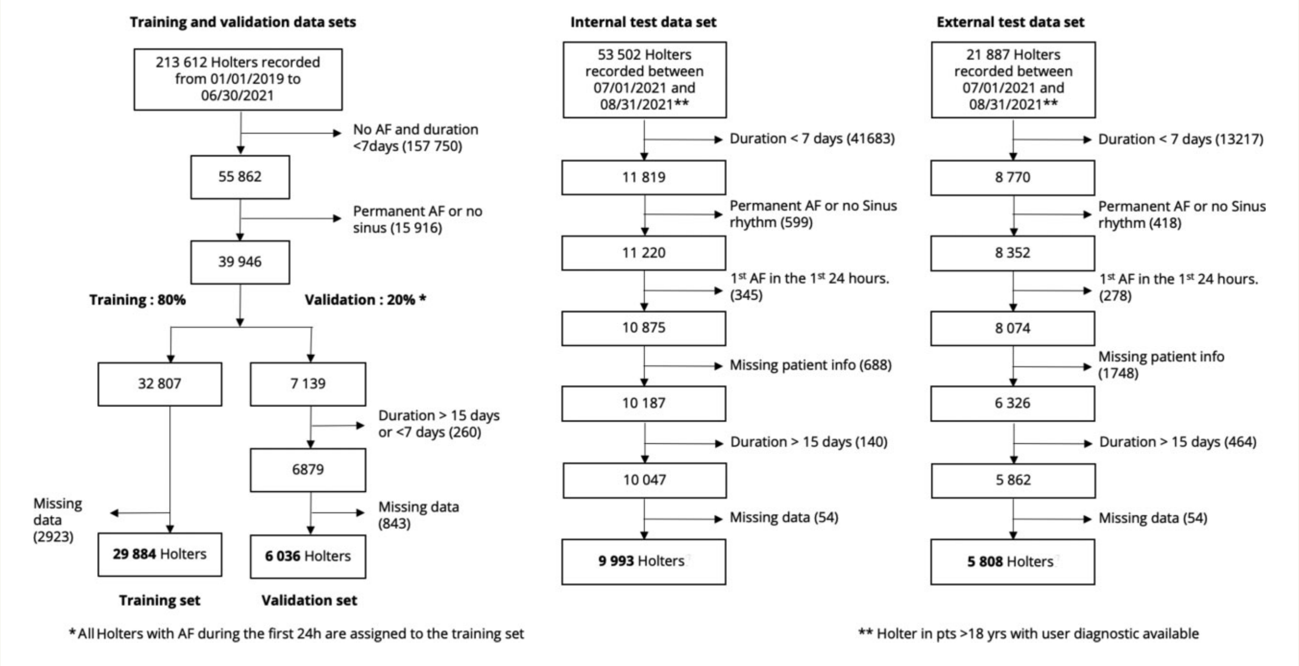

Training and validation data sets
213 612 Holters recorded from 01/01/2019 to 06/30/2021
No AF and duration <7days (157 750)
55 862
Permanent AF or no sinus (15 916)
39 946
Training : 80%
Validation : 20% *
32 807
7 139
Duration > 15 days or <7 days (260)
6879
Missing data (843)
Missing data (2923)
29 884 Holters
6 036 Holters
Training set
Validation set
Internal test data set
53 502 Holters recorded between 07/01/2021 and 08/31/2021**
Duration < 7 days (41683)
11 819
Permanent AF or no Sinus rhythm (599)
11 220
1st AF in the 1st 24 hours. (345)
10 875
Missing patient info (688)
10 187
Duration > 15 days (140)
10 047
Missing data (54)
9 993 Holters
External test data set
21 887 Holters recorded between 07/01/2021 and 08/31/2021**
Duration < 7 days (13217)
8 770
Permanent AF or no Sinus rhythm (418)
8 352
1st AF in the 1st 24 hours. (278)
8 074
Missing patient info (1748)
6 326
Duration > 15 days (464)
5 862
Missing data (54)
5 808 Holters
* All Holters with AF during the first 24h are assigned to the training set
** Holter in pts >18 yrs with user diagnostic available

Figure 1 Patient inclusion diagram. Describes which data are used, which are discarded (no diagnostic, <18 years, Holter of ≤24 h, Holter with atrial fibrillation in the first 24 h, negative Holter of <7 days, Holter with persistent atrial fibrillation or no Normal Sinus Rhythm (NSR)).

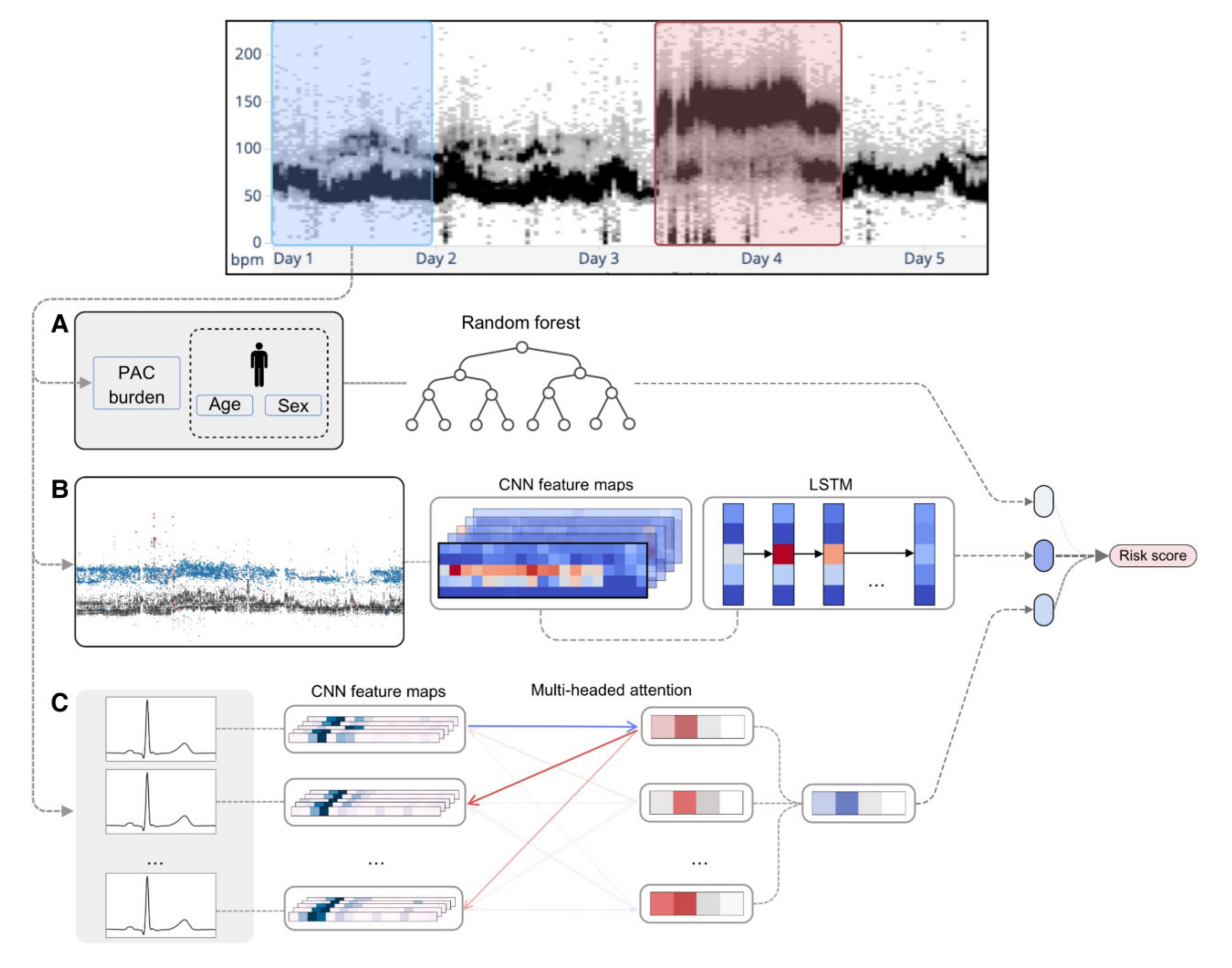


Figure 2 Model's diagram. The first 24 h from a Holter are used to derive inputs for the different models while the following 14 days are used to assign a label. (*A*) Patient age, sex, and 24 h pre-mature atrial contraction burden are passed into a random forest classifier to predict an age, sex, and pre-mature atrial contraction score. (*B*) A convolutional neural network is used to extract feature maps from a heart rate plot, which are passed into a long short-term memory layer as temporal features to predict a heart rate-deep neural network score. (*C*) The Holter is split into sequences of 80 consecutive beats which are individually passed into a convolutional neural network for feature extraction. Then an attention layer identifies relationships between beats within a sequence to predict a score; the average prediction of all sequences is used as the Beat-deep neural network score. The prediction scores from each model are then passed into a logistic regression classifier to generate a final risk score.

term memory network that performs temporal analysis of the feature map by treating it as a time series.

Beat-deep neural network

This second DNN-based model used the raw ECG data as input. To effectively utilize a full 24-h recording, we split the signal into sequences of 80 consecutive beats to use as input. The onsets of the beats were automatically computed by the Cardiologs Holter Platform. The model architecture is composed of two modules: (i) a CNN applied to each beat independently and (ii) an attention module using as input the features extracted from the CNN to exploit relationships between different beats within a sequence. A final fully connected layer was used to predict a risk score from the average output of the attention module. The final Beat-DNN prediction for a recording is taken as the average prediction across all sequences.

Both neural networks were implemented in Keras, with a Tensorflow (Google, Mountainview, CA, USA) backend. The training set was used for model development and model selection. Hyper-parameter tuning was done using the validation set to compare results.

Ensemble model

To aggregate predictions from the different models into a final risk score, a stacking ensemble approach was used. Notably, this has been shown to improve performance when combining multiple models.[10] The ensemble model consists of a meta-learner, which takes as input scores from the Beat-DNN, HR-DNN, and ASP models. The meta-learner chosen for the ensemble model is a logistic regression classifier.

Model interpretability

To better understand the features contributing to the HR-DNN and Beat-model outputs, we used visualization tools highlighting regions of the input signal with a strong positive impact on the predicted score.

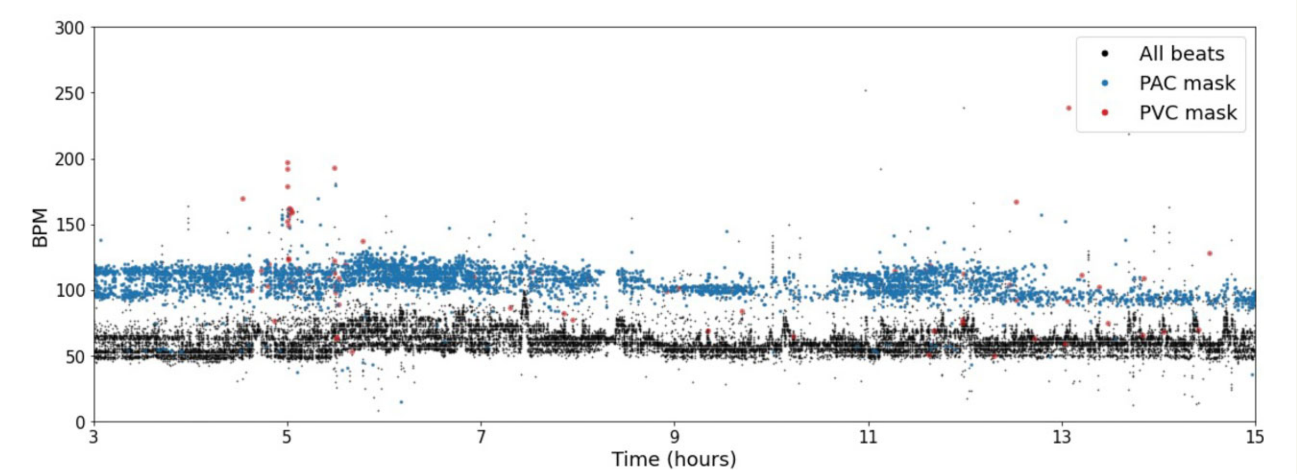


Figure 3 Heart-rate plot. A heart-rate plot is a three-dimensional representation of every beat's instantaneous heart rate during the Holter recording. The *x*-axis represents time, while the *y*-axis represents the heart rate in b.p.m.. The *z*-axis consists of three channels that correspond to each beats classification of either: normal, pre-mature atrial contraction, or pre-mature ventricular contraction. The onset of the beats and classification of normal, pre-mature atrial contraction, or pre-mature ventricular contraction were automatically computed by the Cardiologs Holter Platform. The heart-rate plots cover a 12-h window with a resolution of 36 s per time bin/column and 1 b.p.m./heart rate bin/row. The final input has a size of 1200 × 300 × 3.

Table 1 Population description

	Atrial fibrillation, *n* (%)	**Training (*n* = 29 884)** **3307 (11.1%)**	**Validation (*n* = 6036)** **228 (3.8%)**	**Internal test (*n* = 9993)** **361.0 (3.6%)**	**External test (*n* = 5808)** **218.0 (3.8%)**
Age, *n* (%)	<65 years	14 758 (49.4%)	3334 (55.2%)	5518 (55.2%)	2951 (50.8%)
	≥65 years	11 483 (38.4%)	2702 (44.8%)	4475 (44.8%)	2857 (49.2%)
	Missing	3643 (12.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Sex, *n* (%)	Male	10 196 (34.1%)	2265 (37.5%)	3841 (38.4%)	2179 (37.5%)
	Female	16 428 (55.0%)	3771 (62.5%)	6152 (61.6%)	3629 (62.5%)
	Missing	3260 (10.9%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

For the HR-DNN model, we used LayerCAM for visualization, a method that creates a saliency map for a given prediction.[11] The saliency map is created by combining feature maps of the intermediate convolutional layers and the gradient of the output relative to the feature maps.

For the Beat-DNN model, in order to capture regions around a beat which were of importance to the model's prediction, we focused on Module (i) of the Beat-DNN to consider each beat individually, applying the guided Grad-CAM visualization method on the final convolutional layer of the module.[12]

Statistical analysis

Continuous data are presented as mean values with standard deviations (SDs). All categorical data are presented as proportions. Model performance is presented as sensitivity, specificity, positive predictive value (PPV), and area under the receiver operator curve (AUC). All 95% confidence intervals (CIs) for sensitivity, specificity, and PPV were computed using Wilson score intervals. AUC CIs and differences of AUC between models were calculated using the fast version of Delong's algorithm.[13] We also provide the *P*-value of the *z*-test for comparison of AUCs as defined by Sun and Xu.[13] Age distributions were compared with a Student's *t*-test. A *P*-value <0.05 was considered significant. Statistical analyses and model development were performed using Python 3.8.

Results

Study population

A total of 267 114 patient Holter recordings collected in six different IDTFs were used to form the training, validation, and internal data sets using inclusion and exclusion criteria, as specified in *Figure 1*. About 21 887 other Holter recordings from two independent IDTF were used to build the external data set. A total of 45 913 patient Holter recordings were identified in the internal data set and 5808 recordings in the external data set after applying exclusion criteria (*Figure 1*). Among patients in the internal data set, 35 920 patients were used for model development. In the training data set, 1821 (6.1%) patients were

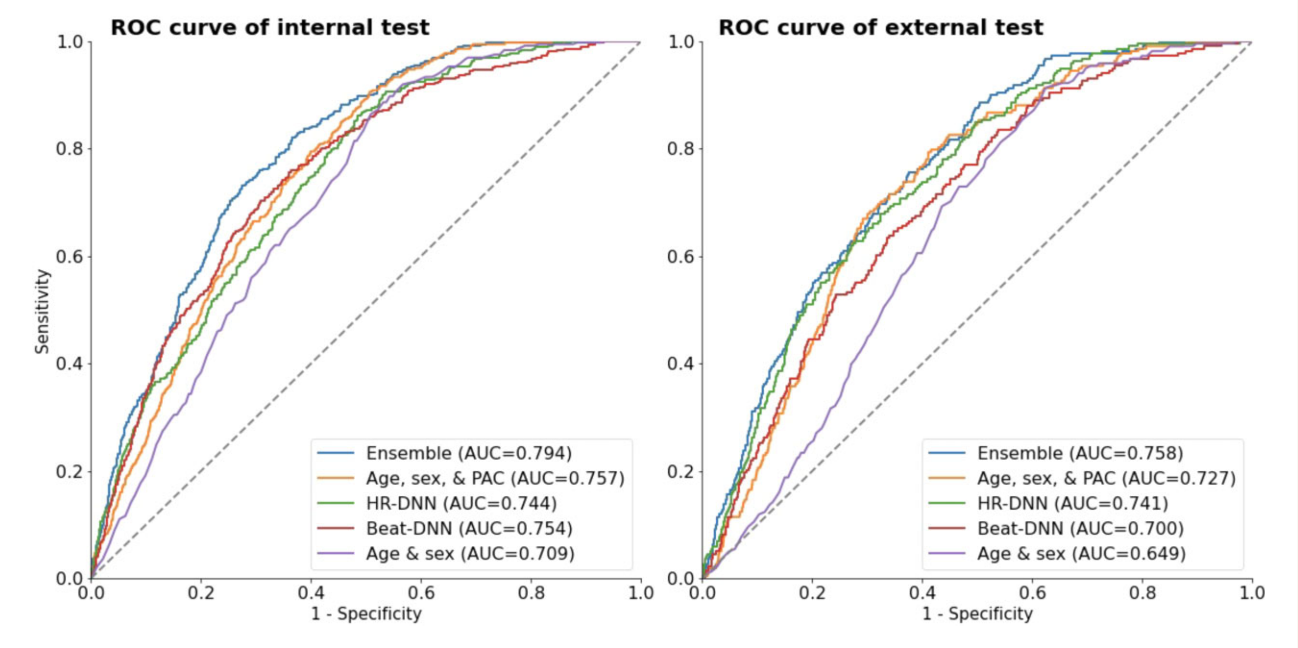


Figure 4 Receiver operating characteristic curves for each of the models on the internal and external test sets.

observed with AF following the first 24 h and an additional 1486 (5.0%) with AF in the first 24 h were included. Following the first 24 h of a recording, we observed 361 (3.6%) patients with AF in the internal test data set and 218 (3.8%) in the external data set.

The mean age (±SD) of patients in the internal test data set was 58.9 ± 17.7 years and 60.5 ± 17.8 years for the external data set. Of note, 61.6% of patients were females in the internal test data set and 62.5% in the external data set. In the internal data set, men had an average age of 61.2 ± 16.5 years and women had 57.5 ± 18.3 years. In the external data set, men had an average age of 62.4 ± 16.8 years and women had 59.4 ± 18.3 years. In both test sets, men were significantly older than women ($P < 0.05$). *Table 1* includes patient demographic characteristics for each data set.

Model performance

The AUCs for each of the models are shown in *Figure 4*. In the internal testing set, AUC was 0.709 for the age and sex model (95% CI 0.688–0.730), 0.757 for the ASP model (0.738–0.777), 0.744 for HR-DNN (0.722–0.766), 0.754 for Beat-DNN (0.731–0.778), and 0.794 for the ensemble model (0.775–0.813). In the external testing set, AUC was 0.649 for the age and sex model (0.620–0.677), 0.727 for the ASP model (0.698–0.755), 0.741 for the HR-DNN model (0.712–0.770), 0.700 for the Beat-DNN model (0.668–0.731), and 0.758 for the ensemble model (0.730–0.785), respectively. We observed similar performances of the ASP, HR-DNN, and Beat-DNN models when evaluating individually, except for the HR-DNN model on the external test set, however, noticed a significant improvement in performance after ensembling scores from each of the models (*Table 2*). We also observed significantly higher performances for all proposed models compared with the age and sex model (*Table 2*).

With an operating point calculated using the F2 score on the validation set, the sensitivity, specificity, PPV, and negative predictive value of the ensemble model were 75.9% (71.2–80.0%), 69.0% (68.1–69.9%), 8.4% (7.5–9.4%), and 98.7% (98.4–99.0%) in the internal testing set and 78.0% (72.0–83.0%), 58.2% (56.9–59.4%), 6.8% (5.9–7.8), and 98.5% (98.1–98.9%) in the external testing set, respectively. Performance of each model in both test data sets is provided in *Table 2*.

While we did notice a significant drop in the performances in the external data set, we still observe a significant improvement in the ensemble model compared with the age and sex model.

Sub-group analysis

Results of the internal testing set across sub-groups defined by age and sex for each of the models are shown in *Table 3*. We observed a significant improvement in AUC across each sub-group for the ensemble model compared with the age and sex model (see Supplementary material online, *Table S1*). We also observed a small but significant improvement in the performances in the female sub-group and the below 65 years sub-group.

Model interpretability

Heart rate plot model interpretability

The saliency map allows us to locate regions of the HR input having a strong positive impact on the predicted score. Among patients with a high likelihood of AF, the most salient features appear to be regions with high volumes of PACs. An example of a Holter with a high predicted score and PAC count is shown in *Figure 5*.

Beat-deep neural network interpretability

With the Grad-CAM visualization method, we observed the importance of the P-wave morphology for the model's predictions as illustrated in the different examples shown in *Figure 6*.

Table 2 Models' performances

	Internal testing set					External testing set				
	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
Ensemble	79.4 (77.5–81.3)*	76.0 (71.3–80.1)	69.0 (68.0–69.9)	8.4 (7.5–9.4)	98.7 (98.4–99.0)	75.8 (73.0–78.5)*	78.0 (72.0–83.0)	58.1 (56.8–59.4)	6.8 (5.9–7.8)	98.5 (98.1–98.9)
ASP	75.7 (73.7–77.6)***	65.8 (60.8–70.5)	70.9 (70.0–71.8)	7.8 (6.9–8.8)	98.2 (97.9–98.5)	72.6 (69.8–75.5)***	68.8 (62.4–74.6)	68.6 (67.3–69.8)	7.9 (6.7–9.2)	98.3 (97.8–98.6)
HR-DNN	74.5 (72.3–76.7)***	92.6 (89.5–94.9)	41.0 (40.0–42.0)	5.6 (5.0–6.2)	99.3 (99.0–99.5)	74.1 (71.2–77.0)***	92.2 (87.9–95.1)	37.3 (36.0–38.6)	5.4 (4.7–6.2)	99.2 (98.7–99.5)
Beat-DNN	75.4 (73.1–77.8)***	57.7 (52.5–62.6)	76.7 (75.8–77.5)	8.5 (7.4–9.6)	98.0 (97.6–98.3)	70.0 (66.8–73.1)***	70.2 (63.8–75.9)	58.6 (57.3–59.8)	6.2 (5.3–7.2)	98.1 (97.5–98.5)
Age and sex	70.5 (68.4–72.6)**	87.4 (83.6–90.4)	47.2 (46.2–48.2)	5.8 (5.3–6.5)	99.0 (98.7–99.3)	64.9 (62.0–67.7)**	82.6 (77.0–87.0)	44.1 (42.8–45.4)	5.4 (4.7–6.3)	98.5 (97.9–98.9)

$^{*}P < 0.05$ for comparison of AUC between ensemble model and other models.
$^{**}P < 0.05$ for comparison of AUC between age and sex model and other models.

Discussion

To the best of our knowledge, we describe the first short-term AF prediction system based on a 24-h ambulatory ECG. Using a deep learning-based meta-model, we showed good performances (AUC 0.79) that were comparable with AF risk score performances on long-term prediction.[14] The association of different models taking advantage of specific structure from an ambulatory ECG (HR-DNN, Beat-DNN, and PAC count) allowed the final ensemble model to perform significantly better than an age- and sex-based model. We also observed that this model performed better in the age group of <65 years and women sub-groups. Using visualization methods, the model decisions were found to be partially based on some previously known risk factors, PAC burden, and P-wave morphology.

Atrial fibrillation is an evolutive disease with a progression from asymptomatic atrial cardiomyopathy to overt or silent paroxysmal then persistent and ultimately permanent AF.[15] For this reason, it is important to find a way to predict it before the first episode or between unrecognized episodes. From the pathophysiological perspective, three classical factors are known to play a role in the genesis of AF, namely the substrate, autonomous nervous system, and triggers. Notably, all these factors can be evaluated on a Holter ECG. Dilatation and progressive fibrosis of the atrium preceding AF episodes may be reflected within various degrees of changes within the P-wave morphology. Abnormalities in the sympathovagal balance can be determined through the assessment of the HR and its variability. PACs that serve as trigger AF have also been shown to be correlated by their number and coupling intervals to the long-term risk of AF.[16]

While a 12-lead ECG gives immediate access to spatial cardiac information on a short period, 24 h Holter includes a specific temporal component that offers additional inputs for models. We used those specificities of a 24 h ambulatory Holter to develop our models. In the ASP model, we use the PAC count assessed on a 24 h period which has been identified as a predictor of future AF.[17] The 24-h HR plot provides information on HR variation over the whole period which can reflect HR variability, known to contain information about future AF.[18] The Beat-DNN approach focuses on the components of the ECG signal itself that have been shown to impact AF risk including P-wave morphology,[19] PR interval,[20] or other signal-related features like QT interval.[21]

We observed differences in some models' performance between the external and internal data sets which can be linked to the following reasons: First, the baseline age and sex model lower performance on the external data set can be explained by a different distribution of age, both for men and women, between the internal and external data sets, which makes age less indicative of AF in the external testing set. The Beat-DNN model also presents a performance discrepancy between data sets. One hypothesis for this behaviour is the fact that centres used in the external data set have not been used in model development. Therefore, factors such as device and lead placement, which have an impact on the ECG signal, can have an impact on the generalizability of a signal-based model. Finally, the ensemble model, leveraging the other models, shows reduced performance on the external data set due to the drop of performance in some of these individual components.

It is well understood that AF is a leading cause of stroke with a substantial morbidity and mortality. Identifying patients at risk is critical to guide anti-coagulation therapy.

After a stroke, it is crucial to identify the aetiology and so an AF diagnosis can lead to anti-coagulation in order to prevent another stroke. It is even more important to have a rapid diagnostic after a transient ischaemic attack where the risk of recurrent stroke is 8.0% at 7 days, 11.5% at 1 month, and 17.3% at 3 months after a transient ischemic attack.[22]

Concerning stroke primary prevention, from the contradictory results of the recent Strokestop[23] and Loop[24] studies, it is not clear today if large screening of patients to identify those at risk of stroke is beneficial. We have developed a model which, from a 24 h Holter, intends to identify patients more likely to present AF in a short-term. This may have several clinical implications, which could include recommending an extended Holter or monitoring with a higher expected yield. If and when proven effective, additional short-term risk prediction of AF may allow for early intervention and mitigation strategies.

To alleviate the black-box effect of neural networks, we sought to visualize regions in the different inputs that have a high impact on the network's decision using saliency maps. Those methods revealed that the two DNN-based models rely, at least partially, on features previously studied (PAC count, P-wave morphology) to reach a decision. There may however be some subtle changes in the signal or in the HR plot which are unrecognizable by the human eye that deep learning may consider and that still remains to be explained in the future.

The primary goal of our research is to improve the yield of AF detection in a population indicated for ambulatory monitoring. Extended Holter can be burdensome for patients and costly for hospitals. The 24-h Holter recordings are still widely used when AF is suspected.[5] If a high risk of AF in a short-term can be inferred from this recording, an extended Holter could be selectively proposed to those high-risk patients, optimizing the monitoring resources and avoiding unnecessary long recording for probable negative patients. Moreover, in a context of race against time in the secondary prevention of stroke, a short-term prediction tool used over a short duration record could impact clinical decisions.[5] In the future, this model could be adapted to different modalities also using ECG with a limited number of leads like smartwatches, hand-held devices, or similar wearables.

Our study presents several limitations. First, this study is a retrospective one involving previously collected Holter recordings and would need to be complemented by a prospective validation study to confirm the performances of the developed ensemble model. Second, as Holters were coming from various facilities across different geographies (USA, Europe, UK, India, and

Table 3 **Sub-group analysis (internal data set)**

		AF rate (%)	Age and sex AUC (95% CI)	Beat-DNN AUC (95% CI)	HR-DNN AUC (95% CI)	ASP AUC (95% CI)	Ensemble AUC (95% CI)
Age (years)	<65	1.70%	0.799 (0.758–0.839)	0.812 (0.771–0.854)	0.824 (0.786–0.862)	0.842 (0.812–0.873)	0.867 (0.836–0.897)*
	≥65	5.97%	0.520 (0.485–0.556)	0.652 (0.619–0.685)	0.612 (0.576–0.647)	0.623 (0.591–0.655)	0.674 (0.643–0.706)*
Sex	Female	2.86%	0.731 (0.704–0.759)	0.776 (0.743–0.808)	0.777 (0.749–0.804)	0.786 (0.761–0.811)	0.821 (0.797–0.846)*
	Male	4.82%	0.652 (0.619–0.685)	0.712 (0.675–0.748)	0.691 (0.654–0.729)	0.706 (0.673–0.738)	0.747 (0.714–0.780)*

*$P < 0.05$ for comparison of AUC between ensemble model and age and sex model.

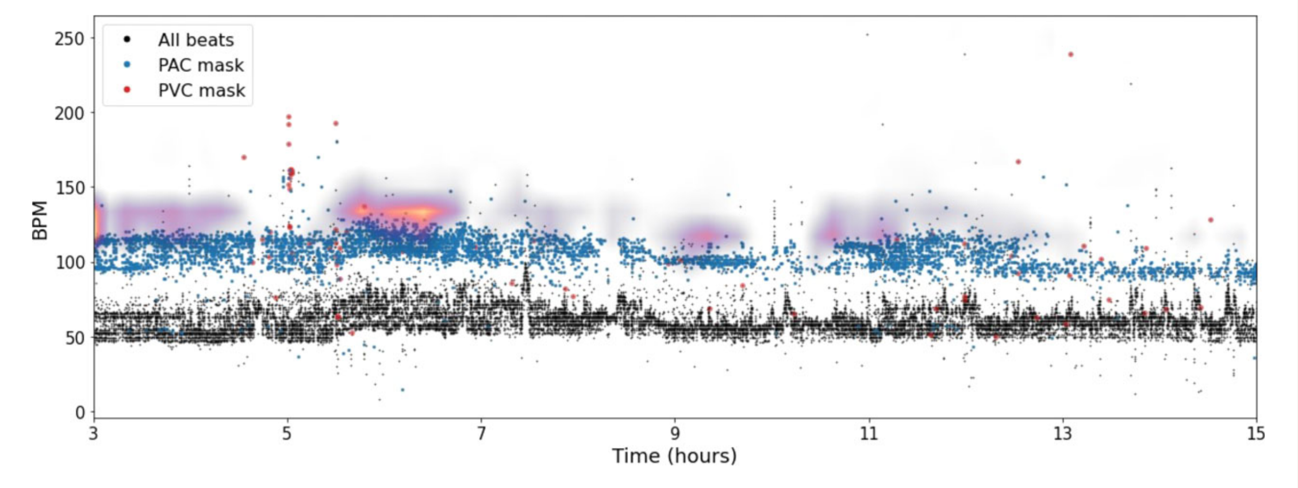


Figure 5 Heart rate-deep neural network interpretability. Saliency map overlaid on a heart-rate plot of the first 15 h of a true-positive Holter (a Holter with atrial fibrillation predicted from the first 24 h and atrial fibrillation documented within 2 weeks after). The highlighted areas indicate regions of the input that influence the prediction. The saliency map activations focus predominantly on regions with high density of pre-mature atrial contractions.

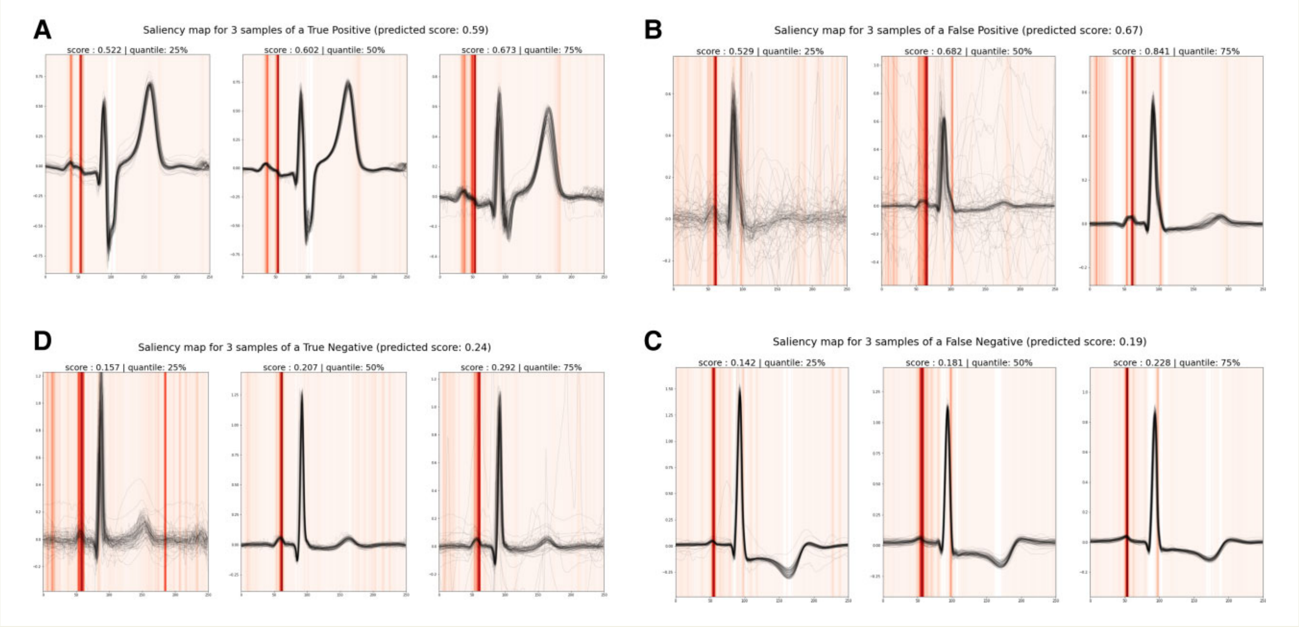


Figure 6 Beat-deep neural network interpretability. Saliency maps are given for four Holters (*A*–*D*). For each of them, three samples are shown corresponding to median, first, and last quartile of prediction scores of the recording's samples. All 80 beats of each sample are plotted in black. The saliency map averaged across the 80 beats is shown in vertical lines (in red). For all recordings, the P wave is the main region of importance. Holters (*C*) and (*D*), respectively, are true negative (Holter with no atrial fibrillation predicted from the first 24 h and no atrial fibrillation observed within the 2 following weeks) and false negative (Holter with no atrial fibrillation predicted from the first 24 h and atrial fibrillation observed within the 2 following weeks), showing low prediction scores and a normal P wave with a single mode on which the network focuses. On the contrary, Holters (*A*) and (*B*), respectively, are true positive and false positive, showing high prediction scores with characteristics of a bifid P-wave. For these recordings, the model focuses on the two modes of the P waves. These examples highlight the importance of the P-wave morphology for the Beat-deep neural network predictions.

South Africa), the lack of attached clinical data makes it difficult to ensure that the data present the necessary diversity and absence of bias which are expected for an AI tool validation. Third, there is a performance difference between the external and internal data sets, which should lead to additional work to improve generalizability. Fourth, even if we ensured that no ECG used for testing of the neural network was used during training, due to the de-identified nature of the Holter recordings, it is still possible that a patient corresponding to a recording in the testing set also contributed to a recording involved in the training set. However, the probability for a patient to have multiple extended Holter is very low. Finally, our classification of positive or negative Holter for AF is based on a continuous recording of only 15 days. Despite being certain of the presence or absence of AF during this period, it is possible that the patient could have presented with AF outside of the recorded window.

Furthermore, the DNNs proposed in this study present several limitations. First, Beat-DNN lacks a global view of the Holter. To aggregate the predictions from each local strip, we took the average, which valued each region of the signal with the same importance. Methods such as multiple instance learning, which has proved successful in identifying discriminating areas of pathophysiology slides, could help improve Beat-DNNs performance by focusing on higher salient periods of the ECG. Furthermore, Beat-DNN and HR-DNN relied on the beat morphology and characteristics of the HR independently, which may have limited each model's ability to leverage relationships between these aspects of the signal. Further work is needed on how to exploit these features together in a single model. Finally, contrary to 12-lead ECG, lead placement for ambulatory ECG is not standardized, which could impact Beat-DNN's ability to generalize due to variability. Exposing Beat-DNN to more training data from different centres could improve generalizability.

Conclusion

In conclusion, our results suggest that a deep learning model using a 24-hour Holter recording can be used to identify patients at risk of developing short-term AF. Future studies will be needed to confirm if this early detection helps optimize resources towards the length of additional monitoring and improve patients' outcomes.

Supplementary material

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

Acknowledgement

The authors thank Christophe Gardella for providing advice for the study design.

Funding

This research is funded by Cardiologs.

Conflict of interest: J.S. Consulting for Biotronik, Boston Scientific, Medtronic, Abbott Inc, EBR systems Inc, Cardiologs Inc, Microport, CVRx, Impulse Dynamics, Medscape Inc, Sanofi Inc, Cardiac Rhythm Group, Implicity Inc., GLG Inc. No stock options or speakers bureau. S.K. No conflicts.

Data availability

Individual data used for this study are not publicly available for proprietary reason. The coding used to train the models is dependent on annotation, infrastructure, and hardware and therefore, cannot be released.

References

1. Kornej J, Börschel CS, Benjamin EJ, Schnabel RB. Epidemiology of atrial fibrillation in the 21st century: novel methods and new insights. *Circ Res* 2020;**127**:4–20.
2. Michaud GF, Stevenson WG. Atrial fibrillation. *N Engl J Med* 2021;**384**:353–361.
3. Gladstone DJ, Spring M, Dorian P, Panzov V, Thorpe KE, Hall J, Vaid H, O'Donnell M, Laupacis A, Côté R, Sharma M, Blakely J, Shuaib A, Hachinski V, Coutts SB, Sahlas DJ, Teal P, Yip S, Spence JD, Buck B, Verreault S, Casaubon LK, Penn A, Selchen D, Jin A, Howse D, Mehdiratta M, Boyle K, Aviv R, Kapral M, Mamdani M, EMBRACE investigators and coordinators. Atrial fibrillation in patients with cryptogenic stroke. *N Engl J Med* 2014;**370**:2467–2477.
4. Sanna T, Diener HC, Passman RS, Di Lazzaro V, Bernstein RA, Morillo CA, Rymer MM, Thijs V, Rogers T, Beckers F, Lindborg K, Brachmann J. Cryptogenic stroke and underlying atrial fibrillation. *N Engl J Med* 2014;**370**:2478–2486.
5. Steinberg JS, Varma N, Cygankiewicz I, Aziz P, Balsam P, Baranchuk A, Cantillon DJ, Dilaveris P, Dubner SJ, El-Sherif N, Krol J, Kurpesa M, La Rovere MT, Lobodzinski SS, Locati ET, Mittal S, Olshansky B, Piotrowicz E, Saxon L, Stone PH, Tereshchenko L, Turakhia MP, Turitto G, Wimmer NJ, Verrier RL, Zareba W, Piotrowicz R. 2017 ISHNE-HRS expert consensus statement on ambulatory ECG and external cardiac monitoring/telemetry. *Heart Rhythm* 2017;**14**:e55–e96.
6. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, Carter RE, Yao X, Rabinstein AA, Erickson BJ, Kapa S, Friedman PA. 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**:861–867.
7. Raghunath S, Pfeifer JM, Ulloa-Cerna AE, Nemani A, Carbonati T, Jing L, vanMaanen DP, Hartzel DN, Ruhl JA, Lagerman BF, Rocha DB, Stoudt NJ, Schneider G, Johnson KW, Zimmerman N, Leader JB, Kirchner HL, Griessenauer CJ, Hafez A, Good CW, Fornwalt BK, Haggerty CM. Deep neural networks can predict new-onset atrial fibrillation from the 12-lead electrocardiogram and help identify those at risk of AF-related stroke. *Circulation* 2021;**143**:1287–1298.
8. Writing Group Members, January CT, Wann LS, Calkins H, Chen LY, Cigarroa JE, Cleveland JC Jr, Ellinor PT, Ezekowitz MD, Field ME, Furie KL, Heidenreich PA, Murray KT, Shea JB, Tracy CM, Yancy CW. 2019 AHA/ACC/HRS Focused Update of the 2014 AHA/ACC/HRS Guideline for the Management of Patients With Atrial Fibrillation. A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society. *Circulation* 2019;**140**:e125–e151.
9. Dewland TA, Vittinghoff E, Mandyam MC, Heckbert SR, Siscovick DS, Stein PK, Psaty BM, Sotoodehnia N, Gottdiener JS, Marcus GM. Atrial ectopy as a predictor of incident atrial fibrillation: a cohort study. *Ann Intern Med* 2013;**159**:721–728.
10. Ting KM, Witten IH. Issues in stacked generalization. *J Artif Intell Res* 1999;**10**:271–289.
11. Jiang PT, Zhang CB, Hou Q, Cheng MM, Wei Y. LayerCAM: exploring hierarchical class activation maps for localization. *IEEE Trans Image Process* 2021;**30**:5875–5888.
12. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. *Grad-CAM: visual explanations from deep networks via gradient-based localization.* In: 2017 IEEE International Conference on Computer Vision, 2017, Venice, Italy, pp. 618–626.
13. Sun X, Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process Lett* 2014;**21**:1389–1393.
14. Poorthuis MHF, Jones NR, Sherliker P, Clack R, de Borst GJ, Clarke R, Lewington S, Halliday A, Bulbulia R. Utility of risk prediction models to detect atrial fibrillation in screened participants. *Eur J Prev Cardiol* 2021;**28**:586–595.
15. Hindricks G, Potpara T, Dagres N, Arbelo E, Bax JJ, Blomström-Lundqvist C, Boriani G, Castella M, Dan G-A, Dilaveris PE, Fauchier L, Filippatos G, Kalman JM, La Meir M, Lane DA, Lebeau J-P, Lettino M, Lip GYH, Pinto FJ, Thomas GN, Valgimigli M, Van Gelder IC, Van Putte BP, Watkins CL, Kirchhof P, Kühne M, Aboyans V, Ahlsson A, Balsam P, Bauersachs J, Benussi S, Brandes A, Braunschweig F, Camm AJ, Capodanno D, Casadei B, Conen D, Crijns HJGM, Delgado V, Dobrev D, Drexel H, Eckardt L, Fitzsimons D, Folliguet T, Gale CP, Gorenek B, Haeusler KG, Heidbuchel H, Iung B, Katus HA, Kotecha D, Landmesser U, Leclercq C, Lewis BS, Mascherbauer J, Merino JL, Merkely B, Mont L, Mueller C, Nagy KV, Oldgren J, Pavlović N, Pedretti RFE, Petersen SE, Piccini JP, Popescu BA, Pürerfellner H, Richter DJ, Roffi M, Rubboli A, Scherr D, Schnabel RB, Simpson IA, Shlyakhto E, Sinner MF, Steffel J, Sousa-Uva M, Suwalski P, Svetlosak M, Touyz RM, Dagres N, Arbelo E, Bax JJ, Blomström-Lundqvist C, Boriani G, Castella M, Dan G-A, Dilaveris PE, Fauchier L, Filippatos G, Kalman JM, La Meir M, Lane DA, Lebeau J-P, Lettino M, Lip GYH, Pinto FJ, Neil Thomas G, Valgimigli M, Van Gelder IC, Watkins CL, Delassi T, Sisakian HS, Scherr D, Chasnoits A, Pauw MD, Smajić E, Shalganov T, Avraamides P, Kautzner J, Gerdes C, Alaziz AA, Kampus P, Raatikainen P, Boveda S, Papiashvili G, Eckardt L, Vassilikos V, Csanádi Z, Arnar DO, Galvin J, Barsheshet A, Caldarola P, Rakisheva A, Bytyçi I, Kerimkulova A, Kalejs O, Njeim M, Puodziukynas A, Groben L, Sammut MA, Grosu A, Boskovic A, Moustaghfir A, Groot Nd, Poposka L, Anfinsen O-G, Mitkowski PP, Cavaco DM, Siliste C, Mikhaylov EN, Bertelli L, Kojic D, Hatala R, Fras Z, Arribas F, Juhlin T, Sticherling C, Abid L, Atar I, Sychov O, Bates MGD, Zakirov NU. 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.
16. Chong B-H, Pong V, Lam K-F, Liu S, Zuo M-L, Lau Y-F, Lau C-P, Tse H-F, Siu C-W. Frequent premature atrial complexes predict new occurrence of atrial fibrillation and adverse cardiovascular events. *Europace* 2012;**14**:942–947.
17. Cabrera S, Vallès E, Benito B, Alcalde Ó, Jiménez J, Fan R, Martí-Almor J. Simple predictors for new onset atrial fibrillation. *Int J Cardiol* 2016;**221**:515–520.
18. Perkiömäki J, Ukkola O, Kiviniemi A, Tulppo M, Ylitalo A, Kesäniemi YA, Huikuri H. Heart rate variability findings as a predictor of atrial fibrillation in middle-aged population. *J Cardiovasc Electrophysiol* 2014;**25**:719–724.
19. Magnani JW, Zhu L, Lopez F, Pencina MJ, Agarwal SK, Soliman EZ, Benjamin EJ, Alonso A. P-wave indices and atrial fibrillation: cross-cohort assessments from the Framingham Heart Study (FHS) and Atherosclerosis Risk in Communities (ARIC) study. *Am Heart J* 2015;**169**:53–61.e1.
20. Nielsen JB, Pietersen A, Graff C, Lind B, Struijk JJ, Olesen MS, Haunsø S, Gerds TA, Ellinor PT, Køber L, Svendsen JH, Holst AG. Risk of atrial fibrillation as a function of the electrocardiographic PR interval: results from the Copenhagen ECG Study. *Heart Rhythm* 2013;**10**:1249–1256.
21. Zhang N, Gong M, Tse G, Zhang Z, Meng L, Yan BP, Zhang L, Wu G, Xia Y, Xin-Yan G, Li G, Liu T. Prolonged corrected QT interval in predicting atrial fibrillation: a systematic review and meta-analysis. *Pacing Clin Electrophysiol* 2018;**41**:321–327.
22. Coull AJ, Lovett JK, Rothwell PM. Population based study of early risk of stroke after transient ischaemic attack or minor stroke: implications for public education and organisation of services. *Br Med J* 2004;**328**:326–328.
23. Svennberg E, Friberg L, Frykman V, Al-Khalili F, Engdahl J, Rosenqvist M. Clinical outcomes in systematic screening for atrial fibrillation (STROKESTOP): a multicentre, parallel group, unmasked, randomised controlled trial. *Lancet* 2021;**398**:1498–1506.
24. Svendsen JH, Diederichsen SZ, Højberg S, Krieger DW, Graff C, Kronborg C, Olesen MS, Nielsen JB, Holst AG, Brandes A, Haugan KJ, Køber L. Implantable loop recorder detection of atrial fibrillation to prevent stroke (The LOOP Study): a randomised controlled trial. *Lancet* 2021;**398**:1507–1516.