European Heart Journal - Digital Health (2021) **2**, 576–585
doi:10.1093/ehjdh/ztab071

ORIGINAL ARTICLE

Atrial fibrillation risk prediction from the 12-lead electrocardiogram using digital biomarkers and deep representation learning

Shany Biton[1], Sheina Gendelman[1], Antônio H. Ribeiro [2], Gabriela Miana[3,4], Carla Moreira[3], Antonio Luiz P. Ribeiro [3,4], and Joachim A. Behar [1]*

[1]Faculty of Biomedical Engineering, Technion-IIT, Haifa, Israel; [2]Department of Information Technology, Uppsala University, Uppsala, Sweden; [3]Telehealth Center, Hospital das Clínicas, Belo Horizonte, Brazil; and [4]Department of Internal Medicine, Faculdade de Medicina, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Received 9 July 2021; revised 20 July 2021; editorial decision 3 August 2021; accepted 4 August 2021; online publish-ahead-of-print 5 August 2021

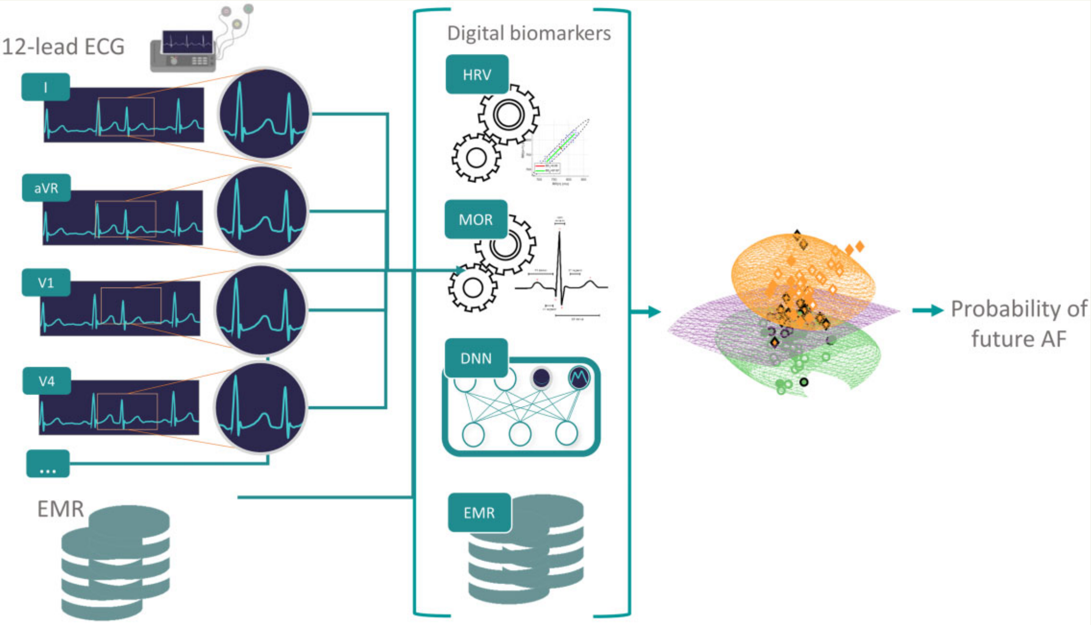
Aims This study aims to assess whether information derived from the raw 12-lead electrocardiogram (ECG) combined with clinical information is predictive of atrial fibrillation (AF) development.

Methods and results We use a subset of the Telehealth Network of Minas Gerais (TNMG) database consisting of patients that had repeated 12-lead ECG measurements between 2010 and 2017 that is 1 130 404 recordings from 415 389 unique patients. Median and interquartile of age for the recordings were 58 (46–69) and 38% of the patients were males. Recordings were assigned to train-validation and test sets in an 80:20% split which was stratified by class, age and gender. A random forest classifier was trained to predict, for a given recording, the risk of AF development within 5 years. We use features obtained from different modalities, namely demographics, clinical information, engineered features, and features from deep representation learning. The best model performance on the test set was obtained for the model combining features from all modalities with an area under the receiver operating characteristic curve (AUROC) = 0.909 against the best single modality model which had an AUROC = 0.839.

Conclusion Our study has important clinical implications for AF management. It is the first study integrating feature engineering, deep learning, and Electronic medical record system (EMR) metadata to create a risk prediction tool for the management of patients at risk of AF. The best model that includes features from all modalities demonstrates that human knowledge in electrophysiology combined with deep learning outperforms any single modality approach. The high performance obtained suggest that structural changes in the 12-lead ECG are associated with existing or impending AF.

* Corresponding author. Tel: (+972) 4 829 4125, Email: jbehar@technion.ac.il

Graphical Abstract



Keywords Atrial fibrillation • Deep learning • Risk prediction

Introduction

Major cardiovascular and cerebrovascular events occur in individuals without known pre-existing cardiovascular conditions. Preventing such events remains a serious public health challenge. For that purpose, clinical risk scores can be used to identify individuals with high cardiovascular risks.[1,2] However, available scoring scales have shown moderate performance.[3] Despite being part of the routine evaluation of many patients in both primary and specialized care, the role of electrocardiogram (ECG) analysis in cardiovascular disease prediction and, hence, prevention is not clear.

Atrial fibrillation (AF) is the most common arrhythmia, with an estimated prevalence of 3% in adults.[4] It is associated with quivering or irregular heartbeat that can lead to blood clots, stroke, heart failure, and other heart-related complications.[5] Much interest has been given to developing novel algorithms for AF detection in the past decade.[6] Because of its prevalence and clinical importance, AF is a good model for the development of new risk assessment algorithms. Wang and Wang[7] highlight AF risk-prediction using big data and Machine learning (ML) as a new opportunity to improve AF management. Two recent work from Christopoulos *et al.*[8] and Raghunath *et al.*[9] made use of a deep learning approach using the raw 12-lead ECG signal as input for AF prediction. However, the deep learning approach limits the interpretation of the features found to be predictive of future AF. Furthermore, combining features engineered from the ECG morphology and heart rate variability (HRV) with deep representation learning features may improve the predictive performance. Overall, little research to date has investigated the feasibility of cardiac abnormality risk prediction, i.e. predicting the future occurrence of a cardiac abnormality, using a data-driven approach based on the raw 12-lead ECG time series. Novel robust predictive models are likely to enhance clinical management of at-risk individuals and may also provide new insights into the aetiology of cardiac abnormalities. We introduce in this work a hybrid data-driven model combining demographics, feature engineering (or 'digital biomarkers') and deep representation learning obtained from the raw 12-lead ECG to predict the future occurrence of AF.

Methods

Database

The 12-lead ECG were obtained from the Telehealth Network of Minas Gerais (TNMG) between 2010 and 2017, a public telehealth system assisting 811 out of the 853 municipalities in the state of Minas Gerais, Brazil. The 12-lead ECG examination were performed mostly in primary care facilities using a tele-electrocardiograph manufactured by Tecnologia Eletrônica Brasileira (São Paulo, Brazil)—model TEB ECGPC—or Micromed Biotecnologia (Brasilia, Brazil)—model ErgoPC 13. Recordings were also performed in emergency departments, ambulances and hospitals. Refer to Ribeiro *et al.*[10,11] for more details. The original database had 1 773 689 patients. Patients less than 16 years old and with invalid ECG recording were excluded, resulting in 1 556 958 patients. The database used in this research is a subset from the TNMG that includes a total of $n = 1\ 130\ 404$ recordings from $m = 415\ 389$ unique patients who had repeated 12-lead ECG examinations within the study period (2010–2017) and while also excluding patients with multiple recordings but all

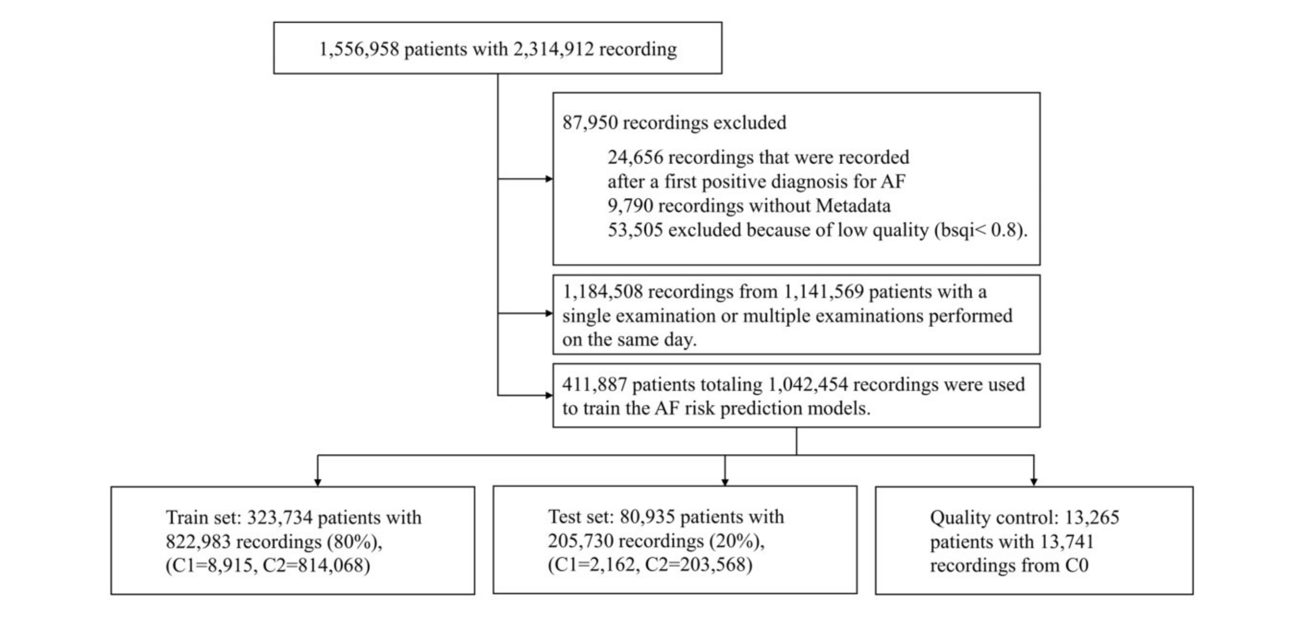

1,556,958 patients with 2,314,912 recording
87,950 recordings excluded
24,656 recordings that were recorded after a first positive diagnosis for AF
9,790 recordings without Metadata
53,505 excluded because of low quality (bsqi< 0.8).
1,184,508 recordings from 1,141,569 patients with a single examination or multiple examinations performed on the same day.
411,887 patients totaling 1,042,454 recordings were used to train the AF risk prediction models.
Train set: 323,734 patients with 822,983 recordings (80%), (C1=8,915, C2=814,068)
Test set: 80,935 patients with 205,730 recordings (20%), (C1=2,162, C2=203,568)
Quality control: 13,265 patients with 13,741 recordings from C0

Figure 1 Block diagram showing the patient inclusion and exclusion criteria and classes definition.

being performed during the same day following the baseline examination. Within the selected subset, the median and interquartile number of recordings per patient was 2 (2–3), see Supplementary material online, *Figure S1*. The median and interquartile age for the recordings were 58 (46–69), see Supplementary material online, *Figure S2*, and 38% of the patients were males. The recording length is between 7 and 10 s and were originally sampled in frequencies ranging from 3 kHz and 1 kHz. We use the same pre-processing as Ribeiro *et al.*[11]: we resample the recording at 400Hz and zero-pad the signal so it becomes 4096 samples long (4096/400 Hz–10.24 s). All ECGs were interpreted by a team of trained cardiologists using standardized criteria,[12] in order to generate an ECG free text report. The reference diagnosis for the 12-lead ECG were originally obtained by testing the agreement between recognition of specific ECG diagnoses from the cardiologist reports (using natural language processing) and the automated diagnosis provided by the Glasgow ECG analysis software. In the case of disagreement, the decision was taken heuristically or through manual review.[10]

Considering the presence or not of AF, we defined C_0 as the class containing the recordings with a positive diagnosis for AF. A recording was included in class C_1 if it had a negative diagnosis for AF and there was, for the same patient, a positive AF diagnosis documented within 5 years, time interval distribution for those patients is presented in Supplementary material online, *Figure S3*. A recording was included in class C_2 if it had a negative diagnosis for AF and there was no documented future positive diagnosis of AF for the corresponding patient. The mean follow-up time was 1.25 years. Recordings that were recorded after a first positive diagnosis for AF were discarded. This resulted in a total of 14 947 recordings in C_0, 12 142 in C_1 and 1 078 545 in C_2—see the block diagram in *Figure 1*.

Data preprocessing

To automatically assess the quality of the raw ECG examples and discard flat or noisy examples, we included a signal quality preprocessing step (see Supplementary material online, *Note SN3*). Before computing the ECG morphological biomarkers prefiltering of the raw ECG time series was performed. Specifically, we used a zero phase second-order infinite impulse response bandpass filter with the passband 0.67–100 Hz[12] to remove baseline wander and high frequency noise. We used a Notch filter at 60 Hz to remove the power-line interference (see Supplementary material online, *Figure S4*).

Digital biomarkers and representation learning features

Digital ECG biomarkers engineered from the heart rate variability (denoted 'HRV') and morphology of the ECG (denoted 'MOR') time series were engineered for each lead. In addition, the pretrained DNN[11] was used as a feature generator to include representation learning features (denoted 'DNN'). Briefly, the DNN structure of Ribeiro *et al.*[11] Consists of convolutional layer followed by four residual blocks with two convolutional layers per block. The output of the last block is fed into a fully connected layer with a Softmax function. The output of each convolutional layer is rescaled using batch normalization and fed into a rectified linear activation unit (ReLU). Finally, metadata (denoted 'META') were included as additional features. A summary diagram illustrating the overall 'hybrid' strategy taken is shown in *Figure 2*. A total of 16 HRV-based features[13] were engineered for each recording (Supplementary material online, *Table S1*). RR intervals were obtained from lead V1 using the epltd R-peak detector. For MOR features, a total of 36 different features were engineered; 12 features extracted from intervals duration (Supplementary material online, *Table S2*) and 24 from waves characteristics (Supplementary material online, *Table S3*) to describe the ECG morphology. These features were extracted for individual leads resulting in a total of 432 features. Some of these features were described in the work of Assaraf *et al.*[14] and are based on the detection of fiducial points on the ECG waveform found using the popular open source wavelet algorithm.[15] In addition, within the MOR feature set (*Table 1*), we included a measure of signal quality, bsqi, for each lead. Overall the number of MOR features per recording was 445. The DNN model presented in the

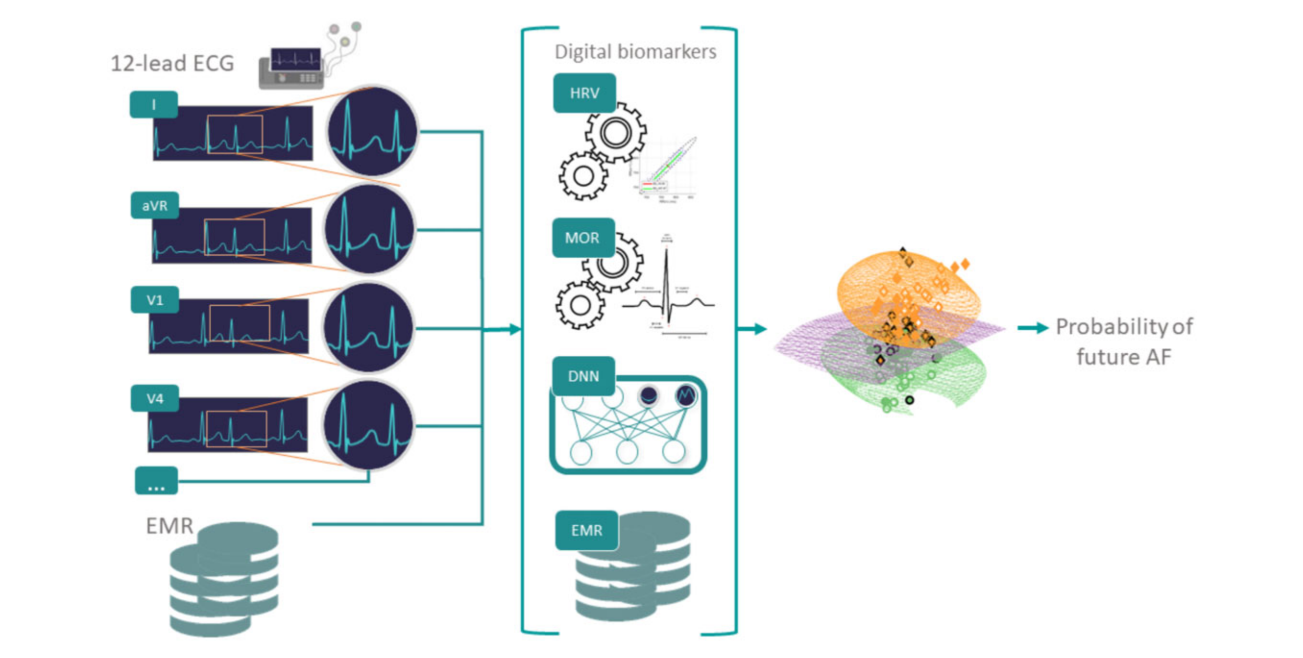


Figure 2 Overview of our experimental setting: digital biomarkers (HRV and MOR) are engineered from the 12-lead ECG. In addition, a pretrained DNN model is used as a feature generator taking as input the raw 12-lead ECG. Finally, EMR information are added as metadata features. The resulting vector of features (HRV, MOR, DNN, and EMR) is used to train random forest classifier in predicting whether an individual will develop AF within 5 years.

Table 1 The feature groups that compose each one of the presented models

	DNN	MOR	HRV	META	Number of features	Number of selected features
Model 1				×	15	—
Model 2			×		16	—
Model 3		×			445	80
Model 4	×				5120	340
Model 5	×	×	×		5581	400
Model 6	×	×	×	×	5596	180

Feature selection was performed using mRMR.

original work by Ribeiro *et al.*[11] was used as feature generators by using the features generated at the level of the first fully connected layer. For that purpose, the DNN was retrained using the 1 141 569 examples that had only 12-lead ECG recording(s) performed during the day of the baseline examination and thus were not used for our AF risk prediction experiment. In other words, the trained risk prediction models did not use any of the recordings used for training the DNN feature generator. Thus, there was no information leakage. For a 12-lead ECG recording, a total of 5120 DNN features were generated (*Table 1*). Finally, 15 META features (Supplementary material online, *Table S4*) were included (*Table 1*). This was to evaluate how patient information typically available from the hospital EMR may improve the model prediction. Although it is important to note that the META features used were self-reported by the patient at the time of the examination and thus are not directly obtained from an hospital EMR.

Performance statistics

The nonparametric Mann–Whitney rank test was used to determine whether individual features were significantly different between recordings in class C_1 and C_2. The lower the *P*-value the stronger the evidence against the null hypothesis. For all features, a *P*-value cut-off at $P < 0.05$ was used for significance testing. In order to assess the performance of the models in correctly classifying individual examples the following statistics will be computed: sensitivity (Se), specificity (Sp), positive predictive value (PPV), the harmonic mean between the Se and PPV termed F_1 measure and the area under the receiver operating characteristic curve (AUROC). We estimated the confidence interval for AUROC using bootstrapping. That is, the AUROC was repeatedly computed on randomly sampled 80% of the test set (with replacement). The procedure was repeated 1000 times and used to obtain the intervals.

ML for risk prediction

Examples from C_1 and C_2 were assigned to train-validation and test sets in an 80:20% split and with stratification by class, age and sex, see Supplementary material online, *Table S5*. ML models were trained to assess AF risk prediction defined as the ability of the classifier to predict the future development of AF within 5 years. A random forest (RF) classifier was trained with the biomarkers as input. To account for the class imbalance, the minority class was proportionally over-weighted. Feature selection was performed using the minimum redundancy maximum relevance (mRMR) algorithm.[16] In brief, mRMR selects the set of relevant features while controlling for redundancy within the set of selected features. Hyperparameters were selected using BayesSearchCV algorithm implemented in the scikit-optimize library[17] and five-fold cross-validation on the training set using 100 iterations. Hyperparameter and associated ranges for the search are listed in Supplementary material online, *Table S6*. The behaviour of the model was also assessed on examples belonging to C_0 as it would be expected that the models predict a probability close to one for example in this class. A total of six models, listed in *Table 1* were evaluated. The threshold on the RF probabilistic output was defined on the ROC curves obtained in cross-validation and so that $Sp = 0.95$. Indeed, having some false positive (FP) predictions is acceptable because the consequences of interventions such as change in lifestyle, increases of some drugs to better treat risk factors such as diabetes will be minimal. Yet, the number of FP needs to be reasonable otherwise it would overburden the healthcare system with unnecessary follow-up examinations.

Survival analysis

We also performed Kaplan–Meier incidence-free survival analysis with the patients included in the test set and considering a new AF as the clinical endpoint. In order to build a per patient model (versus per recording), we only considered the model prediction for the baseline, AF free, 12-lead ECG ($n = 26\ 982$). Patients who did not developed new AF were censored at the most recent encounter. A Cox proportional hazard ratio (HR) model was trained using as input variable the AF risk prediction model-predicted probability output. We computed the area under the curve (AUC) for 5 years of new AF risk prediction, adjusted for age and gender. In a subsequent model, we adjusted for comorbidities and cardiovascular risk factors variables. In a first analysis, the patients were divided in five groups, according with quintiles of probability output of the ML model: (i) those with probability less than 0.2; (ii) those with probability between 0.2 and 0.4; (iii) those with probability between 0.4 and 0.6; (iv) those with probability between 0.6 and 0.8; and (v) those with probability greater than 0.8. In a second analysis, we considered two groups, based on the decision threshold found for the model.

Results

Overall, 53 505 out of 1 095 942 recordings (i.e. 4.9%) were excluded from the database after the signal quality step (*Figure 1*). *Table 2* summarizes the META data statistics for the patient's baseline recording. The patients in C_1 (72.0 ± 13.17) were significantly older than those in C_2 (55.0 ± 16.27) and there were significantly more males in C_1 (51.55%) vs. C_2 (37.61%).

Statistical analysis of the features

The Mann–Whitney rank test performed for individual features for C_1 and C_2 rejected the null hypothesis for 360 out of 445 MOR features, for 14 out of 16 HRV features and 13 out of 15 META features. Among the HRV features the minRR and PNN50 yielded the lowest *P*-values. Among the MOR features the Q, ST, J, and PR yielded the lowest *P*-values. Violin plots in Supplementary material online, *Figures S5 and S6* show the distributions of the HRV and MOR features with the lowest *P*-values. The median and standard deviation for HRV features and the top-20 MOR features are presented in Supplementary material online, *Tables S7 and S8*, respectively. META features are summarized in Supplementary material online, *Table S9*.

Table 2 The table summarizes the META data statistics

	C_0	C_1	C_2	*P*-value
Age (years), mean (SD)	72.00 (12.98)	72.00 (13.17)	55.00 (16.27)	<0.001*
Amiodarone, *n* (%)	197 (2.73)	165 (2.53)	2013 (0.51)	<0.001*
Diuretics, *n* (%)	2334 (32.3)	2072 (31.78)	85 025 (21.36)	<0.001*
Sex, male, *n* (%)	4011 (55.52)	3361 (51.55)	149 755 (37.61)	<0.001*
PAH, *n* (%)	3504 (48.5)	3198 (49.05)	143 902 (36.14)	<0.001*
Chagas, *n* (%)	432 (5.98)	384 (5.86)	9634 (2.42)	<0.001*
Beta blockers, *n* (%)	878 (12.15)	740 (11.35)	31 419 (7.89)	<0.001*
Family CHD, *n* (%)	1005 (13.91)	923 (14.16)	50 140 (12.59)	<0.001*
MI, *n* (%)	121 (1.67)	95 (1.46)	3207 (0.81)	<0.001*
COPD, *n* (%)	86 (1.19)	57 (0.87)	2423 (0.61)	<0.001*
DM, *n* (%)	561 (7.76)	542 (8.31)	28 633 (7.19)	<0.001*
Obesity, *n* (%)	335 (4.64)	391 (6.0)	22 868 (5.74)	0.001*
Dyslipidaemia, *n* (%)	287 (3.97)	268 (4.11)	15 363 (3.86)	0.031*
Calcium blockers, *n* (%)	0 (0.0)	0 (0.00)	95 (0.01)	0.155
Smoking, *n* (%)	367 (5.08)	387 (5.94)	24 627 (6.19)	0.486

It includes a summary for the baseline recordings in C_0, C_1, C_2. With examples in C_0 being AF, examples in C_1 developing AF within 5 years and examples in C_2 not developing AF within 5 years. Mann–Whitney rank test was performed between C_1 and C_2. Features ordered from most to least significant according to their *P*-value. The sign * indicates statistical significance (Mann–Whitney rank test, $P < 0.05$).

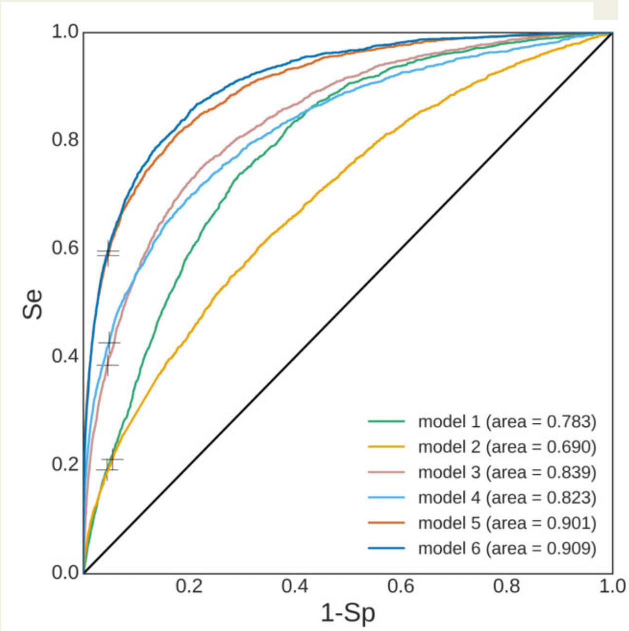


Figure 3 ROC curves for RF classifiers. The '+' symbol on each curve marks the Se-Sp trade-off for the selected probability decision threshold.

Classification

Using mRMR, a subset of the features was selected for models 3–6 (*Table 1*). The RF classifier performed the best for model 6 with AUROC = 0.909 [95% confidence interval (CI): 0.903–0.914] on the test set. The decision threshold at 0.387 was found so that Sp was 0.95 on the validation sets. *Figure 3* shows the ROC curves for the RF classifiers for models 1–6. Results are shown on the test set that consists of 205 730 recordings among which 2162 that belonged to C_1 and 203 568 to C_2. *Table 3* summarizes the performance statistics of the RF model. *Figure 4* shows the probability distribution outputted by model 6 for all examples belonging to the test set and for classes C_0, C_1, and C_2. The probability distribution for C_1 was significantly ($P < 0.05$) higher than for class C_2.

Features importance

Supplementary material online, *Figure S7* shows the feature importance ranking for the top-30 features of the RF classifier for model 6. The most important feature was age. Overall the top-30 features included 1 META feature, 3 HRV features, 9 MOR features, and 17 DNN features. Looking at the top-100 features, these included 2 META features, 3 HRV features, 23 MOR features, and 72 DNN features. Supplementary material online, *Figure S8* shows the feature importance for model 6 while the colours reflect the category a feature belongs to (i.e. MOR, HRV, META or DNN). Supplementary material online, *Figure S9* shows the feature importance ranking for the top-20 features of the RF classifier for all models.

Table 3 **Performance statistics for the RF classifier models (1–6 refer to *Table 1*) for the training, validation and test sets**

Model #	**1**		**2**		**3**		**4**		**5**		**6**	
	Valid	**Test**	**Valid**	**Test**	**Valid**	**Test**	**Valid**	**Test**	**Valid**	**Test**	**Valid**	**Test**
Acc	0.942±0.000	0.762	0.943±0.002	0.937	0.944±0.000	0.948	0.944±0.000	0.945	0.946±0.000	0.949	0.946±0.000	0.949
F1-score	0.072±0.003	0.071	0.072±0.012	0.066	0.137±0.010	0.135	0.143±0.002	0.141	0.187±0.005	0.196	0.189±0.002	0.199
Se	0.208±0.009	0.192	0.203±0.031	0.211	0.408±0.031	0.385	0.429±0.006	0.427	0.573±0.018	0.588	0.581±0.009	0.596
Sp	0.950±0.000	0.955	0.951±0.002	0.945	0.950±0.000	0.954	0.950±0.000	0.951	0.950±0.000	0.953	0.950±0.000	0.953
PPV	0.044±0.002	0.043	0.044±0.007	0.039	0.082±0.006	0.082	0.086±0.001	0.085	0.111±0.003	0.118	0.113±0.001	0.119
NPV	0.991±0.000	0.991	0.991±0.000	0.991	0.993±0.000	0.993	0.993±0.000	0.994	0.995±0.000	0.995	0.995±0.000	0.996
AUROC	0.774±0.011	0.783	0.676±0.028	0.690	0.840±0.019	0.839	0.820±0.002	0.823	0.894±0.006	0.901	0.900±0.004	0.909
		(0.773 – 0.790)		(0.684 - 0.706)		(0.826 - 0.843)		(0.814 - 0.833)		(0.893 - 0.906)		(0.903–0.914)

For train and validation, the mean±standard deviation is reported. For test AUROC, 95% CI is reported as (lower, upper) near the reported score. Best AUROC performance is obtained for model 6 (underlined).

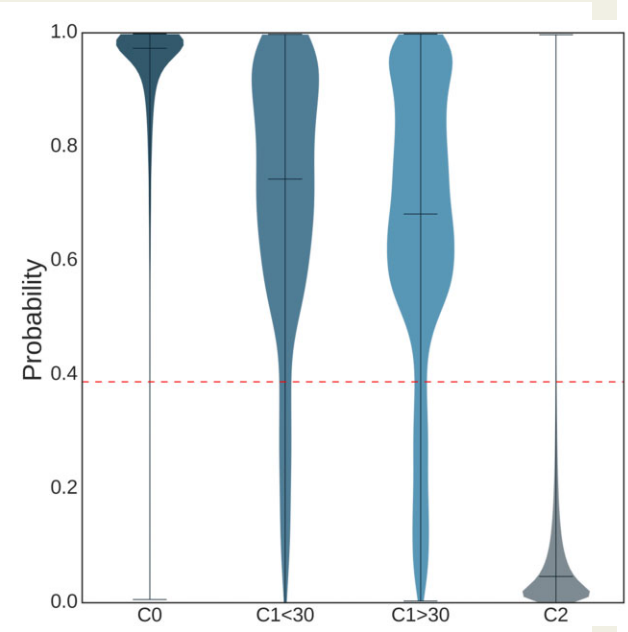


Figure 4 Violin plot showing the probability distribution for the output of the RF classifiers according to the reference class label, for 13 739 recordings in C_0, 214 recordings in C_1<30, 1948 recordings in C_1>30 and 203 568 recordings in C_2. The dotted red horizontal line depicts the probability decision threshold to classify an example as future AF or not.

Error analysis

Considering model 6, there was a total of 9517 FP and 873 false negative (FN). Violin plots representing the probability outputs for all FP per arrhythmia class are presented in Supplementary material online, *Figure S10*. In particular, we observed that a large relative proportion of examples with 1dAVb (18.0%) and left bundle branch block (LBBB) (14.8%) rhythms were FP. The mean ± standard deviation(SD) of time between the examination and a follow-up positive AF examination for true positive (TP) is 1.48 ± 1.23 years. Mean ± SD for FN is 1.57 ± 1.20 years. A cardiologist (G.M.) with 5 years of clinical practice carefully reviewed 100 TP recordings within the test set blinded to any clinical or diagnosis information. Among the 100 TP, the cardiologist identified 50 recordings with a form of hypertrophy [Left ventricular hypertrophy (LVH)/Right ventricular hypertrophy (RVH)/Left atrial hypertrophy (LAH)/LVH], 24 recordings with AF/AFL, 16 recordings with premature atrial contraction (PAC)/premature ventricular contraction (PVC), 10 recordings with a form of bundle branch block (BBB) (LBBB/right bundle branch block), 25 recordings with other cardiac abnormalities, 4 recordings that were too noisy and could not be interpreted, and 2 recordings were normal sinus rhythm (NSR). It is important to note that a recording may present multiple cardiac abnormalities.

Survival analysis

The risk of development of AF for the two evaluated settings are shown in *Table 4* and adjusted survival curves in *Figure 5*. In (*A*) the model indicates that patients with probability >0.8 had higher risk of development AF (HR 43.77, 95% CI: 36.19–52.93; $P < 0.001$), considering as the reference those with probability less than 0.2. For b) the survival curve for probabilities >0.387 had a higher risk of development AF (HR 10.72, 95% CI: 9.39–12.24; $P < 0.001$), compared to those with probability less or equal to the decision threshold. The Cox models adjusted by age and gender present a good performance in the prediction of new AF risk prediction, with an AUC of 0.87 (95% CI: 0.86–0.90) and (*B*) 0.84 (95% CI: 0.82–0.86). The adjusted survival curves for the model by different selections of comorbidities and cardiovascular risk factors are given in Supplementary material online, *Table S10*. This analysis had similar results compared to the models adjusted by age and gender.

Table 4 Risk of new AF

	HR	CI 95%		P − value
Adjusted by age and gender				
prob (0.2, 0.4]	5.222	4.343	6.279	<0.001
prob (0.4, 0.6]	12.063	9.933	14.651	<0.001
prob (0.6, 0.8]	20.967	17.053	25.779	<0.001
prob (0.8, 1]	43.764	36.186	52.928	<0.001
prob (0.387, 1]	10.719	9.389	12.237	<0.001

The table shows the hazard ratios (HRs) with for the different probability groups. The models were adjusted by different selection of variables. For the quintiles, the reference class is prob (0, 0.2]; for the dichotomic analysis, the reference class is (0, 0.387].

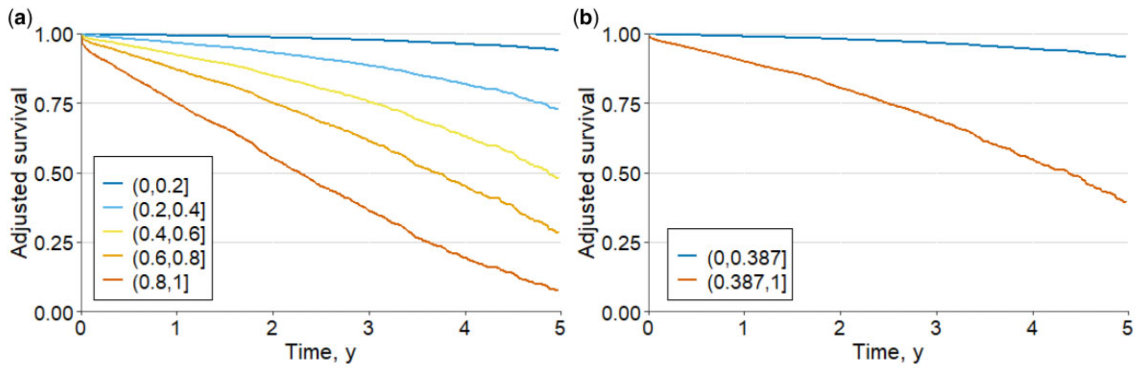
4. Discussion

Models performance

The results indicate that the ML models are indeed learning residual features that are predictive of the future development of AF. The best model was combining features from all modalities (META, HRV, MOR, and DNN) with AUROC = 0.909 (95% CI : 0.903 − 0.914). This highlights the value in combining engineered features that encapsulate Human knowledge in electrophysiology acquired over the past two centuries, DNN features which are less interpretable but enable to let the model to include features that are not intelligible to an human observer but are yet important for the task at hand, and finally metadata which encode demographic and medical information about the patient. The test set *Se* for individuals in C_0 was 0.970 and for individuals in C_1 it was 0.596 when using the decision threshold of 0.387. Although we found a high gap between training and test set performances, this is not uncommon when using RF and the performance between validation sets and the test set were close as seen on *Table 3*. Furthermore, the Cox analysis showed that for different groups of probability results from the ML model further indicated that the highest probability resulted in the highest risk of AF future development (see *Table 4* and *Figure 5*).

Error analysis

The cardiologist review of 100 TP examples highlighted that a large number (24%) of the recordings correctly predicted as future AF were actually AF at the time of the examination and thus were likely



No. at risk

(0,0.2]	22095	13810	7492	3557	1398	542
(0.2,0.4]	2947	1842	1010	490	184	63
(0.4,0.6]	1072	684	396	190	85	28
(0.6,0.8]	517	329	180	98	30	5
(0.8,1]	396	233	142	75	28	2

No. at risk

(0,0.387]	24930	15579	8466	4031	1575	603
(0.387,1]	2097	1319	754	379	150	37

Figure 5 Adjusted survival curves for the risk of new AF. The plots display the survival curves for the different cohorts. The curves were computed from probability results of ML model. The models were adjusted by age and gender. In (*A*) the model with quintiles of probability and (*B*) the model with probability less or equal than 0.387 and those with probability greater than 0.387.

misdiagnosed. This number is coherent with the sensitivity of 0.769 reported in Ribeiro *et al.*[10] for AF diagnosis by a cardiologist. Within this context, we foresee that our model may support the diagnosis of AF and avoid a significant number of misdiagnosis. Our second observation is that a number of TP had a type of hypertrophy, BBB or premature beats. This observation is coherent with previous reports that hypertrophy,[18] ectopic beats,[19] and BBB[20] are predictive of AF. Examples of three TP are displayed in *Figure 6*.

Clinical usability of the new ML model

The best model obtained $\mathrm{AUROC} = 0.909$ $(95\%\ \mathrm{CI}:\ 0.903 - 0.914)$, $\mathrm{Se} = 0.596$, and $\mathrm{Sp} = 0.953$. Practically speaking, with such a high Sp, and considering that a significant number of FP have another cardiac abnormality, this means that the number of FP is reasonable since intervention in FP will have limited consequences and the added need from the healthcare system is acceptable. The *Se* is more modest, yet it means that >50% of patients that will develop (or have a missed diagnosis of) AF can be identified which is a significant added value for the better management of these patients. Since examples in C_0 have AF then it is expected that the model will predict a high probability of AF risk and this was confirmed by our observations (*Figure 4*) where *Se* for individuals in C_0 was 0.970. Furthermore, when separating C_1 into two subclasses C_1(<30days) and C_1(≥30 days) we observed that the model indeed had a significantly higher probability distribution for C_1 (<30 days) than C_1(>30 days)—see *Figure 4* for the violin plot. In the work by Attia *et al.*[21] individuals that developed AF within 31 days were considered as actually having AF at the baseline recording time but likely paroxysmal so that AF was not exerted at this time.

Overall, our findings are very promising. Predicting AF in a 12-lead ECG, a low cost and widely available exam, can help physicians in decision making. The model prediction could be used in the following fashion: patients with a high probability of future AF should first undergo a careful review for the presence of AF. Eventually a second 12-lead ECG may be acquired at the same examination date in the case of doubt. This should result in lowering the number of misdiagnosis. Patients with no visible AF but with a high probability of future AF are at risk of imminent appearance of AF in a likely already diseased patient with paroxysmal AF and thus should be referred for a Holter ECG. If no Holter is available then a follow-up 12-lead ECG within the next few days should be scheduled. Individuals with no visible AF but predicted as future AF have a high risk to develop AF within the next 5 years and should regularly be examined. Although the model prediction has a good performance, the results are preliminary and it is still premature to make any changes in the current AF recommendations. Further research and external validation are necessary, but our data are a step forward in the field of AF screening and diagnosis.

Features importance

Through the analysis of feature importance in model 3, that includes features encoding prior known physiological information, we confirm the importance of the PR interval length[22,23] and importance of QRS width and the R-wave amplitude as predictors of AF. AF is associated with functional or structural changes in the atrial myocardium.[24] In the ECG, these changes are often recognized in the P-wave abnormalities. Therefore, abnormal P-wave duration, P-axis and beat-to-beat P-wave variability have been described as AF predictors.[25] The association of the PR interval with AF is controversial in the literature. PR interval components (P-wave onset to P-wave peak duration, P-wave peak to P-wave end duration and PR segment) have different impacts on AF prediction. A prolongation of the PR-interval will be predictive of AF if the prolongation is mainly due to the P-wave duration. This also explains the observation that a short PR-segment is more predictive of AF, as in this case, the P-wave duration would contribute more to the overall PR interval.[26] Although AF is part of a

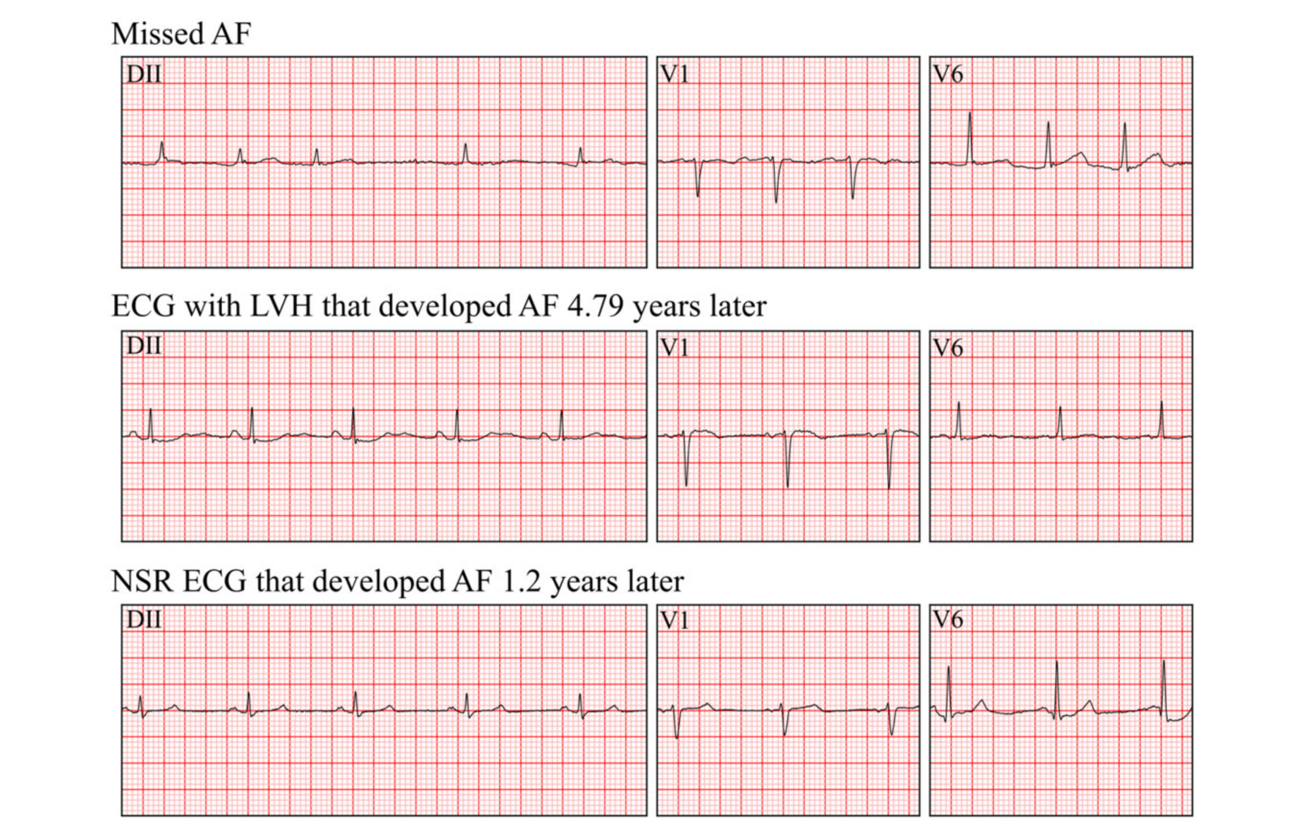

Missed AF
DII
V1
V6
ECG with LVH that developed AF 4.79 years later
DII
V1
V6
NSR ECG that developed AF 1.2 years later
DII
V1
V6

Figure 6 Example for three recordings from lead DII, V1, V6. All were re-annotated by cardiologist with 5 years of experience.

syndrome called atrial cardiomyopathy, some ventricular ECG abnormalities have been reported to be indirect predictors of atrial disease, QRS duration is a marker of structural modifications and ventricular remodelling, therefore, can be associated with left atrial size, supporting the concept of simultaneous processes in the atria and ventricles. Prolonged QRS duration reflected as fragmented QRS[27] and left ventricular hypertrophy[18] have been associated with AF. Sokolow-Lyon voltage product $([SV1 + RV5/\ RV6]*$ QRS duration $\geq$ 371 000 µVms) was related to incident AF, showing that both R-wave and QRS duration have an impact on AF prediction. When combining all features with model 6, age was the most important feature which is consistent with the knowledge that the likelihood of having AF significantly increases as a function of age.[28]

Limitations

The main limitation of our study relates to the lack of evaluation of our model on a longitudinal cohort that was consistently and regularly followed-up for years, i.e. with a baseline examination and then followed-up regular visits for all the cohort. Also, it will be important to evaluate the models performance on external datasets in order to assess their generalization performances. Although it may not be possible to obtain the exact same metadata for external datasets we may be able to benchmark the performances of models 1–4 which only use the raw ECG signal as input. The second main limitation of our study is the lack of a better mapping with the patient's EMR. Indeed, only a limited number of EMR information were available as metadata and these were self-reported by patients. A better access to EMR variables would also enable benchmarking our model against state-of-the-art AF risk scoring systems such as CHARGE-AF.[29] Our experimental setting including patients with repeated 12-lead ECG recordings and excluding those who did not, has an intrinsic bias towards individuals already having a cardiac abnormality or at risk for a cardiovascular disease. The clinical endpoints used in this study was documented AF development within 5 years. With documented corresponding to a future 12-lead ECG examination with a positive AF diagnosis. This label is not necessarily a gold standard as there exists a number of possible bias in that registration such as individuals that would develop silent AF after their latest documented visit but would not perform an additional 12-lead ECG examination. Finally, the model needs to be further validated on individuals presenting no cardiac abnormality at the time of the baseline examination but who later on developed AF.

Conclusion

It is the first study integrating feature engineering, deep learning and EMR metadata to create a risk prediction tool for the management of patients at risk of AF. This new hybrid model may also be extended for the risk prediction of other cardiac pathologies. The high performance obtained suggest that structural changes in the 12-lead ECG are associated with existing or impending AF. Our study has important clinical implications for AF management. The system has the potential to be rapidly incorporated in the clinical practice, helping to detect those prone to develop AF and better manage these at-risk patients to prevent complications. For that purpose further validation of the model on the external test set longitudinal cohort is needed.

Supplementary material

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

Funding

This research was supported for S.B., S.G., and J.A.B. by a grant (3-17550) from the Ministry of Science & Technology, Israel & Ministry of Europe and Foreign Affairs (MEAE) and the Ministry of Higher Education, Research and Innovation (MESRI) of France. A.L.P.R. is supported in part by CNPq (310679/2016-8 and 465518/2014-1) and by FAPEMIG (PPM-00428-17 and RED-00081-16). C.M. is supported by CNPq (381810/2020-8). A.H.R. is partially supported by CAPES (no 88887.474452/2020-00) and by the Kjell och Märta Beijer Foundation.

Conflict of interest: none declared.

Data availability

Researchers affiliated to educational or research institutions might make requests to access the CODE dataset. Requests should be made to the corresponding author of this paper. They will be forwarded and considered on an individual basis by the Telehealth Network of Minas Gerais. If approved, any data use will be restricted to non-commercial research purposes. The data will only be made available on the execution of appropriate data use agreements.

References

1. Anderson KM, Odell PM, Wilson PWF, Kannel WB. Cardiovascular disease risk profiles. *Am Heart J* 1991;**121**:293–298.
2. Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998; **97**:1837–1847.
3. Brindle P, Beswick A, Fahey T, Ebrahim S. Accuracy and impact of risk assessment in the primary prevention of cardiovascular disease: a systematic review. *Heart* 2006;**92**:1752–1759.
4. Haim M, Hoshen M, Reges O, Rabi Y, Balicer R, Leibowitz M. Prospective national study of the prevalence, incidence, management and outcome of a large contemporary cohort of patients with incident non-valvular atrial fibrillation. *J Am Heart Assoc* 2015;**4**:1–12.
5. Wolf PA, Abbott RD, Kannel WB. Atrial fibrillation as an independent risk factor for stroke: the Framingham study. *Stroke* 1991;**22**:983–988.
6. Sörnmo L, Petrėnas A, Marozas V. *Atrial Fibrillation from an Engineering Perspective*. Berlin: Springer, 2018.
7. Wang Q-C, Wang Z-Y. Big data and atrial fibrillation: current understanding and new opportunities. *J Cardiovasc Transl Res* 2020;**13**:944–952.
8. Christopoulos G, Graff-Radford J, Lopez CL, Yao X, Attia ZI, Rabinstein AA, Petersen RC, Knopman DS, Mielke MM, Kremers W, Vemuri P, Siontis KC, Friedman PA, Noseworthy PA. Artificial intelligence-electrocardiography to predict incident atrial fibrillation: a population-based study. *Circ Arrhythmia Electrophysiol* 2020;**13**:e009355.
9. 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 ECG and Help Identify Those at Risk of Atrial Fibrillation-Related Stroke. Circulation *Lippincott Williams and Wilkins; 2021;1287–1298.*
10. Ribeiro ALP, Paixão GMM, Gomes PR, Ribeiro MH, Ribeiro AH, Canazart JA, Oliveira DM, Ferreira MP, Lima EM, Moraes JL de, Castro N, Ribeiro LB, Macfarlane PW., .. Tele-electrocardiography and bigdata: the CODE (Clinical Outcomes in Digital Electrocardiography) study. *J Electrocardiol*Churchill Livingstone Inc.; 2019;**57**:S75–S78.
11. Ribeiro AH, Ribeiro MH, Paixão GMM, Oliveira DM, Gomes PR, Canazart JA, Ferreira MPS, Andersson CR, Macfarlane PW, Wagner M, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun* 2020;**11**:1–9.
12. Kligfield P, Gettes LS, Bailey JJ, Childers R, Deal BJ, Hancock EW, van Herpen G, Kors JA, Macfarlane P, Mirvis DM, Pahlm O, Rautaharju P, Wagner GS; American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; American College of Cardiology Foundation; Heart Rhythm Society, M Josephson, JW Mason, P Okin, B Surawicz, H. Wellens Recommendations for the standardization and interpretation of the electrocardiogram: part I: The electrocardiogram and its technology: a scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society: endorsed by the International Society for Computerized Electrocardiology. *Circulation* 2007;**115**:1306–1324.
13. Chocron A, Oster J, Biton S, Franck M, Elbaz M, Zeevi YY, Behar J. Remote atrial fibrillation burden estimation using deep recurrent neural network. *IEEE Trans Biomed Eng* 2021;**68**:2447–2455.
14. Assaraf D, Levy J, Singh J, Chocron A, Behar JA. Classification of 12-lead ECGs using digital biomarkersand representation learning. *Comput Cardiol* 2020: 47.
15. Pablo Martínez J, Almeida R, Olmos S, Rocha AP, Laguna PA Wavelet-based ECG delineator: evaluation on standard databases. *IEEE Trans Biomed Eng* 2004; **51**: 570–581.
16. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol* 2005;**3**:185–205.
17. Head T, MechCoder GL, Shcherbatyi I, others. scikit-optimize/scikit-optimize: v0.5.2. Zenodo 2018;,
18. Akkaya M, Higuchi K, Koopmann M, Burgon N, Erdogan E, Damal K, Kholmovski E, McGann C, Marrouche NF. Relationship between left atrial tissue structural remodelling detected using late gadolinium enhancement MRI and left ventricular hypertrophy in patients with atrial fibrillation. *Europace* 2013;**15**:1725–1732.
19. German DM, Kabir MM, Dewland TA, Henrikson CA, Tereshchenko LG. Atrial fibrillation predictors: importance of the electrocardiogram. *Ann Noninvasive Electrocardiol* 2016;**21**:20–29.
20. Nielsen JB, Olesen MS, Tangø M, Haunsø S, Holst AG, Svendsen JH. Incomplete right bundle branch block: a novel electrocardiographic marker for lone atrial fibrillation. *Europace* 2011;**13**:182–187.
21. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, Carter RE, Yao X, Rabinstein AA, Erickson BJ 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**:861–867.
22. Nielsen JB, Pietersen A, Graff C, Lind B, Struijk JJ, Olesen MS, Haunsø S, Gerds TA, Ellinor PT, Køber L et al. Risk of atrial fibrillation as a function of the electrocardiographic PR interval: results from the Copenhagen ECG Study. *Heart Rhythm* 2013;**10**:1249–1256.
23. Bidstrup S, Olesen MS, Svendsen JH, Nielsen JB. Role of PR-interval in predicting the occurrence of atrial fibrillation. *J Atr Fibrillation* 2013;**6**:90–94.
24. Goette A, Kalman JM, Aguinaga L, Akar J, Cabrera JA, Chen SA, Chugh SS, Corradi D, D'Avila A, Dobrev D et al. EHRA/HRS/APHRS/SOLAECE expert consensus on atrial cardiomyopathies: definition, characterization, and clinical implication. *Europace* 2016;**18**:1455–1490.
25. Conte G, Luca A, Yazdani S, Caputo ML, Regoli F, Moccetti T, Kappenberger L, Vesin J-M, Auricchio A. Usefulness of P-wave duration and morphologic variability to identify patients prone to paroxysmal atrial fibrillation. *Am J Cardiol* 2017;**119**:275–279.
26. Smith JW, O'Neal WT, Shoemaker MB, Chen LY, Alonso A, Whalen SP, Soliman EZ. PR-interval components and atrial fibrillation risk (from the Atherosclerosis Risk in Communities Study). *Am J Cardiol* 2017;**119**:466–472.
27. Laureanti R, Conte G, Corino VDA, Osswald S, Conen D, Roten L, Rodondi N, Ammann P, Meyer-Zuern CS, Bonati L et al. Sex-related electrocardiographic differences in patients with different types of atrial fibrillation: results from the SWISS-AF study. *Int J Cardiol* 2020;**307**:63–70.
28. Feinberg WM, Blackshear JL, Laupacis A, Kronmal R, Hart RG. Prevalence, age distribution, and gender of patients with atrial fibrillation: analysis and implications. *Arch Intern Med* 1995;155(5):469–73
29. Alonso A, Krijthe BP, Aspelund T, Stepas KA, Pencina MJ, Moser CB, Sinner MF, Sotoodehnia N, Fontes JD, Janssens AC Kronmal RA.. Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the CHARGE-AF consortium. *Journal of the American Heart Association* 2.2 2013: e000102.