

Prediction of reduced left ventricular ejection fraction using atrial fibrillation or flutter electrocardiograms: A machine-learning study

DIGITAL HEALTH
Volume 11: 1–14
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076241311460
journals.sagepub.com/home/dhj



Soonil Kwon^{1,*,†}, SooMin Chung^{2,‡}, So-Ryoung Lee^{3,4} , Kwangsoo Kim^{4,5},
Junmo Kim², Dahyeon Baek⁶, Hyun-Lim Yang⁷, Eue-Keun Choi^{3,4}
and Seil Oh^{3,4}

Abstract

Objective: Although the evaluation of left ventricular ejection fraction (LVEF) in patients with atrial fibrillation (AF) or atrial flutter (AFL) is crucial for appropriate medical management, the prediction of reduced LVEF (<50%) with AF/AFL electrocardiograms (ECGs) lacks evidence. This study aimed to investigate deep-learning approaches to predict reduced LVEF (<50%) in patients with AF/AFL ECGs and easily obtainable clinical information.

Methods: Patients with 12-lead ECGs of AF/AFL and echocardiography were divided into those with LVEF <50% and ≥50%. A convolutional neural networks-based model customized to the study (AFibEFNet) and other deep-learning models were investigated. Electrocardiogram signals, ECG features, and clinical features (demographic information, comorbidities, blood cell counts, and blood test results) were collected for training. A hold-out test dataset was constructed using a different recruitment period. Five-fold cross-validation and calibration plots were used to evaluate performance.

Results: A total of 15,683 patients were analyzed (mean age, 70.0 ± 11.7 years; 61.2% men), with 82.2% having LVEF ≥50% and 17.8% having LVEF < 50%. Among the learning models, the AFibEFNet outperformed other models regarding area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), and F1-score. Using ECG signals alone, the AFibEFNet model predicted reduced LVEF with AUROC of 0.798 (95% confidence interval [CI], 0.767–0.829) and AUPRC of 0.508 (95% CI, 0.434–0.564). For the AFibEFNet model, additional training with ECG and clinical features significantly improved AUROC (0.816 vs. 0.798, $p = 0.04$) and AUPRC (0.547 vs. 0.508, $p < 0.001$). The AFibEFNet model primarily focused on the R-wave, QRS onset and offset, and T-wave in ECG signals.

Conclusions: Among the patients with AF/AFL, machine learning may predict reduced LVEF with 12-lead ECGs of AF/AFL.

Keywords

Atrial fibrillation, atrial flutter, machine learning, ejection fraction, heart failure

Submission date: 1 May 2024; Acceptance date: 16 December 2024

¹Division of Cardiology, Department of Internal Medicine, SMG-SNU Boramae Medical Center, Seoul, Republic of Korea

²Interdisciplinary Program in Bioengineering, Seoul National University, Seoul, Republic of Korea

³Department of Internal Medicine, Seoul National University Hospital, Seoul, Republic of Korea

⁴Department of Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea

⁵Department of Transdisciplinary Medicine, Institute of Convergence Medicine with Innovative Technology, Seoul National University Hospital, Seoul, Republic of Korea

⁶Industrial and Management Engineering, POSTECH, Pohang, Republic of Korea

⁷Office of Hospital Information, Seoul National University Hospital, Seoul, Republic of Korea

[†]The two authors contributed equally.

[‡]This work was conducted while the author was affiliated with Seoul National University Hospital.

Corresponding authors:

So-Ryoung Lee, Department of Internal Medicine, Seoul National University Hospital, 101 Daehak-ro, Jongno-gu, Seoul 03080, South Korea.
Email: minerva1368@gmail.com

Kwangsoo Kim, Department of Transdisciplinary Medicine, Institute of Convergence Medicine with Innovative Technology, Seoul National University Hospital, 101 Daehak-ro, Jongno-gu, Seoul 03080, South Korea.
Email: kksoo716@gmail.com



Introduction

Atrial fibrillation (AF) is the most prevalent cardiac arrhythmia and is often accompanied by atrial flutter (AFL).¹ For optimal management of patients with AF or AFL, measuring left ventricular ejection fraction (LVEF) is essential, as reduced LVEF limits the medications that can be safely prescribed. European guidelines advise against the use of Vaughan-Williams class I antiarrhythmic drugs or dronedarone in patients with AF and reduced LVEF, due to increased risks of proarrhythmia and adverse outcomes.^{2,3} Additionally, nondihydropyridine calcium channel blockers should be avoided for the heart rate control in these patients, as they may further decrease the cardiac output.^{2,3} The presence of reduced LVEF also influences a stroke risk and prevention strategies.^{2,4} However, an accurate measurement of LVEF typically requires an echocardiographic assessment.⁵

Deep learning has been introduced as an effective tool for analyzing 12-lead electrocardiograms (ECGs) and detecting the underlying medical condition.⁶ Accordingly, some reports stated that deep learning is feasible for detecting underlying left ventricular dysfunction by analyzing ECGs for the general population.^{7–10} However, most previous studies examined sinus rhythm ECGs; thus, their results may not apply to AF/AFL ECGs since it is often difficult to define the ST-segment or T waves during AF/AFL. Therefore, for patients with AF/AFL ECGs, we must develop a new deep-learning model to predict a reduced LVEF.

This study aimed to investigate a deep-learning approach to detect reduced LVEF (<50%) by analyzing 12-lead ECGs of AF/AFL and easily obtainable clinical data such as age, sex, and comorbidities.

Methods

Study design and population

This was a single-center retrospective cohort study. The enrollment flow of the study population is shown in Figure 1. Patients aged ≥ 18 years and underwent echocardiography between January 2003 and July 2022 was identified. This study included patients who underwent AF/AFL ECG and echocardiography at an interval of <1 month. The exclusion criteria were as follows: (1) no available ECG ($n = 13,011$); (2) no AF/AFL ECG ($n = 231,557$); (3) no echocardiographic data for analysis ($n = 29,060$); (4) missing values for LVEF ($n = 3514$); (4) an interval of ≥ 1 month between echocardiography and AF/AFL ECG ($n = 10,086$); and (5) outliers for study variables ($n = 384$). The outliers are defined in Supplementary Table 1. Consequently, 15,683 patients with AF/AFL ECG and echocardiographic LVEF data were investigated. The study population was divided into the control group

(those with $\text{LVEF} \geq 50\%$; $n = 12,899$) and the reduced LVEF group (those with $\text{LVEF} < 50\%$, $n = 2784$).

Data acquisition

This study used clinical data retrieved from the Seoul National University Hospital Patient Research Environment (SUPREME) system. Patients' demographic information, comorbidities, blood test results, and echocardiography data were extracted from the SUPREME system, while raw ECG data were retrieved from the MUSE Cardiology Information System (GE Healthcare, WI, USA). Physicians at the Seoul National University Hospital diagnosed AF or AFL based on their direct evaluation of the patients. A complete list of the features and their definitions is presented in Supplementary Table 2. The index date was defined as the date of the earliest AF/AFL ECG. Demographic and ECG features were acquired from the index date, whereas comorbidities, blood test results, and echocardiographic features were acquired within three months of the index date.

Echocardiographic features were compared between groups to delineate the baseline characteristics of the population further (Supplementary Table 3). This study used index ECGs which confirmed AF or AFL. The index ECGs were initially screened by diagnostic labels within the MUSE Cardiology Information System. If multiple echocardiographic studies were available within one month of the index AF/AFL ECG, the study with the closest date to the index AF/AFL ECG was chosen for analysis. Both the signal data and features were obtained for each ECG. Electrocardiogram features included heart rate, QRS duration, QT interval, R-axis, T-axis, Q onset, Q offset, T offset, and selected cardiac rhythm or conduction abnormalities (AF, AFL, bundle branch block, and atrioventricular block). The ECG parameters were measured from global fiducial points of all 12 simultaneous leads using the Marquette 12SL algorithm (GE Healthcare, Chicago, IL, USA). Numerous studies have utilized the Marquette 12SL algorithm due to its stability and accuracy in measuring ECG parameters such as amplitudes, durations, and intervals.¹¹ Previous publications have comprehensively documented the specific details and criteria for ECG diagnostic statements and measurements generated by the Marquette 12SL algorithm.^{11,12} The ECG data comprised signals from eight leads: I, II, and V1-V6. Each signal was recorded for 10 s at a sampling rate of 500 Hz. Initially encoded in base64 within XML files, the signal data were decoded into numerical arrays using a custom Python script. If missing values existed among continuous features, imputation was performed using the median value of each feature. For all tabular data, robust regularization was performed using the median and interquartile range values for reliable training.

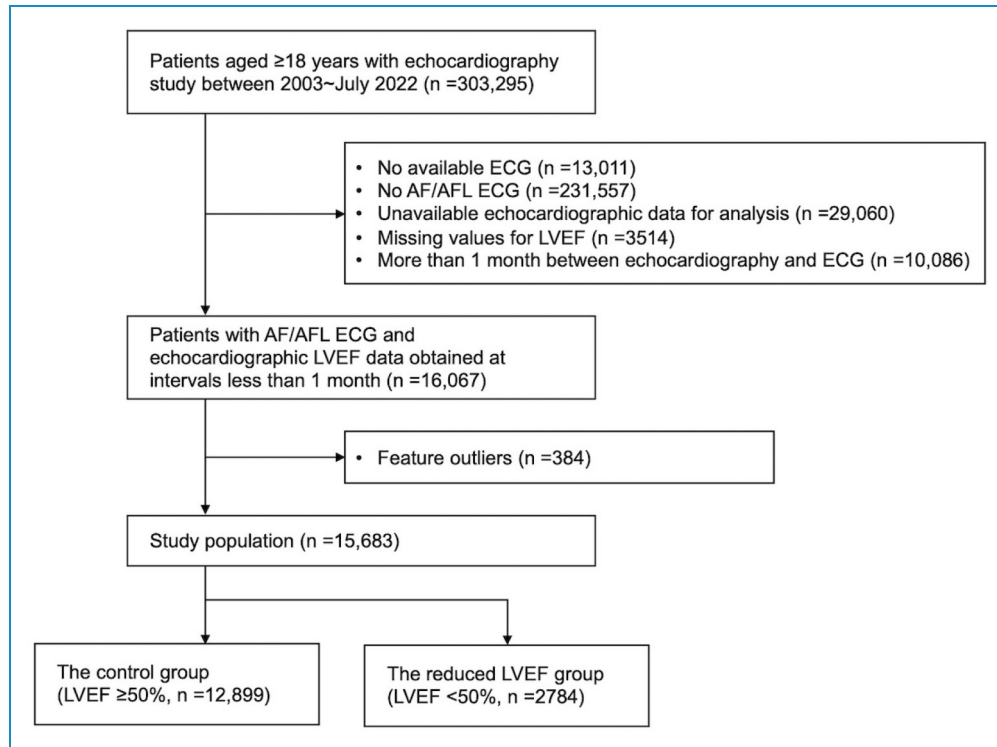


Figure 1. The flowchart of the study population. The study population was patients with AF/AFL ECGs and echocardiographic LVEF data within a one-month interval. The population was categorized into the control group with normal LVEF ($\geq 50\%$) and the reduced LVEF ($< 50\%$) group. AF: atrial fibrillation; AFL: atrial flutter; ECG: electrocardiogram; LVEF: left ventricular ejection fraction.

Deep-learning training and evaluation process

This study categorized training data as follows: (1) ECG signals that could be acquired from raw ECG data, (2) clinical features composed of demographic features (age, sex, height, body weight, and body mass index), comorbidities (hypertension, diabetes mellitus, ischemic heart disease, dyslipidemia, chronic obstructive pulmonary disease (COPD), chronic kidney disease (CKD), liver disease, stroke, thyroid disease), and blood test results (blood urea nitrogen (BUN), serum creatinine, glomerular filtration rate (GFR), high-sensitive C-reactive protein (hs-CRP), hemoglobin, serum levels of sodium, potassium, white blood cell and platelet counts), and (3) ECG features. Deep learning was designed to identify subjects with reduced LVEF ($< 50\%$). Tabular data, including the ECG features, were used in the LightGBM model,¹³ which has been shown to perform well among tree-based algorithms. For the deep-learning model, the ECG signal was processed through the convolutional layer, and the remaining data were concatenated with feature embeddings of the signal in the third fully connected layer on a convolutional neural network (CNN) model with a residual block.

We compared three CNN architectures. Inspired by architectures originally designed for two-dimensional image analysis, these models were modified to 1-D convolution to handle ECG data. All models were trained from

scratch using our ECG dataset. Three types of CNN-based deep-learning models were compared. The first model is a CNN with a residual block based on nEMGNet (the number of parameters: 46.3 M).¹⁴ The kernel sizes of some of the convolution layers were modified to fit the input shape of the ECG signals, and the clinical data (tabular data) were concatenated after the third fully connected layer. The second model was a modified version of ResNet50 (the number of parameters: 15.9 M).¹⁵ The kernel size was modified to fit the input shape and to concatenate the clinical data with the embedding vector of the ECG signal in the fully connected layer. The third model is EfficientNet b5 (the number of parameters: 41.7 M), another CNN-based model that achieves high efficiency and accuracy with fewer parameters.¹⁶ Also, we analyzed a long short-term memory (LSTM) model (the number of parameters: 353.2 M) as a representative of recurrent neural network (RNN)-based models.¹⁷ The architectures of the four deep-learning models are presented in Supplementary Figure 1. After analyses, a CNN with a residual block model exhibiting the highest performance was selected as the final model and named AFibEFNet to distinguish it from other CNN-based models. To assess the appropriateness of the final model selection, we further investigated its interpretability through various tests, including subgroup and sensitivity analyses.

In our analysis, the training and hold-out test datasets were split based on patients, meaning no patient was included in both datasets simultaneously. Among the study population, those before 2021 ($n=14,247$) were used to develop and train the model, and those after 2021 ($n=1436$) were used as the hold-out test dataset. A five-fold cross-validation was performed with patient-based splits for the performance evaluation, and the F1-score was used as the model optimization metric. The success criteria for predicting reduced LVEF was defined as achieving an area under the receiver operating characteristic curve (AUROC) with a lower limit of 95% confidence interval (CI) above 0.5, using the hold-out test dataset. All experiments were conducted in Python v.3.6.9 and Pytorch v.1.10.2 environments, and Pycaret library v.2.3.10 was used for the LightGBM model. Detailed information on the hyperparameter settings is presented in Supplementary Table 4.

This study adheres to the guidelines for machine learning application in biomedical research.¹⁸

Calibration plot

A calibration plot was constructed to evaluate the model's performance by dividing the test dataset into six bins according to the predicted probability of a reduced LVEF. The plot depicts the relationship between the predicted and actual probabilities for a given class.

Evaluating feature attributions

The guided Grad-CAM method¹⁹ was used to visualize the ECG segments important for the deep-learning model in predicting reduced LVEF. Sensitivity maps for the average beat of each lead were plotted for the five patients with the highest probability of reduced LVEF. For the LightGBM model, we investigated feature importance. Gini importance was calculated for the feature importance.

Statistical analyses

For the performance evaluation, various diagnostic parameters were evaluated (AUROC, area under the precision-recall curve (AUPRC), F1-score, sensitivity, and specificity). We compared the diagnostic parameters across the following models: (1) four deep-learning models with training ECG signals, clinical features, and ECG features; (2) four deep-learning models with training ECG signals alone; (3) the LightGBM model with training clinical features and ECG features; (4) the LightGBM model with training ECG features; and (5) the LightGBM model with training clinical features. Owing to the inherent randomness of deep-learning models, we repeated the measurement of the diagnostic parameters five times. The average results were reported with 95% CIs. For the threshold-dependent

metrics, we used the Youden J-index. Area under the receiver operating characteristic curves were compared using the DeLong test. Two-sided values of $p<0.05$ rejected the null hypothesis. All statistical analyses were performed using the R software and Python.

Subgroup and sensitivity analyses

We performed the subgroup analysis according to sex, the timing of ECG and echocardiography (ECG first/echocardiography first), hypertension, ischemic heart disease, diabetes mellitus, stroke, thyroid disease, COPD, liver disease, CKD, left atrial diameter (<40 mm/ ≥ 40 mm), and heart rate (<100 beats/min/ ≥ 100 beats/min). To compare the model's performance according to the reduced LVEF criteria, we performed a sensitivity analysis by modifying the definition of reduced LVEF to $<40\%$ and $<35\%$ and compared the diagnostic parameters with the main results.

Results

Baseline characteristics of the study population

A total of 15,683 patients with AF/AFL were analyzed (12,899 in the control group and 2784 in the reduced LVEF group). Compared to the control group, the reduced LVEF group was more likely to be male (65.5% vs. 60.3%, $p<0.001$). In general, the reduced LVEF group had more prevalent comorbidities, such as diabetes mellitus, ischemic heart disease, COPD, CKD, and liver disease, than the control group (Table 1). Accordingly, the reduced LVEF group had significantly higher BUN, serum creatinine, hs-CRP, and potassium levels and lower GFR, hemoglobin, and platelet counts (all $p<0.001$) (Table 1). For ECG parameters, the reduced LVEF group had a significantly higher heart rate (103 vs. 91 beats/min), wider QRS duration (98 vs. 92 ms), shorter QT interval (360 vs. 368 ms), and higher incidences of AFL (13.8% vs. 10.3%) and LBBB (4.7% vs. 0.9%) (all $p<0.001$ except for QT interval, which was $p=0.001$) (Table 1).

Deep-learning performance for predicting reduced LVEF

Table 2 summarizes deep-learning performance across the algorithms and training datasets. For the LightGBM model, training ECG features yielded a higher AUROC than training clinical features (0.758; 95% CI, 0.727–0.790; and 0.670; 95% CI, 0.633–0.708, respectively). In general, training both ECG and clinical features resulted in a higher diagnostic performance than training either feature set alone (AUROC, 0.752; 95% CI, 0.717–0.786; AUPRC, 0.423; 95% CI, 0.356–0.490, F1-score, 0.430; 95% CI, 0.390–0.476; sensitivity, 0.645; 95% CI, 0.543–0.782; and specificity, 0.735; 95% CI, 0.607–0.838).

Table 1. Baseline characteristics.

	Control group (LVEF \geq 50%, $n = 12,899$)	Reduced LVEF group (LVEF <50%, $n = 2784$)	p
Demographics			
Age (year)	70 (62–77)	70 (61–77)	0.56
Men (%)	7774 (60.3)	1823 (65.5)	<0.001
Height (cm)	163.4 (156.0–169.8)	164.1 (157.0–170.0)	0.002
Body weight (kg)	63.9 (55.9–72.0)	63.0 (55.0–71.1)	0.002
Body mass index (kg/m ²)	24.1 (22.0–26.3)	23.7 (21.5–26.0)	<0.001
Comorbidity			
Hypertension (%)	2566 (19.9)	528 (19.0)	0.27
Diabetes mellitus (%)	1841 (14.3)	539 (19.4)	<0.001
Ischemic heart disease (%)	329 (2.6)	326 (11.7)	<0.001
Dyslipidemia (%)	1828 (14.2)	421 (15.1)	0.19
COPD (%)	434 (3.4)	126 (4.5)	0.003
Chronic kidney disease (%)	1444 (11.2)	472 (17.0)	<0.001
Liver disease (%)	929 (7.2)	155 (5.6)	0.002
Stroke (%)	1541 (11.9)	316 (11.4)	0.38
Thyroid disease (%)	557 (4.3)	102 (3.7)	0.12
Laboratory blood test			
BUN (mg/dL)	17 (14–23)	20 (15–29)	<0.001
Serum creatinine (mg/dL)	0.97 (0.80–1.20)	1.10 (0.88–1.47)	<0.001
Estimated GFR (mL/kg/1.73m ²)	72.2 (56.3–86.8)	64.9 (43.8–81.3)	<0.001
hs-CRP (mg/dL)	0.31 (0.08–2.42)	0.59 (0.15–3.30)	<0.001
Hemoglobin (mg/dL)	13.1 (11.2–14.6)	12.7 (10.9–14.5)	<0.001
Sodium (mg/dL)	140 (137–142)	139 (136–141)	<0.001
Potassium (mg/dL)	4.3 (4.0–4.7)	4.4 (4.0–4.8)	<0.001
White blood cell (k/mL)	6.41 (5.01–8.20)	6.73 (5.08–8.95)	<0.001
Platelet (k/mL)	201 (159–249)	192 (148–242)	<0.001
Electrocardiogram			
Heart rate (/min)	91 (74–116)	103 (83–127)	<0.001

(continued)

Table 1. Continued.

	Control group (LVEF $\geq 50\%$, $n = 12,899$)	Reduced LVEF group (LVEF $< 50\%$, $n = 2784$)	p
QRS duration (ms)	92 (84–100)	98 (88–114)	<0.001
QT interval (ms)	368 (328–404)	360 (324–400)	0.001
R axis (degree)	44 (11–69)	35 (–8–73)	<0.001
T axis (degree)	38 (6–73)	73 (11–145)	<0.001
Q onset (ms)	221 (215–226)	219 (214–225)	<0.001
Q offset (ms)	266 (261–273)	268 (263–282)	<0.001
T offset (ms)	405 (386–424)	402 (382–421)	<0.001
Atrial fibrillation (%)	11,564 (89.7)	2400 (86.2)	<0.001
Atrial flutter (%)	1335 (10.3)	384 (13.8)	<0.001
Right bundle branch block (%)	1282 (9.9)	287 (10.3)	0.56
Left bundle branch block (%)	122 (0.9)	132 (4.7)	<0.001

Data are N (%) or median (interquartile range).

BUN: blood urea nitrogen; COPD: chronic obstructive pulmonary disease; GFR: glomerular filtration rate; hs-CRP: high-sensitive C-reactive protein; LVEF: left ventricular ejection fraction.

However, the LightGBM model was inferior to the AFibEFNet model even if it was trained with ECG signals alone (Table 2 and Figure 2).

Similarly, it was generally observed in other deep-learning models that diagnostic performance improved as more datasets were trained (Table 2). Among the four deep-learning models, the AFibEFNet outperformed other deep-learning models regarding AUROC, AUPRC, and F1-score (Table 2). Across the various models, the final model (the AFibEFNet trained with all datasets which included ECG signals, ECG features, and clinical features) achieved the highest diagnostic performance in general (AUROC, 0.816; 95% CI, 0.787–0.845; AUPRC, 0.547; 95% CI, 0.481–0.594; F1-score, 0.492; 95% CI, 0.461–0.536; sensitivity, 0.765; 95% CI, 0.692–0.833; and specificity, 0.738; 95% CI, 0.699–0.799).

Subgroup analysis

Subgroup analyses were performed using the final model (Table 3). Among the subgroups, the model showed significantly higher performance in patients without liver disease or CKD; the AUROC was 0.826 (95% CI: 0.797–0.855) versus 0.648 (95% CI: 0.495–0.801) for liver disease ($p = 0.03$) and 0.826 (95% CI: 0.797–0.855) versus 0.618 (95% CI: 0.447–0.788) for CKD ($p = 0.02$). Also, the

subgroup of slower heart rates ($<100/\text{min}$) showed a significantly higher AUROC compared to its counterpart ($\geq 100/\text{min}$); AUROCs of 0.851 and 0.769, respectively; $p = 0.009$. The final model showed marginally increased AUROCs for those with ischemic heart disease ($p = 0.14$) and left atrial diameter of ≥ 40 mm ($p = 0.10$) versus their counterparts. However, no significant performance differences were observed for other comorbidities. The model showed the best overall performance in patients with ischemic heart disease (AUROC of 0.868, AUPRC of 0.905, F1-score of 0.791, and sensitivity of 0.944). However, the highest specificity (0.800) was observed in the subgroup with heart rates <100 beats/min.

Sensitivity analysis

As the cutoff for reduced LVEF became lower ($<40\%$ or $<35\%$), the AUROC generally increased (0.859 vs. 0.868), while the AUPRC and F1-score decreased (0.459 vs. 0.406 and 0.352 vs. 0.274 for LVEF $<40\%$ and $<35\%$, respectively) (Supplementary Table 5).

Evaluation of feature attributions for predicting reduced LVEF

Figure 3 illustrates the feature attributions of the LightGBM model. Among the ECG features, information on the QRS morphology (R-axis, Q onset, Q offset, and QRS duration),

Table 2. The deep-learning performance for the prediction of reduced LVEF among patients with AF/AFL.

Algorithms	Training datasets	AUROC	AUPRC	F1-score	Sensitivity	Specificity
Machine learning	ECG features	0.758 (0.727–0.790)	0.362 (0.302–0.436)	0.411 (0.382–0.468)	0.752 (0.594–0.915)	0.629 (0.480–0.799)
Machine learning	Clinical features	0.670 (0.633–0.708)	0.325 (0.260–0.382)	0.348 (0.317–0.379)	0.726 (0.436–0.825)	0.524 (0.452–0.809)
Machine learning	All features	0.752 (0.717–0.786)	0.423 (0.356–0.490)	0.430 (0.390–0.476)	0.645 (0.543–0.782)	0.735 (0.607–0.838)
ResNet50	ECG signals	0.766 (0.730–0.803)	0.480 (0.406–0.540)	0.457 (0.420–0.514)	0.684 (0.581–0.786)	0.745 (0.641–0.836)
ResNet50	ECG signals and all features	0.775 (0.739–0.810)	0.503 (0.425–0.558)	0.480 (0.436–0.521)	0.684 (0.615–0.795)	0.773 (0.665–0.805)
AFibEFNet	ECG signals	0.798 (0.767–0.829)	0.508 (0.434–0.564)	0.464 (0.420–0.513)	0.697 (0.628–0.850)	0.745 (0.594–0.811)
AFibEFNet (the final model)	ECG signals and all features	0.816 (0.787–0.845)	0.547 (0.481–0.594)	0.492 (0.461–0.536)	0.765 (0.692–0.833)	0.738 (0.699–0.799)
EfficientNet b5	ECG signals	0.782 (0.749–0.815)	0.500 (0.422–0.552)	0.473 (0.420–0.526)	0.658 (0.551–0.774)	0.795 (0.656–0.872)
EfficientNet b5	ECG signals and all features	0.790 (0.758–0.823)	0.507 (0.427–0.562)	0.492 (0.451–0.530)	0.697 (0.632–0.765)	0.786 (0.731–0.815)
LSTM	ECG signals	0.584 (0.544–0.624)	0.211 (0.174–0.256)	0.313 (0.284–0.342)	0.603 (0.410–0.765)	0.579 (0.377–0.734)
LSTM	ECG signals and all features	0.716 (0.679–0.754)	0.403 (0.330–0.459)	0.409 (0.373–0.455)	0.637 (0.466–0.709)	0.714 (0.666–0.866)

Data are mean (95% CI). The numbers of training and test samples were 14,247 and 1436, respectively. Compared to the final model, the other models showed significantly lower AUROCs; all $p < .05$.

AF: atrial fibrillation; AFL: atrial flutter; AUPRC: the area under the precision-recall curve; AUROC: the area under the receiver operating characteristics curve; ECG: electrocardiogram; LSTM: long short-term memory; LVEF: left ventricular ejection fraction.

ventricular repolarization (T-axis and T offset), and heart rate were considered important for the LightGBM model. Among the clinical features, blood cell count, demographic features (BMI, body weight, and age), and serologic biomarkers (hs-CRP, creatinine, sodium, and blood urea nitrogen) are important. If all the features were trained together, some of the most important were the R-axis, creatinine, heart rate, and T-axis. For deep learning, the important ECG signals were mainly the R-waves (especially lead V6), QRS onset and offset (leads V1, V2, and V3), and minorly T-waves (leads V1 and V2) (Figure 4).

Discussion

This study demonstrated the feasibility of using deep learning to detect underlying reduced LVEF in patients with AF/AFL, employing 12-lead ECG and clinical information. Our principal findings are: First, deep learning achieved

the highest diagnostic performance when the AFibEFNet model was trained with raw ECG signals and features. Second, learning raw ECG signals provided better predictions of reduced LVEF compared to learning ECG or clinical features alone, emphasizing the diagnostic importance of raw AF/AFL ECG signals. Compared to previous reports,^{7–10} our study's strengths include (1) focusing on patients with AF/AFL ECG, (2) using a large-scale dataset (~15,000 patients), and (3) analyzing multiple diagnostic parameters (AUROC, AUPRC, F1-score, etc.) for a balanced interpretation of the model's performance.

Machine learning can be used to detect reduced LVEF or heart failure.^{20,21} However, most studies investigated general cardiovascular patients and did not focus on those with AF/AFL. Attia et al. validated a deep-learning prediction model of reduced LVEF using ECGs of general cardiovascular patients.⁹ They showed that the deep-learning model achieved a sensitivity of 82.5% and specificity of

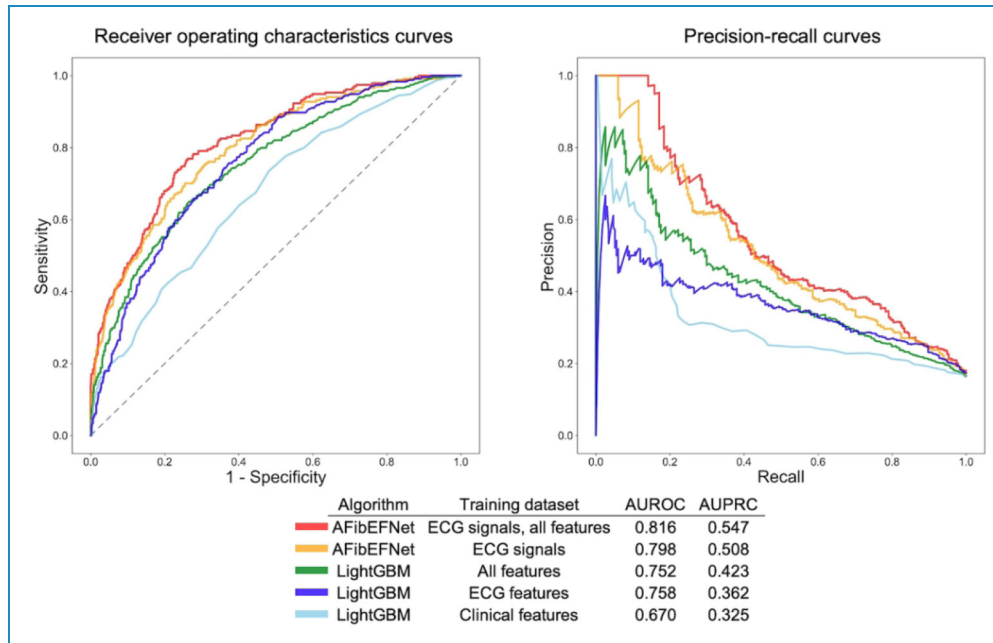


Figure 2. The performance of the AFibEFNet and LightGBM models for predicting reduced LVEF. The performance of the two models was evaluated by comparing their AUROCs and AUPRCs on different training datasets. AUPRC: the area under the precision-recall curve; AUROC: the area under the receiver operating characteristics curve; ECG: electrocardiogram; LVEF: left ventricular ejection fraction.

86.8%, which seems better than our results (sensitivity of 76.5%, specificity of 73.8%). A possible explanation for the differences between the previous study and our study is as follows: First, the previous study used a relatively more severe definition of a reduced LVEF ($\leq 35\%$) than our study ($< 50\%$). We found that AUROC tended to increase as the cutoff value for reduced LVEF decreased, but this benefit was counterbalanced by decreased AUPRC (Supplementary Table 5). This could be because ECG features for left ventricular dysfunction, such as pathologic Q-waves, ST-segment changes, or T-wave abnormalities, become more robust as left ventricular dysfunction increases. Second, previous studies focused on general cardiovascular patients, whereas our study focused on those with AF/AFL. Predicting a reduced LVEF using AF/AFL ECGs might be challenging because the signs of left ventricular dysfunction, such as ST segments and T-wave abnormalities, are often difficult to characterize because of fibrillatory or flutter waves. Therefore, deep-learning models trained with sinus rhythm ECGs may yield different results if applied to patients with AF/AFL.

To date, few studies have targeted patients with AF to predict heart failure using machine learning. Hamatani et al. investigated the machine-learning prediction of heart failure in patients with AF.²² The study evaluated seven clinical variables (age, history of heart failure, creatinine clearance, the cardiothoracic ratio on X-ray, LVEF, left ventricular end-systolic diameter, and left ventricular asynergy) to predict heart failure among patients with AF. However, the clinical utility of

machine learning appears limited because the previous model requires a prior echocardiographic study to perform machine learning analysis. In addition, the study did not analyze AF/ECG signals, which could have greater diagnostic importance for reduced LVEF than clinical variables, as shown in our analysis (Table 2). Compared with a previous study, we also focused on patients with AF/AFL and utilized AF/AFL ECG signals, ECG features, and clinical variables to predict a reduced LVEF.

Our data emphasizes the importance of ECG to predict reduced LVEF with deep learning. The ECG features were considered more important than clinical features (Figure 3 and Table 2). When the total feature importance scores were evaluated, the 23 clinical features had a total score of 922, whereas 12 ECG features had a total score of 673. Therefore, each ECG feature had more weight on average, with a mean importance score of 56.08, compared to 40.09 for clinical features. These results supported the finding that ECG features hold a higher feature importance than clinical features in our model. Also, we conducted a feature ablation analysis in which LightGBM's performances were compared by sequentially excluding each feature (Supplementary Figure 2). The result was generally coherent, suggesting that selected ECG features (particularly heart rate, R-, Q-, and T-wave) were crucial to classifying the reduced LVEF group. However, deep learning of raw ECG signals may outperform machine learning of human-invented ECG features for detecting underlying heart disease.^{23,24} Accordingly, we also observed that deep learning of raw ECG signals resulted in a better

Table 3. Subgroup analysis.

Subgroup	Training samples	Test samples	AUROC	AUPRC	F1-score	Sensitivity	Specificity	<i>p</i> ^a
Sex								
Men	14,247	931	0.823	0.611	0.513	0.753	0.736	0.61
Women	14,247	505	0.808	0.360	0.445	0.797	0.741	
LVEF cutoff								
50%	14,247	1436	0.816	0.547	0.492	0.765	0.738	0.05
40%	14,247	1436	0.859	0.459	0.352	0.872	0.706	
35%	14,247	1436	0.868	0.406	0.274	0.879	0.692	
Timing of ECG and echocardiography								
ECG first	14,247	585	0.839	0.581	0.512	0.844	0.735	0.24
Echocardiography first	14,247	851	0.803	0.536	0.478	0.715	0.740	
Hypertension								
Yes	14,247	370	0.844	0.621	0.536	0.778	0.769	0.24
No	14,247	1066	0.806	0.518	0.477	0.760	0.727	
Ischemic heart disease								
Yes	14,247	66	0.868	0.905	0.791	0.944	0.467	0.14
No	14,247	1370	0.799	0.470	0.452	0.732	0.745	
Diabetes mellitus								
Yes	14,247	239	0.868	0.755	0.632	0.820	0.815	0.06
No	14,247	1197	0.802	0.485	0.468	0.766	0.735	
Stroke								
Yes	14,247	128	0.838	0.592	0.549	0.778	0.864	0.67
No	14,247	1308	0.814	0.547	0.498	0.773	0.740	
Thyroid disease								
Yes	14,247	50	0.815	0.478	0.455	0.875	0.810	0.99
No	14,247	1386	0.816	0.553	0.496	0.770	0.744	
COPD								
Yes	14,247	42	0.899	0.828	0.741	0.923	0.897	0.12

(continued)

Table 3. Continued.

Subgroup	Training samples	Test samples	AUROC	AUPRC	F1-score	Sensitivity	Specificity	<i>p</i> ^a
No	14,247	1394	0.812	0.541	0.486	0.765	0.745	
<i>Liver disease</i>								
Yes	14,247	78	0.648	0.263	0.341	0.750	0.667	0.03
No	14,247	1358	0.826	0.566	0.513	0.775	0.763	
<i>CKD</i>								
Yes	14,247	67	0.618	0.466	0.462	0.750	0.667	0.02
No	14,247	1369	0.826	0.558	0.499	0.775	0.752	
<i>Left atrial diameter</i>								
<40 mm	14,247	335	0.770	0.347	0.436	0.720	0.723	0.10
≥40 mm	14,247	1083	0.832	0.597	0.509	0.787	0.744	
<i>Heart rate</i>								
<100 per minute	14,247	755	0.851	0.517	0.456	0.756	0.800	0.009
≥100 per minute	14,247	681	0.769	0.567	0.515	0.770	0.66	

^aFor the comparison of AUROCs between subgroups.
Data are mean. The analysis was performed with the final model (the AFibEFNet model trained with ECG signals, ECG features, and clinical features).
AUPRC: the area under the precision-recall curve; AUROC: the area under the receiver operating characteristics curve; CKD: chronic kidney disease; COPD: chronic pulmonary obstructive disease; ECG: electrocardiogram; LVEF: left ventricular ejection fraction.

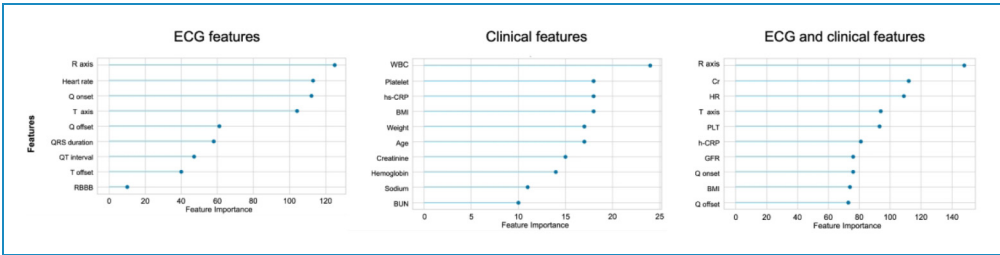


Figure 3. Important features for the LightGBM model to predict reduced LVEF. The importance of ECG and clinical features was calculated and compared. BMI: body mass index; BUN: blood urea nitrogen; ECG: electrocardiogram; GFR: glomerular filtration rate; hs-CRP: high-sensitive C-reactive protein; LVEF: left ventricular ejection fraction; RBBB: right bundle branch block.

predictive performance than the LightGBM model training with ECG features (Table 2). This finding suggests that, in addition to the analysis of human-invented ECG features, that of whole ECG signals using deep learning would be beneficial for predicting a reduced LVEF.

Among the deep-learning models, we observed that the CNN-based models (AFibEFNet, ResNet50, and EfficientNet b5) performed better than the LSTM model. However, we acknowledge that our findings do not definitively establish the superiority of CNN over RNN for all

applications. This is primarily because the comparative efficacy of these models was not the primary focus of our study, and other RNN-based models may yield different results. Nonetheless, our findings were consistent with existing literature that CNN-based models showed robust performance in ECG signal analysis,²⁵ which may enhance the robustness and credibility of our study.

Besides the model's performance, the interpretability of the model is crucial, especially in clinical applications. To address this concern, we have visualized the feature

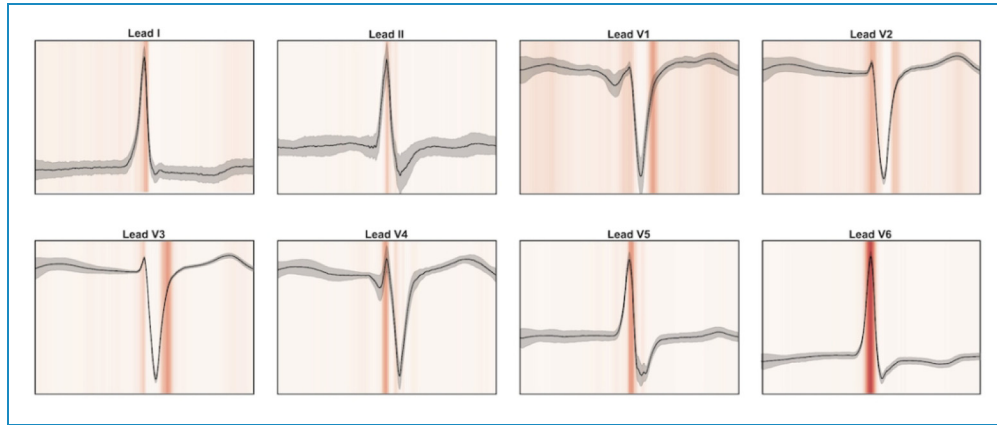


Figure 4. Visualization of feature attributions of ECG signals to predict reduced LVEF. Darker shades in Guided Grad-CAM visualize critical ECG features essential for model assessment. Black lines represent the averaged ECG of leads for five patients with the highest likelihood of reduced LVEF. Grey areas represent the standard deviation of the beats. ECG: electrocardiogram, LVEF: left ventricular ejection fraction.

attributions of the ECG signals that the model uses to predict reduced LVEF, as shown in Figure 4. The visualization indicates that the model prioritizes the R-wave, QRS onset and offset, and T-wave. This focus aligns with clinical knowledge, as left ventricular dysfunction in heart failure patients often leads to observable changes in QRS morphology. The QRS complex, which represents ventricular activation, is typically affected in conditions such as myocardial infarction and cardiomyopathy, presenting with widened QRS duration, reduced R-wave amplitude, alterations in the R-wave axis, pathological Q-waves, and ST-segment changes. Therefore, the model's emphasis on the QRS complex is clinically reasonable. Further analysis of feature importance (Figure 3) corroborates the findings from the ECG signal attributions. The model consistently considers factors such as the R-wave axis and heart rate, along with Q-wave onset and offset, as significant for identifying reduced LVEF. Among the clinical features, variables such as blood cell count, high-sensitivity C-reactive protein, body mass index, and age were deemed important. These factors are clinically relevant as systemic inflammation and advanced age are recognized risk factors for heart failure.²⁶ Therefore, the model's prioritization of these specific features for predicting reduced LVEF suggests that it operates on medically relevant criteria, thus providing some interpretability.

In the context of deep-learning classification for medical purposes, it is frequently observed that there are significantly more normal samples than diseased ones. In such cases where the classes are imbalanced, as in our study, training deep-learning models to produce accurate results is challenging, often leading to a bias toward the majority class in predicted class probabilities.²⁷ The calibration plot (Figure 5) was employed to evaluate the model's probabilistic outputs considering the class imbalance present in our dataset. A well-calibrated model should provide outputs

that closely approximate the actual likelihood of an event. According to Figure 5, our final model exhibited under-confidence, which was inclined to incorrectly predict the absence of reduced LVEF, especially within subsets of data with a higher prevalence of the condition. Therefore, if our model is used in clinical practice, physicians need a cautious interpretation when the predicted probability of reduced LVEF is above 0.5.

Limitations

First, there was a class imbalance (12,899 and 2784 patients in the control and reduced LVEF groups, respectively). To assess the impact of class imbalance, we conducted a sensitivity analysis by systematically adjusting the class ratio within our dataset to observe the corresponding effects on the model's performance, thereby investigating the impact of class imbalance (Supplementary Table 6). As the class imbalance increased, there was a general trend of an increasing AUROC, and a decreasing F1-score balanced this improvement. Despite the varying levels of imbalance, the model maintained a generally acceptable performance. Another approach to addressing class imbalance is to use a weighted loss function to penalize the majority class during training. However, the model's performance with the weighted loss function did not improve compared to the default loss function (Supplemental Table 7). We hypothesize that the default loss function may have allowed the model to learn more generalized patterns without overemphasizing the minority class, resulting in better overall performance. Second, we were unable to perform external validation for this study. However, we implemented a temporal validation approach to mitigate the risk of overfitting and enhance our findings' validity. We trained our model on patient data collected until 2021 and validated it on a subsequent dataset comprising patients

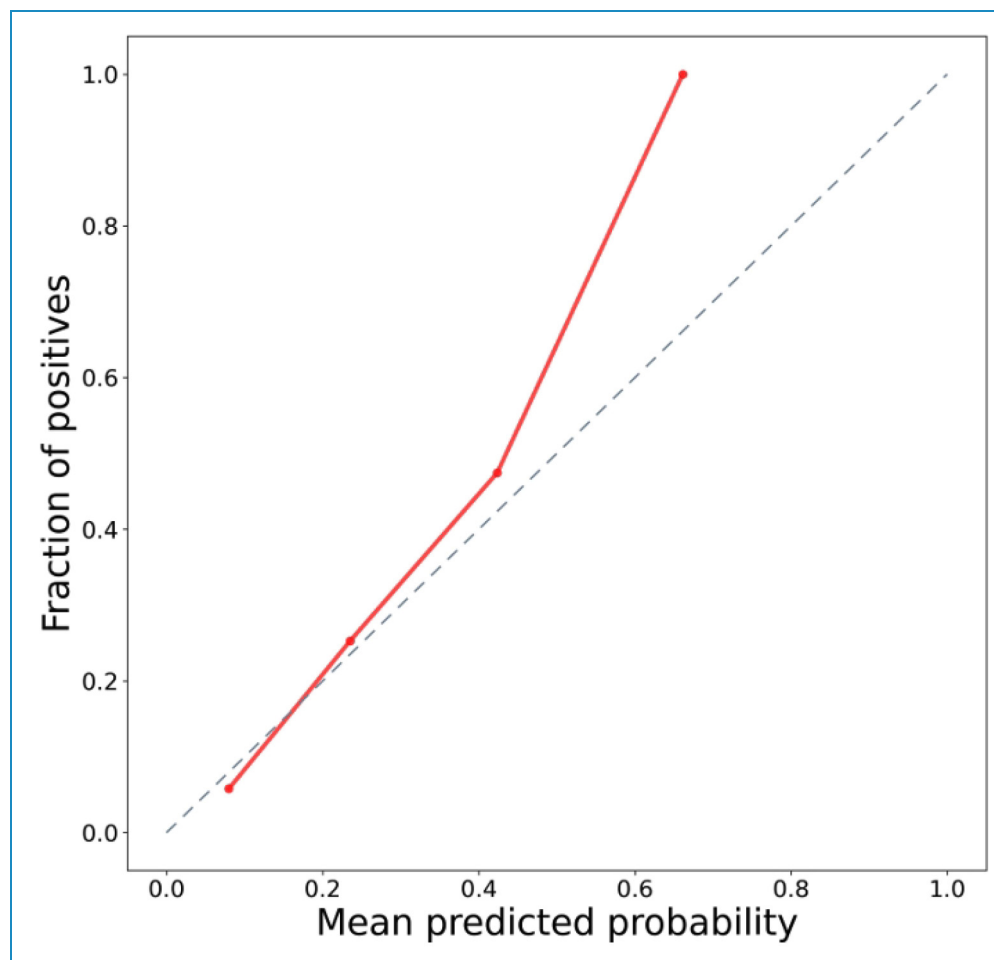


Figure 5. A calibration plot of the final model. A calibration plot of the final model (AFibEFNet trained with ECG signals and all features) showed under-confidence. ECG: electrocardiogram.

from 2021 onward. This approach ensures a clear temporal separation between the training and validation sets, thereby reducing the potential for overfitting. Third, the operational definitions of comorbidities might have over- or under-estimated their prevalence, although they were validated and peer-reviewed elsewhere.^{28–30} Also, as the data were obtained from a single tertiary hospital, the patient profile may differ from that of the general AF/AFL population. Fourth, although the deep-learning model was able to detect a reduced LVEF in patients with AF/AFL ECG, its performance was not validated for other underlying cardiac conditions, such as heart failure with preserved LVEF or structural heart disease. Fifth, variability introduced by different ECG equipment and the varying expertise of technicians could affect signal quality. We evaluated the model's performance across test datasets sorted by ECG devices and technicians, as shown in Supplementary Table 8. Despite the limited subgroup sizes, the general preservation of robustness across devices and technicians was observed. Sixth, individual variations may affect the

model's performance. Although it is challenging to evaluate every individual factor, the subgroup analysis investigated the impact of selected variables on performance. Seventh, although our model demonstrated the feasibility of predicting reduced LVEF, its current performance remains modest, limiting its immediate clinical applicability. Further development and validation in diverse populations, datasets, and racial groups are essential to enhance its generalizability and potential clinical utility. Finally, our results are primarily based on the Korean population. Thus, the extrapolation of our model to other races needs further validation.

Conclusions

Machine learning prediction of reduced LVEF using AF/AFL ECGs was feasible. However, the model's modest performance may limit its current clinical applicability. Further model development and validation in broader AF and AFL

populations are necessary to assess its potential clinical utility.

Abbreviations: AF: atrial fibrillation; AFL: atrial flutter; AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve; CI: confidence interval; ECG: electrocardiogram; LVEF: left ventricular ejection fraction; HR: hazard ratio.

Consent to participate: The Seoul National University Hospital Institutional Review Board waived the requirement for informed consent since the study used anonymized data, and thus, consent would be impossible or impracticable to obtain.

Contributorship: Kwon S and Chung S contributed equally and are co-first authors. Lee SR and Kim K are the co-corresponding authors. Conceptualization: Kwon S, Chung S, Kim J, Baek D, Yang HL, Lee SR, Kim K; Data curation: Kwon S, Chung S, Kim J, Baek D; Formal analysis: Kwon S, Chung S, Kim J, Baek D; Funding acquisition: Lee SR, Kim K; Investigation: Kwon S, Chung S, Kim J, Baek D, Yang HL, Lee SR, Choi EK, Oh S, Kim K; Methodology: Kwon S, Chung S, Kim J, Baek D, Yang HL, Lee SR; Project administration: Lee SR, Kim K; Supervision: Yang HL, Lee SR, Kim K; Validation: Kwon S, Chung S, Kim J, Baek D, Yang HL, Lee SR, Kim K; Visualization: Kwon S, Chung S; Writing original draft: Kwon S, Chung S; Writing review & editing: Kwon S, Chung S, Yang HL, Lee SR, Choi EK, Oh S, Kim K.

Data and code availability: The raw data used in this study are available upon reasonable request for research purposes only. The code for this study is publicly accessible at <https://github.com/SoominChung/AFibEFNet>, under the CC BY-NC 2.0 license.

Declaration of conflicting interests: The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Choi EK: Research grants from Bayer, BMS/Pfizer, Biosense Webster, Chong Kun Dang, Daiichi-Sankyo, Samjinpharm, Sanofi-Aventis, Seers Technology, Skylabs, and Yuhan. No fees are received personally.


Ethical approval: The study protocol conformed to the Declaration of Helsinki (revised in 2013) and was reviewed and approved by the Seoul National University Hospital Institutional Review Board (no. H-2207-001-1336).

Funding: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: HI20C1662, 1711138358, KMDF_PR_20200901_0173). The funding source had no roles in the study.

Guarantor: KS and CS.

Supplemental material: Supplemental material for this article is available online.

ORCID iDs: So-Ryoung Lee  <https://orcid.org/0000-0002-6351-5015>

Eue-Keun Choi  <https://orcid.org/0000-0002-0411-6372>

References

1. Lee SR, Choi EK, Han KD, et al. Trends in the incidence and prevalence of atrial fibrillation and estimated thromboembolic risk using the CHA₂DS₂-VASc score in the entire Korean population. *Int J Cardiol* 2017; 236: 226–231.
2. Hindricks G, Potpara T, Dagres N, et al. 2020 ESC guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS): the Task Force for the diagnosis and management of atrial fibrillation of the European Society of Cardiology (ESC) Developed with the special contribution of the European Heart Rhythm Association (EHRA) of the ESC. *Eur Heart J* 2021; 42: 373–498.
3. Gopinathannair R, Chen LY, Chung MK, et al. Managing atrial fibrillation in patients with heart failure and reduced ejection fraction: a scientific statement from the American Heart Association. *Circ Arrhythm Electrophysiol* 2021; 14: HAE0000000000000078.
4. Lip GY, Nieuwlaat R, Pisters R, et al. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest* 2010; 137: 263–272.
5. Richards M, Maskell G, Halliday K, et al. Diagnostics: a major priority for the NHS. *Future Healthc J* 2022; 9: 133–137.
6. Attia ZI, Noseworthy PA, Lopez-Jimenez F, 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.
7. Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med* 2019; 25: 70–74.
8. Adedinsowo D, Carter RE, Attia Z, et al. Artificial intelligence-enabled ECG algorithm to identify patients with left ventricular systolic dysfunction presenting to the emergency department with dyspnea. *Circ Arrhythm Electrophysiol* 2020; 13: e008437.
9. Attia ZI, Kapa S, Yao X, et al. Prospective validation of a deep learning electrocardiogram algorithm for the detection of left ventricular systolic dysfunction. *J Cardiovasc Electrophysiol* 2019; 30: 668–674.
10. Kashou AH, Medina-Inojosa JR, Noseworthy PA, et al. Artificial intelligence-augmented electrocardiogram detection of left ventricular systolic dysfunction in the general population. *Mayo Clin Proc* 2021; 96: 2576–2586.

11. Marquette™ 12SL™ ECG analysis program - statement of validation and accuracy, <https://www.gehealthcare.com/support/manuals?search=eyJzZWZyY2hUZlJtJoiNDE2NzkyLTAwMyIsImxhbmd1YWdlTmFtZSI6IkVuZ2xpc2ggKEVOKSJ9> (2008, accessed 25 October 2024).
12. Marquette™ 12SL™ ECG analysis program - physician's guide, <https://www.gehealthcare.com/support/manuals?search=eyJzZWZyY2hUZlJtJoiMjA1NjI0Ni0wMDciLCJsYW5ndWFnZU5hbWUiOiJFbmdsaXNoIChFTikifQ%3D%3D> (2022, accessed 25 October 2024).
13. Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* 2017; 30: 3146–3154.
14. Yoo J, Yoo I, Youn I, et al. Residual one-dimensional convolutional neural network for neuromuscular disorder classification from needle electromyography signals with explainability. *Comput Methods Programs Biomed* 2022; 226: 107079.
15. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Paper/poster presented at: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.
16. Tan M and Le Q. Efficientnet: rethinking model scaling for convolutional neural networks. In: Kamalika C and Ruslan S (eds) *Proceedings of the 36th international conference on machine learning*. Proceedings of Machine Learning Research: PMLR, Cambridge, MA, USA, 2019, pp.6105–6114.
17. Hochreiter S and Schmidhuber J. Long short-term memory. *Neural Comput* 1997; 9: 1735–1780.
18. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res* 2016; 18: e323.
19. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Paper/poster presented at: 2017 IEEE international conference on computer vision (ICCV), 2017.
20. Olsen CR, Mentz RJ, Anstrom KJ, et al. Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure. *Am Heart J* 2020; 229: 1–17.
21. Sanders WE Jr, Burton T, Khosousi A, et al. Machine learning: at the heart of failure diagnosis. *Curr Opin Cardiol* 2021; 36: 227–233.
22. Hamatani Y, Nishi H, Iguchi M, et al. Machine learning risk prediction for incident heart failure in patients with atrial fibrillation. *JACC Asia* 2022; 2: 706–716.
23. Abubaker M and Babayigit B. Detection of cardiovascular diseases in ECG images using machine learning and deep learning methods. *IEEE Trans Artif Intell* 2022; 3: 248–260.
24. Kwon S, Hong J, Choi EK, et al. Deep learning approaches to detect atrial fibrillation using photoplethysmographic signals: algorithms development study. *JMIR mHealth uHealth* 2019; 7: e12770.
25. Siontis KC, Noseworthy PA, Attia ZI, et al. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat Rev Cardiol* 2021; 18: 465–478.
26. Adamo L, Rocha-Resende C and Prabhu SD. Reappraising the role of inflammation in heart failure. *Nat Rev Cardiol* 2020; 17: 269–285.
27. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019; 17: 230.
28. Choi EK. Cardiovascular research using the Korean national health information database. *Korean Circ J* 2020; 50: 754–772.
29. Lee SR, Choi EK, Han KD, et al. Edoxaban in Asian patients with atrial fibrillation: effectiveness and safety. *J Am Coll Cardiol* 2018; 72: 838–853.
30. Lee SR, Choi EK, Jung JH, et al. Lower risk of stroke after alcohol abstinence in patients with incident atrial fibrillation: a nationwide population-based cohort study. *Eur Heart J* 2021; 42: 4759–4768.