European Heart Journal - Digital Health (2021) **2**, 626–634
doi:10.1093/ehjdh/ztab080

ORIGINAL ARTICLE

# ECG-AI: electrocardiographic artificial intelligence model for prediction of heart failure

**Oguz Akbilgic [1,2]*, Liam Butler [1], Ibrahim Karabayir [1,3], Patricia P. Chang[2], Dalane W. Kitzman[2], Alvaro Alonso[5], Lin Y Chen[6], and Elsayed Z. Soliman [2,7]**

[1]Department of Health Informatics and Data Science, Parkinson School of Health Sciences and Public Health, Loyola University Chicago, 2160 S 1st Street, Maywood, IL 60153, USA; [2]Sections on Cardiovascular Medicine and Geriatrics, Department of Internal Medicine, Wake Forest School of Medicine, 475 Vine Street, Winston-Salem, NC 27101, USA; [3]Departmet of Econometrics, Kirklareli University, 3 Kayalı Kampüsü Kofçaz, Kirklareli, Turkey; [4]Department of Medicine, Division of Cardiology, University of North Carolina at Chapel Hill, 160 Dental Circle, Chapel Hill, NC 27599, USA; [5]Department of Epidemiology, Rollins School of Public Health, Emory University, 1518 Clifton Rd. NE Atlanta, GA, 30322, USA; [6]Cardiovascular Division, Department of Medicine, University of Minnesota Medical School, 401 East River Parkway, Minneapolis, MN 55455, USA; and [7]Internal Medicine, Epidemiological Cardiology Research Center, Sections on Cardiovascular Medicine, Wake Forest School of Medicine, 525 Vine Street, Winston-Salem, NC 27101, USA

*Received 14 June 2021; revised 19 August 2021; editorial decision 31 August 2021; accepted 1 September 2021; online publish-ahead-of-print 9 October 2021*

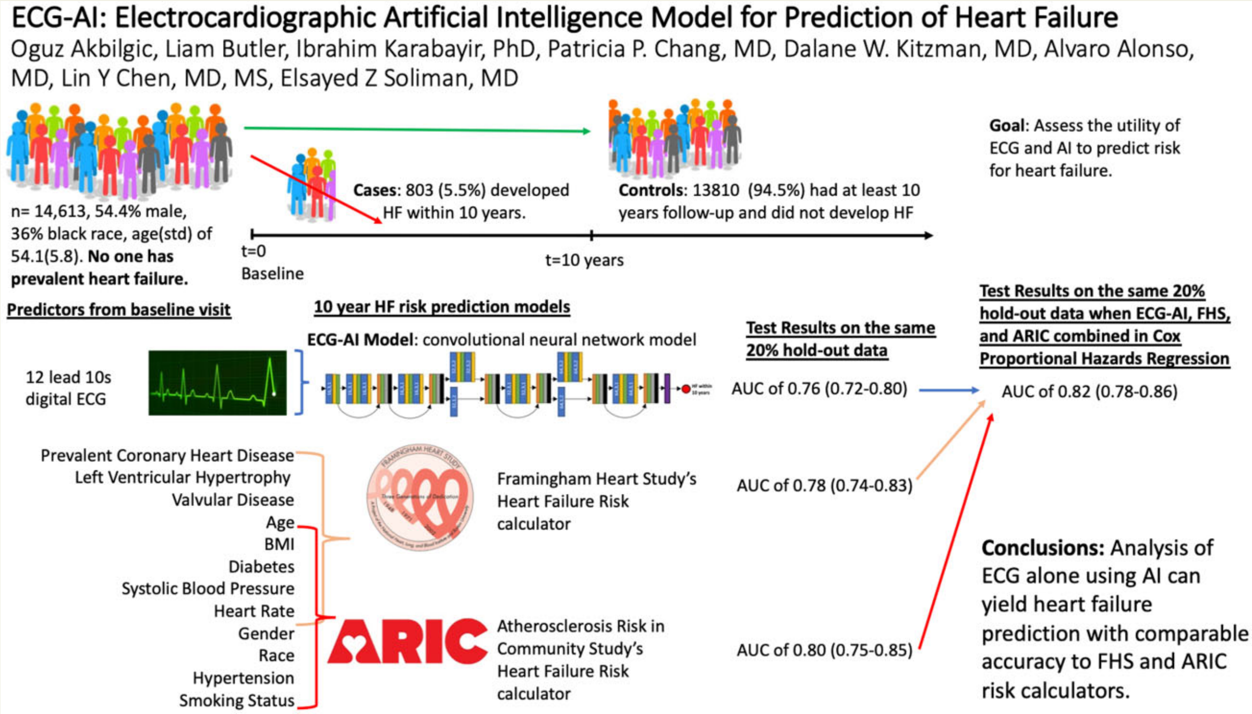
**Aims** Heart failure (HF) is a leading cause of death. Early intervention is the key to reduce HF-related morbidity and mortality. This study assesses the utility of electrocardiograms (ECGs) in HF risk prediction.

**Methods and results** Data from the baseline visits (1987–89) of the Atherosclerosis Risk in Communities (ARIC) study was used. Incident hospitalized HF events were ascertained by ICD codes. Participants with good quality baseline ECGs were included. Participants with prevalent HF were excluded. ECG-artificial intelligence (AI) model to predict HF was created as a deep residual convolutional neural network (CNN) utilizing standard 12-lead ECG. The area under the receiver operating characteristic curve (AUC) was used to evaluate prediction models including (CNN), light gradient boosting machines (LGBM), and Cox proportional hazards regression. A total of 14 613 (45% male, 73% of white, mean age ± standard deviation of 54 ± 5) participants were eligible. A total of 803 (5.5%) participants developed HF within 10 years from baseline. Convolutional neural network utilizing solely ECG achieved an AUC of 0.756 (0.717–0.795) on the hold-out test data. ARIC and Framingham Heart Study (FHS) HF risk calculators yielded AUC of 0.802 (0.750–0.850) and 0.780 (0.740–0.830). The highest AUC of 0.818 (0.778–0.859) was obtained when ECG-AI model output, age, gender, race, body mass index, smoking status, prevalent coronary heart disease, diabetes mellitus, systolic blood pressure, and heart rate were used as predictors of HF within LGBM. The ECG-AI model output was the most important predictor of HF.

**Conclusions** ECG-AI model based solely on information extracted from ECG independently predicts HF with accuracy comparable to existing FHS and ARIC risk calculators.

* Corresponding author. Tel: +1 708 216 8971, Fax: +1 708 216 8216, Email: oakbilgic@luc.edu

## Graphical Abstract



**ECG-AI: Electrocardiographic Artificial Intelligence Model for Prediction of Heart Failure**

Oguz Akbilgic, Liam Butler, Ibrahim Karabayir, PhD, Patricia P. Chang, MD, Dalane W. Kitzman, MD, Alvaro Alonso, MD, Lin Y Chen, MD, MS, Elsayed Z Soliman, MD

n= 14,613, 54.4% male, 36% black race, age(std) of 54.1(5.8). **No one has prevalent heart failure.**

**Cases:** 803 (5.5%) developed HF within 10 years.

**Controls:** 13810 (94.5%) had at least 10 years follow-up and did not develop HF

t=0
Baseline

t=10 years

**Goal:** Assess the utility of ECG and AI to predict risk for heart failure.

**Predictors from baseline visit**

**10 year HF risk prediction models**

**Test Results on the same 20% hold-out data**

**Test Results on the same 20% hold-out data when ECG-AI, FHS, and ARIC combined in Cox Proportional Hazards Regression**

12 lead 10s digital ECG

**ECG-AI Model:** convolutional neural network model

AUC of 0.76 (0.72-0.80)

AUC of 0.82 (0.78-0.86)

Prevalent Coronary Heart Disease
Left Ventricular Hypertrophy
Valvular Disease
Age
BMI
Diabetes
Systolic Blood Pressure
Heart Rate
Gender
Race
Hypertension
Smoking Status

Framingham Heart Study's Heart Failure Risk calculator

AUC of 0.78 (0.74-0.83)

ARIC

Atherosclerosis Risk in Community Study's Heart Failure Risk calculator

AUC of 0.80 (0.75-0.85)

**Conclusions:** Analysis of ECG alone using AI can yield heart failure prediction with comparable accuracy to FHS and ARIC risk calculators.

**Keywords** Heart failure • ECG • Electrocardiogram • Deep learning • Artificial intelligence • ARIC

### Translational perspective

This study investigates whether electrocardiogram (ECG) alone, when processed via artificial intelligence, can accurately predict the risk of heart failure (HF). ECG-artificial intelligence deep learning models using only standard 10 s 12-lead ECG data from 14 613 participants from the Atherosclerosis Risk in Communities (ARIC) study cohort could predict future HF with comparable accuracy to the HF risk calculators from ARIC study and Framingham Heart Study. Artificial intelligence is capable of using ECG tracings to predict incident HF. This can also enable pre-screening of large patient populations for risk of HF remotely when adapted into smartwatches with ECG functionality.

# Introduction

There are ∼6.5 million adults, with more than 550 000 yearly diagnoses, in the USA that are reported to suffer or have suffered heart failure (HF).[1,2] Heart failure is a progressive complex condition that is often terminal and is a major public health concern and burden.[3,4] Heart failure often results in structural or functional cardiac disorders that impair the pumping of blood between cardiac compartments and the rest of the body.[5] Early signs and symptoms of HF can substantially vary between different groups of patients, which can further complicate diagnosis and treatment.[5,6]

Heart failure was mentioned in 13.4% of all death certificates in the USA in 2018. While there have been advances in diagnoses and management, outcomes in patients with HF are still largely variable, and risks among different subgroups can substantially change over time.[2,6] Early diagnosis and treatment can significantly improve HF prognosis,[7] and subsequently help reduce the health and economic burdens of HF. Although HF therapy has somewhat improved survival rates, greater efforts are needed toward early detection of cardiac disorders and prevention of HF.[8] Thus, it is essential to develop HF pre-screening tools that rely on a minimal amount of data that are easy to obtain, low cost with accessibility and promise of future remote applications. At this point, better utilization of electrocardiograms (ECGs) beyond their current clinical use and interpretation has the potential to lead to the development of such HF pre-screening tools.

Several recent studies have shown the utility of artificial intelligence (AI) applied to digital ECGs (time-voltage signal) in detecting and predicting cardiovascular disease. Specifically, such AI models utilizing digital ECGs were used in prediction of atrial fibrillation,[9]

**Table 1 Study cohort characteristics and risk factors**

| **Risk factors** | ***n* (%) or mean (SD)** | | **$\chi^2$ or *T*-test** |
|---|---|---|---|
| | **HF in 10 years (*n* = 13 810)** | **HF in 10 years (*n* = 803)** | ***P*-values** |
| Gender (male)[a] | 6179 (44.7) | 456 (57.2) | <0.001 |
| Race (Black)[a] | 3559 (25.8) | 289 (36.0) | <0.001 |
| Age at visit 1[a,b] (years) | 53.9 (5.7) | 57.2 (5.2) | <0.001 |
| BMI (kg/m$^2$)[a,b] | 27.4 (5.2) | 29.5 (6.3) | <0.001 |
| Smoking status[a] | | | <0.001 |
| Former | 4407 (31.9) | 284 (35.4) | |
| Current | 3485 (25.2) | 304 (37.9) | |
| Prevalent coronary heart disease[b] | 458 (3.3) | 138 (17.2) | <0.001 |
| Diabetes mellitus[a,b] | 1326 (9.6) | 286 (35.6) | <0.001 |
| Systolic blood pressure (mmHg)[a,b] | 120.5 (18.4) | 131.2 (22.9) | <0.001 |
| Hypertension medication[a] | 3566 (25.8) | 420 (52.3) | <0.001 |
| Left ventricular hypertrophy[b] | 253 (1.9) | 50 (6.4) | <0.001 |
| Valvular disease[b] | 33 (0.2) | 9 (1.1) | <0.001 |
| Heart rate (ventricular, beats per minute)[a,b] | 66.4 (10.0) | 70.5 (12.3) | <0.001 |

ARIC, Atherosclerosis Risk in Communities; FHS, Framingham Heart Study; HF, heart failure; SD, standard deviation.

[a] Variables used in ARIC risk calculator.

[b] Variables used in FHS risk calculator.

cardiomyopathy,[10,11] and all-cause mortality.[12] We hypothesize that standard 10 s 12-lead ECG alone can predict HF risk within 10 years with moderately high accuracy. We utilized data from the *Atherosclerosis Risk in Communities* (ARIC) study cohort to test this hypothesis.

# Methods

## Cohort

The ARIC is an ongoing prospective epidemiologic study conducted in four communities in the USA (Forsyth County, NC; Jackson, MS; Washington County, MD; and the northwest suburbs of Minneapolis, MN) and designed to investigate the aetiology of atherosclerosis and its clinical outcomes, and cardiovascular risk factors associated with demographics, race, gender, and time. From 1987 to 1989 (visit 1, the baseline for our analysis), a total of 15 792 participants (8710 women and 4266 of black race) were enrolled and completed a home interview and clinic visit. In this analysis, we utilized data from visit 1 and follow-up visit 2 to visit 4 (visit 2: 1990–92, visit 3: 1993–96, visit 4: 1996–98) in AI-based models while using the entire follow-up in survival analysis up to 2019.

## Outcomes

Our main outcome was predicting new-onset HF events within 10 years from visit 1 baseline examination. Heart failure was defined by hospitalization and HF as a hospital discharge diagnosis [International Classification of Disease, Ninth Revision, Clinical Modification (ICD-9-CM), code 428], or in-hospital or out-of-hospital deaths attributed to HF (deaths coded as ICD-9-CM code 428 or International Classification of Disease, Tenth Revision, code 150, without a previous record of hospitalization with ICD-9-CM code 428).[13]

## Risk factors

We used a total of 12 risk factors which were used in the ARIC HF risk calculator[8] and Framingham Heart Study (FHS) HF risk calculator.[3] These clinical risk factors included in the 'ARIC' model in this study were gender, race, age, diabetes, hypertension medication, body mass index (BMI; kg/m$^2$), systolic blood pressure (mmHg), prevalent coronary heart disease, smoking status, and heart rate (beats per minute, b.p.m.). The clinical risk factors included in the 'Framingham' model were age, diabetes, BMI (kg/m$^2$), systolic blood pressure (mmHg), prevalent coronary heart disease, heart rate (b.p.m.), left ventricular hypertrophy (LVH), and valvular disease (see *Table 1*) (see Supplementary material online, *Section S1* for details).

## Electrocardiogram data

Raw digital ECG data (time-voltage) for 12 leads from the baseline (visit 1) were used. A supine 12-lead ECG at 250 Hz frequency of 10 s at rest was used. The ECGs were initially obtained from the MAC PC10 personal cardiogram (Marquette Electronics, Milwaukee, WI, USA). In this study, ECG data are used as indicators for possible subclinical HF risk.

## Inclusions/exclusion criteria

All ARIC participants with good quality ECG data at baseline as well as information on all relevant risk factors and HF events during the study's long-term follow-up were eligible for inclusion in this analysis. Participants with prevalent HF (*n* = 739) at the baseline visit, missing HF data during follow-up, and missing or poor-quality ECGs were excluded.

## Study design

We randomly split our study cohort into 80% for model building and 20% as hold-out test data. Heart failure prediction models were built using different machine learning and statistical methods with five-fold cross-validation using the 80% model building dataset. During five steps of five-fold cross-validation, we built five independent models from

**Table 3** Cox proportional hazards regression model modelling heart failure risk

| Covariate | Coefficient | Hazard ratio | 95% CI | *P*-value |
|---|---|---|---|---|
| ECG-AI outcome | 5.05 | 155.61 | 58.93–410.92 | <0.01 |
| Gender | 0.31 | 1.37 | 1.14–1.65 | <0.01 |
| Race | 0.14 | 1.15 | 0.94–1.40 | 0.176 |
| Age | 0.08 | 1.09 | 1.07–1.11 | <0.01 |
| Diabetes | 0.96 | 2.60 | 2.14–3.17 | <0.01 |
| Hypertension medication | 0.49 | 1.62 | 1.35–1.96 | <0.01 |
| BMI | 0.04 | 1.04 | 1.02–1.05 | <0.01 |
| Systolic blood pressure | 0.01 | 1.017 | 1.00–1.01 | <0.01 |
| Prevalent coronary heart disease | 0.89 | 2.44 | 1.89–3.14 | <0.01 |
| Ventricular rate | 0.02 | 1.02 | 1.02–1.03 | <0.01 |
| Left ventricular hypertrophy | 0.35 | 1.42 | 1.00–2.02 | 0.049 |
| Valvular disease | 1.35 | 3.86 | 1.98–7.53 | <0.01 |
| Smoking status | 0.56 | 1.75 | 1.56–1.96 | <0.01 |

BMI, body mass index; CI, confidence interval; ECG-AI, electrocardiographic artificial intelligence.

**Table 4** Response of electrocardiographic artificial intelligence and Cox proportional hazards regression models to follow-up electrocardiograms

| Mean Δrisk with 95% CI as a percentage | Controls | Cases |
|---|---|---|
| ECG-AI model | 0.235 (0.178–0.291) | 1.414 (0.912–1.917) |
| Cox model | 0.061 (0.031–0.0915 | 2.568 (1.883–3.252) |

CI, confidence interval; ECG-AI, electrocardiographic artificial intelligence.

**Table 5** Response of electrocardiographic artificial intelligence and Cox proportional hazards regression models to follow-up electrocardiograms

| Scenarios | ECG time | TP | FP | TN | FN | Specificity | Sensitivity | Negative predictive value | Positive predictive value |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Baseline | 116 | 764 | 1819 | 2 | 0.7042 | 0.9831 | 0.9990 | 0.1318 |
| | Follow-up | 116 | 764 | 1819 | 2 | 0.7042 | 0.9831 | 0.9990 | 0.1318 |
| 2 | Baseline | 108 | 515 | 2068 | 10 | 0.8006 | 0.9153 | 0.9952 | 0.1734 |
| | Follow-up | 116 | 528 | 2055 | 2 | 0.7956 | 0.9831 | 0.9990 | 0.1801 |
| 3 | Baseline | 93 | 261 | 2322 | 25 | 0.8990 | 0.7881 | 0.9893 | 0.2627 |
| | Follow-up | 111 | 258 | 2325 | 7 | 0.9001 | 0.9407 | 0.9970 | 0.3008 |
| 4 | Baseline | 77 | 127 | 2456 | 41 | 0.9508 | 0.6525 | 0.9836 | 0.3775 |
| | Follow-up | 95 | 123 | 2460 | 23 | 0.9524 | 0.8051 | 0.9907 | 0.4358 |

ECG, electrocardiogram; FN, false negative; FP, fasle positive; TN, true negative; TP, true positive.

ECG-AI CNN model which only uses digital 12-lead ECG data alone as input, resulting in an AUC of 0.756 on hold-out dataset, which was not significantly different than the AUC (0.778) of the FHS risk calculator (DeLong test, $P = 0.180$). However, the AUC of the ECG-AI model was lower than the AUC (0.778) of the ARIC risk calculator (DeLong test, $P = 0.034$). In an additional analysis, we experimented with applying the same ECG-AI architecture using only lead I data. Interestingly, we obtained an AUC of 0.754 (0.709–0.798), similar to the 12-lead version.

We also built traditional Cox proportional hazards regression to model time from baseline to incident HF up to 2018 follow-up. Cox model, utilizing all ARIC and FHS risk calculator variables as well as the outcome of ECG-AI, resulted in a concordance of 0.826 (0.804–0.848). For a fair comparison with the other three models, we set $t = 10$ and calculated the cumulative risk for HF within 10 years and obtained an AUC of 0.821 (0.781–0.861), sensitivity of 0.711, sensitivity of 0.752, positive predictive value of 0.132, and negative predictive value of 0.980. The AUC of the Cox model was higher than both AUC of the FHS risk calculator (DeLong test, $P < 0.01$) and the ARIC risk calculator (DeLong test, $P < 0.01$). The details of the Cox model provided in *Table 3* revealed that the ECG-AI outcome was the most important predictor of HF. This is also confirmed by the variable importance analysis on the LGBM model utilizing the outcome of the ECG-AI model and ARIC variables as inputs, which provided an AUC of 0.818 (see Supplementary material online, *Figure S2 and Section S3*).

## Subgroup analysis

Cox model yielded an AUC of 0.818 (0.781–0.858) for black, 0.816 (0.776–0.857) for white, 0.828 (0.788–0.868) for male, and 0.810 (0.769–0.851) for female participants.

## Sensitivity analysis over time

Our analysis was based on ECGs recorded at baseline exams. However, we also had access to ECGs recorded over follow-up exams. We used these follow-up exams to assess the sensitivity of our model on follow-up ECGs closer to the HF events. For the 20% hold-out dataset, we run our ECG-AI and final Cox models on the ECGs collected after baseline yet still preceding HF event. For controls, we used the latest available ECG. Next, for each patient with available follow-up ECGs, we calculated Δrisk as the difference between risk from original and follow-up ECG divided by the time between two ECGs. Therefore, Δrisk represents the change in predicted risk per year for each patient (*Table 4*).

## Clinical utility

We further assess the possible clinical utility of our final Cox model to identify patients at risk for HF who may benefit from cardiac imaging. *Table 5* presents four different scenarios of specificity (0.70, 0.80, 0.90, 0.95) and corresponding accuracy metrics.

The results in *Table 4* show that for scenario 1 corresponding to the specificity of 0.7, 32.5% (880 of 2701) patients would be predicted at high risk for HF, and among these high risk predicted patients, 13.2% (116 of 880) would develop HF within 10 years. For Scenario 4 corresponding to a specificity of 0.95, our model would identify 7.5% (204 of 2701) of the general population at high risk for HF where 37.7% (77 of 204) of them indeed would develop HF. Interestingly, if we would use follow-up ECGs for the same scenario, we could identify 8.1% (218 of 2701) of the patients at high risk for HF and 43.6% (95 of 218) of them indeed would develop HF.

# Discussion

Heart failure prevalence is increasing globally and is more commonly experienced by older persons. This can cause both monetary and personal burden. It is not uncommon for HF to be diagnosed at a late-stage, past pharmacological intervention.[7] It is therefore of high importance to predict HF at early stages and provide timely interventions. If detection and/or prediction are performed early, it can substantially reduce the overall burden. The FHS and ARIC HF risk calculators[3,8] examined existing HF risk factors and proposed simple and effective HF risk calculators that would facilitate the primary prevention and early diagnosis of HF in general practice. More complex models were then developed using additional data that can add to the potential of early identification of HF. The FHS HF calculator[3] uses a standard pooled logistic regression model to identify the risk of HF within 4 years, while the ARIC HF risk calculator[8] uses a Cox regression model. The latter was also applied in this study. In addition to using clinical variables, this research also used 12-lead ECGs to predict HF within 10 years, aiming to obtain comparable results to that using clinical risk factors.

Several recent studies have shown the utility of AI on digital ECGs (time-voltage signals) in the detection and prediction of arrhythmias and cardiovascular disease.[17,18] A range of AI models has been developed to predict the risk of abnormal heart conditions, including HF, atrial fibrillation.[19–21] There has also been an effort to use machine learning models to diagnose[22,23] and predict the possibility of readmission and mortality following HF using solely risk factors.[22] While some recent research proposes the use of AI in the prediction of HF using both or a collection of risk factors and 12-lead ECG information, there is rarely a comparative time window, and if so, it is within a relatively short period of time, e.g. present to 5 years.[24–26] Recent studies have used ECG waveform data to develop AI networks to identify specific cardiac abnormalities such as ejection fraction,[27] left ventricular systolic dysfunction,[28] and mitral regurgitations[29] all of which are directly or indirectly related to HF. However, a key component not addressed is the time window for early identification of the possibility of HF. A meta-analysis by Grün *et al.*[30] involving five main publications[31–33] reported an almost perfect prediction of congestive HF using a 2 s ECG (ROC > 0.98). These studies, however, do not provide information on the time window considered in developing the model and how early it can detect HF. This is a very important component to achieve the best results for diagnoses and precision medicine as opposed to identifying whether a person already has developed HF. It is thus essential to develop models that consider a trade-off of accuracy with timeliness of early diagnosis.

Results obtained in this research show that existing ARIC and FHS HF risk calculators utilizing a total of 12 clinical risk factors can predict HF with AUCs of 0.80 (0.75–0.85) and 0.78 (0.74–0.83), respectively. Our ECG-AI model (model 2) utilizing solely 12-lead ECGs yielded a comparable AUC of 0.756 (0.717–0.795) and AUC of 0.780 (0.737–0.823) when combined with age and gender (Model 3 in Supplementary material online, *Table S1*). Also, the lead I version of ECG-AI provided a comparable accuracy to the standard 12-lead-based ECG-AI model. Although our solely ECG-based model does not improve performance over existing ARIC and FHS risk calculators, our proposed ECG-AI model may be more applicable in a clinical setting since it relies only on ECG data. Considering the widespread use and availability of ECGs, such models can facilitate future automated pre-screening tools running on cardio-servers or electronic health records (EHR). This helps identify patients who may benefit from close monitoring or cardiac imaging, such as an echocardiogram or cardiac magnetic resonance imaging (MRI). The development of these AI-based models may ease the burden on healthcare