Forecasting left ventricular systolic dysfunction in heart failure with artificial intelligence

Payal Chandak[1,2,3] Abena Kyereme-Tuah[4], Judy Hung [2,5], Hanna Gaggin [2,5], Isaac S. Kohane[1,3,5,*], Collin M. Stultz[1,2,4,5,6,*]

1. Harvard-MIT Program in Health Sciences and Technology, Cambridge Ma. 02139
2. Division of Cardiology, Department of Medicine, Massachusetts General Hospital, Boston Ma. 02114
3. Department of Biomedical Informatics, Harvard Medical School, Boston Ma. 02115
4. Department of Electrical Engineering and Computer Science, Research Laboratory for Electronics, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge Ma. 02139
5. Harvard Medical School, Boston Ma. 02115
6. Institute for Medical Engineering and Science, MIT, Cambridge, Ma. 02129

*Corresponding Authors:

Collin M. Stutz
50 Vassar St. Room 36-736
Cambridge, Ma. 02139
cmstultz@csail.mit.edu

Isaac S. Kohane
10 Shattuck Road
Harvard Medical School
Boston, MA 02115
Isaac_Kohane@hms.harvard.edu

Abstract

Background. Objective assessment of left ventricular function remains a key prognosticator that is used to guide therapeutic decisions for patients with heart failure (HF). However, the Left ventricular ejection fraction (LVEF) is dynamic, with worsening LVEF linked to increased morbidity and mortality. Identifying patients at risk of LVEF decline would improve prognostication and enable timely therapeutic intervention.

Methods. We developed a deep learning model to Predict changes in left ventricULar Systolic function from ECGs of patients who have Heart Failure (PULSE-HF). The model integrates 12-lead ECG waveforms with a patient's history of prior LVEF measurements to calculate the likelihood that the LVEF will be less than 40% during the year after the ECG is obtained. The model is retrospectively developed and tested using data from one hospital and validated on retrospective cohorts from two different hospitals.

Findings. PULSE–HF demonstrates strong discriminatory ability with respect to forecasting whether the LVEF would fall below 40% within the next year, achieving AUROCs of 92% across all three HF cohorts. Among HF patients with baseline LVEF above 40%, PULSE–HF effectively identified those at risk of worsening LVEF with AUROCs of 89%, across all three datasets. PULSE–HF's discriminatory ability remained consistently high across a range of subgroups with different comorbidities and regardless of medical therapy. Assuming an underlying prevalence of LVEF worsening of 10% per year, PULSE-HF negative predictive values over 97% assuming an underlying sensitivity of 80%. Lastly, we demonstrate that a lead I version of PULSE–HF has a performance virtually indistinguishable from the model that uses all 12 ECG leads

Interpretation. PULSE–HF robustly predicts worsening LVEF in patients who have a prior diagnosis of HF. The method provides a platform for identifying patients who are at increased risk of worsening systolic dysfunction.

Funding: This work was supported, in part, by a grant from Quanta Computers.

Research in context

Evidence before this study:
Heart failure (HF) remains a leading cause of morbidity and mortality across the globe. Left ventricular ejection fraction (LVEF) plays an important role in prognostication and treatment of HF. Moreover, LVEF is dynamic, with worsening LVEF being associated with increased mortality and poor outcomes compared to HF patients with stable LVEF. Identifying HF patients at risk of future decline in LVEF would improve prognostication and facilitate the timely delivery of appropriate medical therapy. However, models that use the electronic health record (EHR) to forecast declines in LVEF have demonstrated modest discriminatory ability. For instance, a model incorporating echocardiographic features, demographics, and comorbidities from the EHR had AUROCs between 54%–68% for predicting a 30% decline in LVEF within one year after HF diagnosis. Electrocardiograms (ECG) may contain information about LVEF, as deep learning models of ECGs have been used to estimate the contemporaneous LVEF. Nevertheless, to date, the ECG has not been used to forecast forecast future changes in LVEF have not been described. Overall, no robust method exists for identifying HF patients at high risk of future LVEF decline.

Added value of this study:
We developed a deep learning model, called PULSE–HF, to Predict changes in left ventricULar Systolic function from ECGs of patients who have Heart Failure. The model leverages both the 12-lead ECG and history of prior LVEF measurements to identify patients at risk of having a reduced LVEF (<40%) within the year after the ECG was obtained. We also developed a single lead version of PULSE-HF, based on lead I, that has similar performance relative to the model that uses all 12 leads.

Implications of all the available evidence:
The ECG contains information that can be leveraged to forecast changes in LVEF. As ECGs are typically obtained for HF patients during yearly follow-up visits, we envision that the method could be used annually to identify patients who are at the highest risk of future LVEF decline. Such screening could be routinely integrated into existing clinical workflow with minimal changes to clinical practice. The prospect that similar performance can be achieved with a single lead model (lead-I) suggests that it may be possible to do such screening using wearable or pocket ECG monitoring devices. Prospective studies are needed to better define the utility of PULSE–HF in guiding HF management decisions, optimizing prediction windows, and integrating the model into clinical practice.

Introduction

Heart failure (HF) remains a leading cause of morbidity and mortality in individuals over 65 years of age (1). Despite advances in diagnosis and treatment, five-year mortality rates in patients with HF exceed 50% in many cohorts, highlighting the need for improved risk stratification and therapies (1). A cornerstone of HF management is an objective assessment of left ventricular systolic function, typically quantified as the left ventricular ejection fraction (LVEF). The LVEF is used to classify HF into three mutually exclusive subtypes,: HF with reduced ejection fraction (HFrEF, LVEF < 40%); HF with mildly reduced ejection fraction (HFmrEF, LVEF 40–49%); and HF with preserved ejection fraction (HFpEF, LVEF ≥ 50%) (2). Therapeutic decision making is a function of the precise subtype that a patient belongs; e.g., guideline directed medical therapy (GDMT) for patients with HFrEF has been shown to reduce mortality by 73% over two years (3). Moreover, LVEF plays an important role in prognostication, as decreases in LVEF below 45% are associated with proportional increases in the hazard ratio for all-cause mortality (4).

LVEF is dynamic and may increase with appropriate therapy or decrease as HF progresses over time (5,6). For example, longitudinal studies demonstrate that about one out of three patients with HFpEF experience a decline in their LVEF to below 40% within 2 years of diagnosis (7–9). Indeed, worsening LVEF been associated with an approximately 2 fold increase in mortality relative to stable HFpEF diagnosis, as well as worse outcomes compared to a stable HFrEF diagnosis (8–11). A method that accurately identifies HF patients at high risk of LVEF decline would improve prognostication, optimize echocardiogram scheduling, and enable timely initiation of GDMT.

Although a number of studies have identified clinical features that are correlated with decreasing LVEF, risk models that utilize these features to forecast a significant decline in LVEF have demonstrated modest discriminatory ability (9,11). For instance, an extreme gradient boosting (XGBoost) model incorporating electronic health record (EHR) data—including echocardiographic features, demographics, and comorbidities—had AUROCs between 54%–68% for predicting a 30% decline in LVEF within one year after HF diagnosis (12). These results suggest that tabular EHR data, including features known to be correlated with LVEF decline, have limited predictive value for the task of forecasting declines in LVEF.

Prior studies suggest that information about left ventricular systolic function is contained in the electrocardiogram (ECG) and deep learning models that use ECG signals can identify subclinical markers of systolic dysfunction that are not captured in structured EHR data (13,14). However, these models use the ECG to classify the concurrent LVEF rather than forecasting when the LVEF will decline (13,14). As such, they rely on LVEF measurements associated with echocardiograms that are acquired at a time close to when the ECG obtained and consequently are not designed to identify patients who will have a significant decline in their LVEF in the future. Moreover, these models were not specifically developed on patients who have heart failure.

We developed a deep learning model, called PULSE–HF, to Predict changes in left ventricULar Systolic function from ECGs of patients who have Heart Failure. PULSE–HF integrates 12-lead ECG waveforms with LVEF information from echocardiograms obtained before the ECG is acquired (i.e., the LVEF history), to forecast the likelihood that a patient's LVEF will be less than 40% within one year after the ECG. We retrospectively validate the model in HF cohorts from three different hospitals. Notably, among HF patients with baseline LVEF above 40%, PULSE–HF effectively identified those at risk of worsening LVEF in all three datasets. We also

developed a single-lead version of PULSE–HF (based on lead I), and demonstrate that its performance is similar to the 12-lead version. PULSE-HF provides a platform to identify patients with HF who are at increased risk of worsening systolic dysfunction.

Methods

Study design

PULSE-HF was developed and validated using clinical data from three institutions: Massachusetts General Hospital (Hospital 1), Brigham and Women's Hospital (Hospital 2), and a deidentified-dataset from Physionet (MIMIC IV, https://physionet.org/content/mimic-iv-ecg/1.0/). In all cohorts, only patients with a confirmed HF diagnosis, based on ICD codes, and who had at least one ECG and one LVEF measurement were included. Data from Hospital 1 was used to develop and test model, and data from the other two institutions were used to validate the approach. This retrospective study was approved by MGB Institutional Review Board protocol number 2020P000132, with a waiver of informed consent as the data were obtained from routine clinical care.

Model development

PULSE-HF is designed to predict if any LVEF measurements will be under 40% in the year after an ECG is obtained. Consequently, the task is formalized as a supervised learning problem where an ECG is labeled as positive if there is any echocardiogram that indicates an LVEF below 40% within the year of the date that the ECG was obtained, and negative otherwise. Labeled ECGs from Hospital 1 were split by patient into a development set comprising 85% of the data and an internal test set comprising 15%. The development set was further divided into a training dataset that was used to learn model parameters, and a validation dataset that was used for early stopping. ECGs from Hospital 2 and MIMIC served as external test datasets, ensuring no patient overlap during model training.

ECG waveforms across all cohorts were recorded for 10 seconds and sampled at a minimum of 250 Hz with amplitudes converted to millivolts. We used linear interpolation to fit each lead into 2,500 samples, and organize them into a 2500 × 12 matrix with lead order {I, II, III, AVR, AVL, AVF, V1, V2, V3, V4, V5, V6}. ECGs were excluded if any sample exceeded 5 mV in absolute value, if any lead was entirely zero or whose absolute value was below machine precision. If multiple ECGs were recorded within an hour, only the last one was retained.

In addition to the ECG waveform, we also used information from previously recorded LVEF measurements obtained from echocardiograms. More precisely, we define the LVEF history to include the following metrics acquired one year before the ECG: the initial LVEF value at diagnosis of HF, the LVEF value immediately preceding the ECG, the count, mean, minimum, maximum, standard deviation, and range of LVEFs in the past year, whether any LVEFs under ≤40%, between 40–50%, and above 50% in the past year, whether all LVEFs in the past year were classified as under ≤40%, between 40–50%, and above 50%, and the days since diagnosis of heart failure.

We trained a binary classification model using an ECG encoder (Figure S1A,B), a LVEF history encoder (Figure S2), and a prediction module (Figure S3). A ResNet-18 variant was used as the ECG encoder, with 12 input channels and 1D convolutions, producing an ECG embedding. A multilayer perceptron (MLP) encoder processed the LVEF history into another embedding. The ECG and LVEF history embeddings are then added together and then passed to an MLP module, to generate a prediction. A schematic of the final model is shown in Figure S3. We also

trained single-lead versions of the PULSE–HF model where the only change was to the ECG encoder using only one input channel.

Model parameters were optimized using a binary cross entropy loss with 5% weight decay, an AdamW optimizer, batch size of 2048, and learning rate scheduler with an initial learning rate of 10^{-5} and an end learning rate of 10^{-8}. Dropout regularization with probability of 0.3 and early stopping on validation AUROC with patience of 20 epochs was applied to mitigate overfitting. Model development utilized PyTorch and PyTorch Lightning, with source code publicly available.

To compare our approach to a standard, non-deep-learning based method, we trained a logistic regression model using clinical features extracted from ECGs and the same summary statistics of LVEF history that were used in PULSE–HF. Using Scikit-Learn, we performed a grid search over to optimize logistic regression by varying the regularization strength over the range 1E-3 to 1E3 and testing both L1 and L2 regularization using the Stochastic Average Gradient Accelerated (SAGA) solver and selecting the best model based on validation AUROC.

Lastly, to determine the relative importance of different input features, we performed a saliency analysis to interpret the contributions of different features in the Lead I PULSE-HF model (15,16). For each ECG, we computed the absolute gradient of the loss function at each input sample; i.e., a sample from the 12-lead ECG or a LVEF history feature. For ECG saliency values, we analyze the temporal distribution of high-saliency points, we aligned these points with R peaks detected via NeuroKit2. Specifically, we identified the most salient point in each ECG and measured its time distance to the next R peak. For LVEF features, we calculated the mean absolute saliency value for each feature and compared it to the mean top 1 absolute saliency value from ECGs. All analyses were done using Python.

Statistical analysis

Model performance was evaluated using the area under the receiver operating characteristic curve (AUROC). Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were computed across various operating points. Using the best-performing model and the same thresholds as in the development dataset, we assessed the PULSE–HF model on the internal test dataset at Hospital 1 and external test datasets at Hospital 2 and MIMIC. To evaluate robustness, we stratified performance across key clinical subgroups. We examined model performance among ECGs with a baseline LVEF above 40% to determine its ability to forecast worsening LVEF. Single-lead versions of PULSE–HF were similarly evaluated to determine feasibility for remote monitoring. All 95% confidence intervals were estimated via 1000 bootstraps of the respective test datasets.

Role of the funding source
Funding sources provided computational resources and student support. Funders played no role in the design of the work, analysis of the data, or writing the paper.

Results

PULSE-HF is designed to determine whether the LVEF will be below 40% within one year of when an ECG is acquired and previous LVEF measurements (Figure 1). The model is evaluated using retrospective data obtained from three hospitals.

Cohort characteristics

We evaluated PULSE–HF across three institutions; cohort characteristics across demographics, comorbidities, and medications are summarized in Table 1. At Hospital 1, the model was validated on an internal holdout dataset of 144,703 ECGs from 12,698 patients. At Hospital 2 and MIMIC, the model was validated on external holdout datasets of 274,897 ECGs from 26,942 patients and 196,584 ECGs from 20,850 patients, respectively. The cohorts had statistically significant differences in age, sex, LVEF distribution, comorbidity diagnosis, and medication prescription. While these cohorts are broadly representative of heart failure populations, the variation in characteristics underscores the need for a predictive model that can adapt to diverse patient populations.

PULSE–HF forecasts low LVEF in HF patients

We first explored whether future LVEF could be accurately forecasted using ECG features and LVEF without deep learning. A logistic regression model using summary ECG features developed on the same task achieved modest AUROCs (Hospital 1 71.4%; Hospital 2 68.5%). Incorporating the same LVEF history features used in PULSE–HF improved AUROCs (Hospital 1 88.1%; Hospital 2: 87.4%), thereby demonstrating that LVEF history provides important information for identifying low LVEF within a year of the index ECG.

PULSE–HF, which leverages both ECG and LVEF history, demonstrates strong discriminatory ability with respect to forecasting whether the LVEF would fall below 40% within the next year, achieving AUROCs of 92.0%–92.6% across all three institutions (Table 2). Moreover, in all three cohorts, the discriminatory ability remained consistently high across a range of subgroups with different comorbidities and regardless of medical therapy (Table 2). Further, discriminatory ability remained consistent in individual ECGs that exhibited atrial fibrillation or pacing in the Hospital 1 and Hospital 2 datasets (Table 3)

PULSE–HF forecasts worsening LVEF in patients with baseline LVEF above 40%

PULSE-HF detects when the LVEF will fall below 40% in the year after an ECG is obtained. If the LVEF is already below 40% when the initial ECG is acquired, then a positive PULSE-HF prediction corresponds to forecasting that the LVEF will remain below 40% in the following year. By contrast, if the initial LVEF is above 40%, then a positive prediction for PULSE-HF forecasts that the LVEF will decline to below 40% in the subsequent year; i.e, that the LVEF will worsen.

To determine how well PULSE-HF can forecast worsening LVEF, we focus on the subgroup of patients where the most recent echocardiogram is associated with a “baseline” LVEF above 40%. Cohort characteristics across demographics, comorbidities, and medications among patients who had at least one LVEF above 40% after their HF diagnosis are summarized in Table 4. Clinical characteristics within this cohort were similar to the overall population at each corresponding site. Moreover, across the three clinical cohorts, approximately 20% of patients who begin with an LVEF over 40% eventually develop an LVEF below 40% over median and mean follow-up times of 10 and 22 months, respectively. The average rate of LVEF decline per year was about 10%, a finding consistent with prior longitudinal studies (11,12).

For the task of forecasting a decline in the LVEF, PULSE-HF exhibited strong discriminative performance with AUROCs of 89.0%–89.3% (Figure 2A) and AUPRCs of 68.1%–82.0% (Figure 2B). For comparison, regression models achieved AUROCs 60.0%–65.4% when using ECG features only, and AUROCs 67.7%–75.4% when using ECG and LVEF history features. In

addition, the discriminatory ability remained consistently high across a range of subgroups with different comorbidities and regardless of medical therapy (Table 5). PULSE-HF also demonstrated consistent sensitivity–specificity tradeoffs across all three institutions, with specificity ranging from 85.9%–90.9% at 70% sensitivity, 72.3%–84.1% at 80% sensitivity, and 45.7%–69.7% at 90% sensitivity (Figure 2C).

We computed negative predictive values (NPVs) and positive predictive values (PPVs) for PULSE-HF as a function of model sensitivity. As our data, as well as other longitudinal studies (12), suggest that the prevalence of significant decrements in LVEF over one year is approximately ~10%, we calculate predictive values for this underlying prevalence. Calculated NPVs were 96.0%–97.0% at 70% sensitivity, 97.2%–98.0% at 80% sensitivity, and 98.2%–98.9% at 90% sensitivity (Figure 2D). Corresponding PPVs range between 37.6%–44.7%, at 70% sensitivity, 25.7%–35.4% at 80% sensitivity, and 16.3%–24.5% at 90% sensitivity (Figure 2E).

Given the relatively low PPVs, we evaluated how the number of positive predictions correlate with predictive performance. Across various sensitivity thresholds, PPV increased consistently as positive predictions accumulated (Figure 2F). With only one positive prediction over the previous two years, PUSLE–HF's PPVs were 45.8%, 55.7%, and 62.9% at 70% sensitivity in Hospital 1, Hospital 2, and MIMIC. With three positive predictions in that time, these PPVs improved to 59.1%, 61.5%, and 64.7% respectively.

A Lead I version of PULSE-HF effectively forecasts worsening LVEF

To determine the relative importance of each lead for predicting worsening LVEF, we constructed twelve single-lead versions of PULSE-HF, each one leveraging just one lead from the 12-lead ECG, to predict whether the LVEF will be below 40% in the year after the ECG is obtained. Across all twelve single-lead models, discriminatory ability in forecasting LVEF below 40% remained virtually indistinguishable from the full 12-lead model (Figure S4).

To further evaluate whether a single lead model can forecast worsening LVEF, we use the version of PULSE-HF trained using lead-I alone, PULSE-HFL1, as most wearable and pocket devices equipped with ECG sensors record lead 1 (17,18). For the task of forecasting LVEF deterioration among patients with a baseline LVEF above 40%, the Lead I model maintained its discriminative ability relative to the full 12-lead model, achieving AUROCs of 89.2%–89.4% (Figure 3A) and AUPRCs of 68.6%–82.5% (Figure 3B). Sensitivity-specificity tradeoffs remained stable, with specificity ranging from 86.4%–91.2% at 70% sensitivity, 73.0%–83.5% at 80% sensitivity, and 43.7%–66.1% at 90% sensitivity (Figure 3C). At a 10% prevalence of worsening per year, NPVs ranged from 96.2%–97.1% at 70% sensitivity, 97.3%–97.9% at 80% sensitivity, and 98.3%–98.9% at 90% sensitivity (Figure 3D). PPVs were 38.5%–46.0% at 70% sensitivity, 26.1%–34.8% at 80% sensitivity, and 15.9%–22.8% at 90% sensitivity (Figure 3E).

We again computed PPVs for repeated or PULSE-HF positive predictions as a function of model sensitivity. PULSE–HFL1 demonstrated similar improvements to the 12-lead version of PULSE-HF, achieving PPVs of 61.5%–65.9% with three or more positive predictions over two years (Figure 3F). Both PULSE–HF and PULSE–HFL1 respectively achieved PPVs of 61.9%–64.9% and 63.0%–68.1% with three or more positive predictions over one year of monitoring.

Determinants of PULSE-HF performance

To understand what aspects of the input data (ECG and LVEF history) contribute most to PULSE-HFL1's predictive performance we use saliency maps to identify regions of the ECG signal and specific characteristics of prior echocardiograms to identify features that the model tends to "look at" when making predictions. Our saliency analysis reveals that PULSE-HFL1 primarily focuses on ECG features occurring during diastole, particularly within 200ms before the following R-wave, the part of the cardiac cycle when ventricular filling is nearly complete and abnormalities in left ventricular compliance and relaxation may be most pronounced (

Figure 4A). Notably, saliency values for ECG-based features are one order of magnitude larger than those for LVEF history, indicating that subtle temporal variations in the ECG signal provide meaningful prognostic information beyond what LVEF alone can capture (

Figure 4B). Among LVEF history features, the model primarily focuses on the initial LVEF measurement and the LVEF immediately preceding the index ECG. Overall, these findings highlight that while PULSE-HFL1 integrates both ECG and LVEF history, its decisions are largely driven by ECG-derived features, particularly in diastole.

Discussion

PULSE–HF leverages 12-lead ECGs and LVEF history to predict whether a patient's LVEF will decline below 40% within one year after the index ECG. The model is designed specifically for prognostication in patients who have a prior diagnosis of heart failure patients and was validated across three retrospective datasets of patients who have a prior diagnosis of heart failure. Among these cohorts, the discriminatory ability was high (AUROCs ~92%) and maintained across subgroups with different comorbidities and medical therapies.

Arguably, the most important application for PULSE–HF is in identifying patients who will have a decline in their LVEF in the future. In those patients who have an initial LVEF over 40%, PULSE-HF can discriminate between patients who will have a subsequent decline in their LVEF vs those who will not (AUROCs ~89%). The high discriminatory ability is again preserved over all three clinical cohorts and also in across subgroups with different comorbidities. Prior attempts to forecast worsening LVEF using structured EHR data have shown limited success, with AUROCs ranging from 54% to 68%, underscoring the importance of ECGs in capturing subclinical markers of systolic dysfunction. Moreover, a logistic regression trained using ECG features and information about the LVEF history achieves AUROCs of 67.7%–75.4%. Assuming an underlying incidence for LVEF worsening of 10% per year, and a 70% sensitivity, PULSE–HF demonstrated modest PPVs (~45%) and high NPVs (~97%) across all three institutions. Patients who have multiple positive predictions have higher PPVs (~65%). These data suggest that PULSE-HF provides a robust approach for identifying patients at low risk of future LVEF decline.

The ability to predict LVEF decline using single lead ECGs would facilitate the identification of high-risk patients using wearable and/or pocket ECG monitors, which most often record ECGs signals using a single lead. All twelve single lead versions of PULSE-HF had discriminatory ability that was very similar to the performance of the 12-lead model. Moreover, the

performance of Lead I version of PULSE–HF achieved AUROCs, PPVs and NPVs comparable to the full 12-lead model. This highlights the potential for integrating PULSE–HF into remote monitoring devices, enabling continuous risk assessment outside of traditional healthcare environments.

Current AHA guidelines recommend reassessment only when it would meaningfully alter clinical management (19). Patients who are deemed to be low risk of adverse outcome may therefore receive fewer follow-up echocardiograms relative to those who are deemed to be high risk. Our ability to detect LVEF decline promptly is limited by our ability risk stratify based on clinical signs, symptoms and the EHR. While EHR data may be challenging to wrangle, obtaining 12-lead ECGs have become a routine part of annual visits to a cardiovascular clinic. By leveraging ECGs, PULSE-HF provides a low-cost annual screening method to longitudinally identify HF patients at high risk of LVEF decline. Patients identified as high risk by PULSE–HF would be prioritized for follow-up echocardiography, thereby enabling the timely identification of HFrEF. By detecting left ventricular dysfunction sooner, clinicians could initiate GDMT earlier, which has been shown to improve survival, reduce heart failure hospitalizations, slow disease progression, and enhance quality of life for patients with HFrEF (20).

While deep learning models provide a platform for building performant models using ECG data, it is challenging to understand precisely what these models have learned and how they arrive at decisions. Insights into what aspects of the input signal the model tends to focus on when arriving at a decision can sometimes be garnered from a saliency analysis, which strives to understand what input features the model most “looks at” when making a prediction (15,16,21). In the Lead I PULSE-HF version, the model tends to focus on regions of the ECG signal within 200ms before the subsequent R-wave – the time corresponding to end diastole. This suggests a potential link between impairments in left ventricular filling and future declines in systolic function. While LVEF features are significantly less influential than ECG features, the LVEF measurement taken immediately before the ECG is the most relevant.

Our study has limitations. Our method is retrospectively evaluated on three clinical cohorts, corresponding to observational datasets. As such, the timing in which echocardiograms were obtained was determined by subjective clinical assessment and clinical practice, which can vary at different institutions. Differences in care practices and resource availability at other sites could introduce biases that warrant further investigation. Prospective evaluation is needed necessary to assess the utility of PULSE–HF in guiding HF management decisions, optimizing prediction windows, and integrating the model into clinical practice. While the single-lead models developed from 12-lead data are highly effective, it remains to be seen whether the same performance can be achieved with real-world, potentially noisier wearable ECG data.

PULSE–HF is a performant AI-based approach for forecasting worsening systolic function in HF patients and leverages a widely available test – the ECG – for risk stratification. PULSE–HF has the potential to improve HF outcomes by enabling timely interventions before irreversible cardiac deterioration occurs.

Data sharing statement

Raw electrocardiogram and electronic health records data can be shared pending IRB approval at the relevant hospital from which the data was collected. Please reach out to the corresponding authors for more information. PULSE–HF model weights and development code are freely available at https://github.com/payalchandak/PULSE-HF.

Acknowledgements

This work has benefited significantly from many contributions, including collaborations with Leo Anthony Celi, Takeshi Tohyama, Ahram Han, and Brian Gow; and discussions with Matthew McDermott and Noa Dagan; and many others.

Declaration of interests

The authors have no conflicts of interest to declare.

Figures

Table 1. Cohort characteristics.

	Hospital 1	Hospital 2	MIMIC
Patients	83,433 (development 70,918; testing 12,515)	26,942	20,850
Female	43.4%	45.2%	47.2%
Age	70.2 ± 15.6	70.8 ± 14.4	74.2 ± 13.4
Number of ECGs	947,905	274,897	196,584
Number of LVEFs	235,476	192,060	27,312
Diabetes mellitus	41.4%	39.0%	41.2%
Hypertension	83.8%	67.4%	61.4%
Atherosclerotic heart disease	68.3%	60.4%	59.2%
Chronic obstructive pulmonary disease	12.4%	5.7%	10.8%
Atrial fibrillation	49.2%	45.8%	51.3%
ACEi/ARB/ARNi	44.4%	40.1%	34.8%
Beta blockers	52.9%	49.6%	56.0%
MRA	13.2%	8.3%	7.8%
Diuretics	56.5%	60.6%	48.7%

Table 2. Discriminatory ability for forecasting low LVEF in all patients, stratified by important clinical variables.

Subgroup		AUROC		
		Hospital 1	Hospital 2	MIMIC
All patients		92.6 ± 0.2	92.2 ± 0.2	92.0 ± 0.2
Average annual hospitalizations	< 1	92.4 ± 0.3	92.7 ± 0.2	92.2 ± 0.3
	>= 1	92.8 ± 0.4	91.1 ± 0.3	91.8 ± 0.3
Diabetes mellitus	No	92.6 ± 0.3	92.4 ± 0.3	91.7 ± 0.3
	Yes	92.7 ± 0.3	92.1 ± 0.3	92.6 ± 0.3
Hypertension	No	90.4 ± 0.7	92.5 ± 0.4	92.1 ± 0.4
	Yes	92.8 ± 0.2	92.0 ± 0.3	92.0 ± 0.3
Atherosclerotic heart disease	No	91.5 ± 0.6	92.4 ± 0.4	91.4 ± 0.4
	Yes	92.8 ± 0.2	92.0 ± 0.3	92.1 ± 0.3
Chronic obstructive pulmonary disease	No	92.3 ± 0.2	92.3 ± 0.2	92.1 ± 0.2
	Yes	94.2 ± 0.6	91.9 ± 0.9	92.1 ± 1.2
Atrial fibrillation	No	92.8 ± 0.3	93.1 ± 0.3	92.4 ± 0.3
	Yes	92.5 ± 0.3	91.5 ± 0.3	91.7 ± 0.3
ACEi/ARB/ARNi	No	91.9 ± 0.4	92.3 ± 0.2	91.9 ± 0.2
	Yes	92.6 ± 0.3	91.3 ± 0.3	92.1 ± 0.5
MRA	No	91.5 ± 0.3	92.0 ± 0.2	91.8 ± 0.2
	Yes	93.5 ± 0.4	91.7 ± 0.5	96.1 ± 0.7
Beta blockers	No	92.4 ± 0.4	93.0 ± 0.2	91.9 ± 0.2
	Yes	92.4 ± 0.3	91.3 ± 0.3	92.0 ± 0.4
Diuretics	No	91.4 ± 0.5	91.9 ± 0.3	91.8 ± 0.2
	Yes	92.6 ± 0.3	91.9 ± 0.2	92.8 ± 0.5

Table 3. Discriminatory ability of PULSE–HF in ECGs with pacing and atrial fibrillation, as identified by clinician diagnoses accompanied with ECGs. These diagnoses were not available in MIMIC.

			AUROC	
			Hospital 1	Hospital 2
All patients	ECGs contain pacing	No	92.9 ± 0.2	91.9 ± 0.2
		Yes	92.7 ± 0.7	92.4 ± 0.5
	ECGs contain atrial fibrillation	No	93.1 ± 0.2	92.3 ± 0.2
		Yes	92.0 ± 0.8	91.5 ± 0.4
Patients with LVEF above 40%	ECGs contain pacing	No	89.5 ± 0.5	89.2 ± 0.3
		Yes	87.3 ± 1.7	88.3 ± 0.8
	ECGs contain atrial fibrillation	No	89.5 ± 0.4	89.5 ± 0.3
		Yes	88.8 ± 1.3	87.9 ± 0.6

	Hospital 1	Hospital 2	MIMIC
Patients	54,100	13,675	10,795
Female	45.4%	49.4%	52.2%
Age	70.2 ± 15.5	70.1 ± 14.7	74.2 ± 13.5
Number of ECGs	772033	168381	129151
Number of LVEFs	213723	165416	21870
Diabetes mellitus	43.7%	40.5%	42.5%
Hypertension	88.4%	74.0%	70.2%
Atherosclerotic heart disease	70.5%	59.3%	58.0%
Chronic obstructive pulmonary disease	14.9%	7.3%	10.2%
Atrial fibrillation	52.9%	49.1%	53.0%
ACEi/ARB/ARNi	45.8%	38.5%	37.0%
Beta blockers	53.6%	49.5%	58.9%
MRA	13.0%	7.5%	7.3%
Diuretics	58.7%	60.0%	52.9%

Table 4. Cohort characteristics in patients with any LVEF > 40 after diagnosis.

Table 5. Discriminatory ability for forecasting worsening LVEF in patients with LVEF > 40, stratified by important clinical variables.

Subgroup		AUROC		
		Hospital 1	Hospital 2	MIMIC
All patients		89.0 ± 0.4	89.2 ± 0.3	89.3 ± 0.3
Average annual hospitalizations	< 1	90.0 ± 0.4	90.6 ± 0.3	91.3 ± 0.4
	>= 1	86.2 ± 0.9	85.0 ± 0.6	87.0 ± 0.5
Diabetes mellitus	No	89.5 ± 0.6	90.0 ± 0.4	89.6 ± 0.4
	Yes	88.5 ± 0.6	88.1 ± 0.4	89.0 ± 0.5
Hypertension	No	90.7 ± 0.9	90.5 ± 0.5	90.0 ± 0.5
	Yes	88.5 ± 0.5	88.3 ± 0.4	88.8 ± 0.4
Atherosclerotic heart disease	No	89.5 ± 0.9	89.8 ± 0.5	89.2 ± 0.6
	Yes	88.9 ± 0.5	88.6 ± 0.3	88.9 ± 0.4
Chronic obstructive pulmonary disease	No	89.0 ± 0.4	89.3 ± 0.3	89.3 ± 0.3
	Yes	88.8 ± 1.3	87.5 ± 1.1	88.6 ± 1.6
Atrial fibrillation	No	90.5 ± 0.6	91.0 ± 0.4	89.9 ± 0.4
	Yes	87.7 ± 0.6	87.2 ± 0.4	88.8 ± 0.5
ACEi/ARB/ARNi	No	90.6 ± 0.5	90.6 ± 0.3	89.1 ± 0.3
	Yes	87.0 ± 0.7	86.8 ± 0.5	90.5 ± 0.8
MRA	No	89.0 ± 0.4	89.5 ± 0.3	89.2 ± 0.3
	Yes	89.4 ± 1.0	85.8 ± 1.1	94.7 ± 1.2
Beta blockers	No	90.0 ± 0.6	91.1 ± 0.3	89.2 ± 0.4
	Yes	88.1 ± 0.6	86.8 ± 0.4	89.5 ± 0.6
Diuretics	No	90.0 ± 0.7	90.5 ± 0.4	89.1 ± 0.4
	Yes	88.1 ± 0.5	88.2 ± 0.3	90.1 ± 0.7

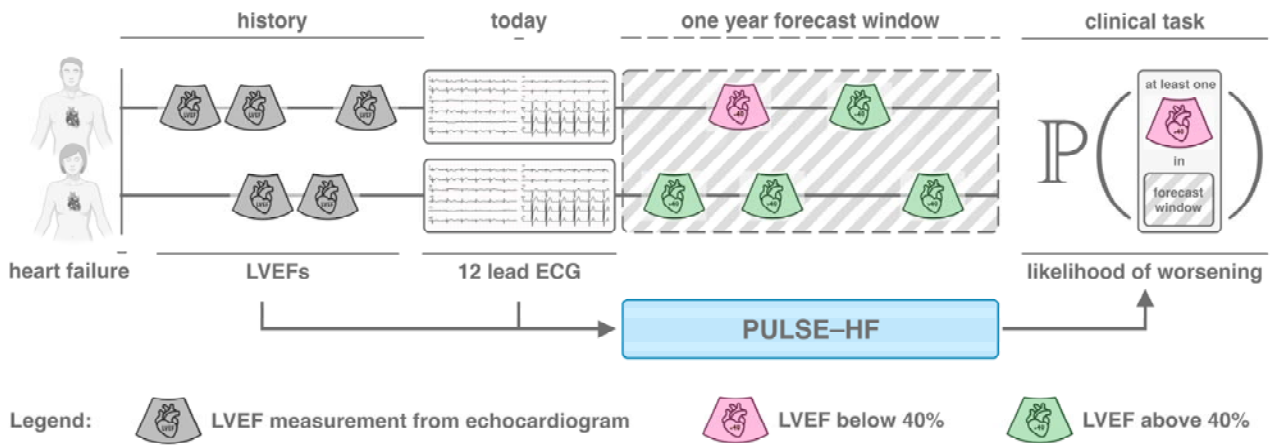

Figure 1. Overview of PULSE–HF. A deep learning model designed to Predict changes in left ventricULar Systolic function from Electrocardiograms of patients with Heart Failure. PULSE–HF integrates 12-lead ECGs with LVEF history to forecast the likelihood of LVEF falling below 40% within one year after the ECG.

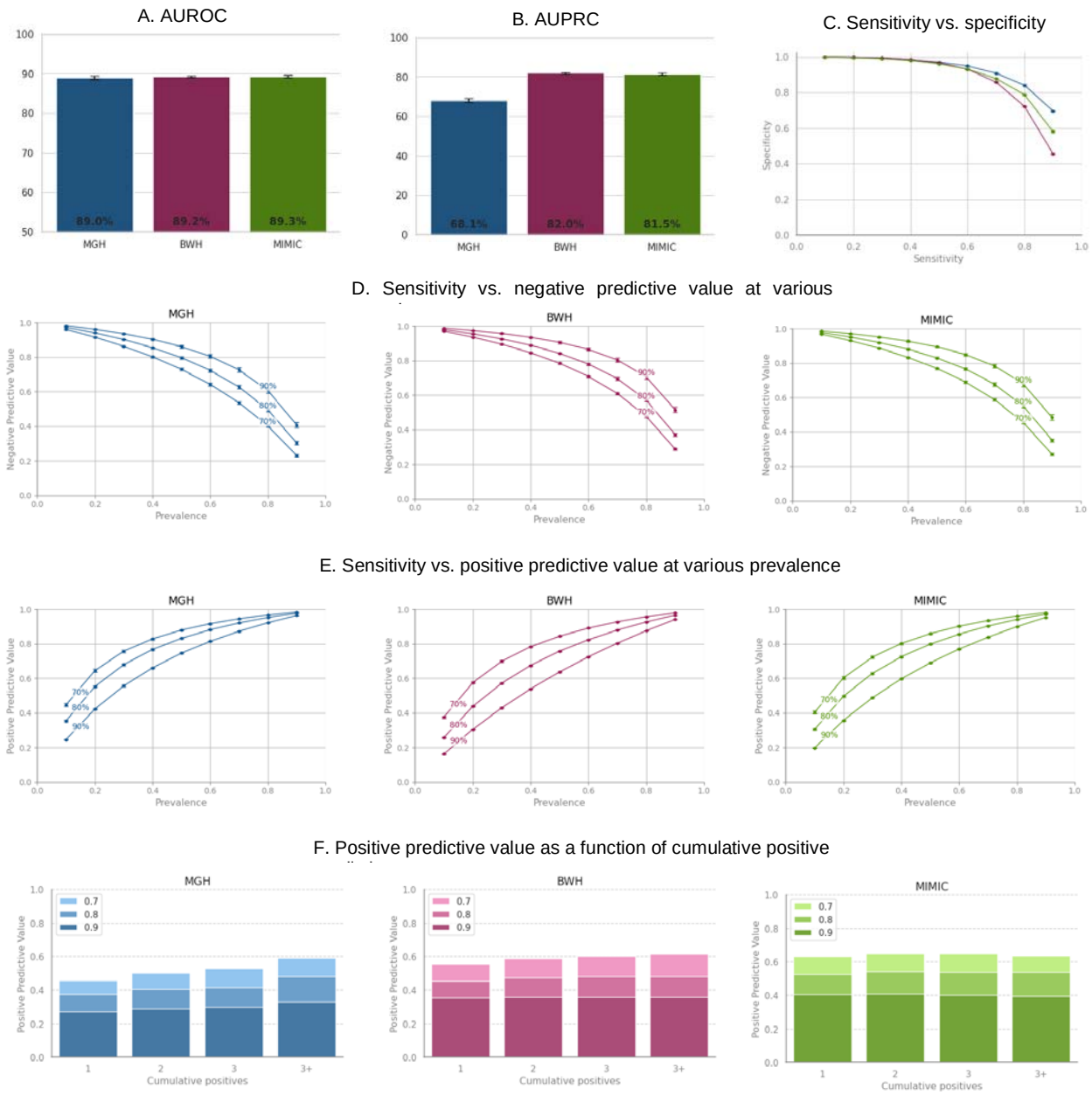

Figure 2. PULSE-HF performance in patients who have a baseline LVEF above 40%. Results are shown for three institutions: the internal test set at Hospital 1 and external validation at Hospital 2 and MIMIC. (A) Area under the receiver operating characteristic curve (AUROC). (B) Area under the precision-recall curve (AUPRC). (C) Sensitivity vs. specificity trade-off curves. (D) Sensitivity vs. negative predictive value (NPV) at various assumed prevalences of one-year incidence of LVEF decline below 40%. (E) Sensitivity vs. positive predictive value (PPV) at different assumed prevalences of one-year incidence of LVEF decline below 40%. (F) PPV as a function of cumulative positive predictions, demonstrating that multiple positive predictions improve PULSE–HF's risk stratification.

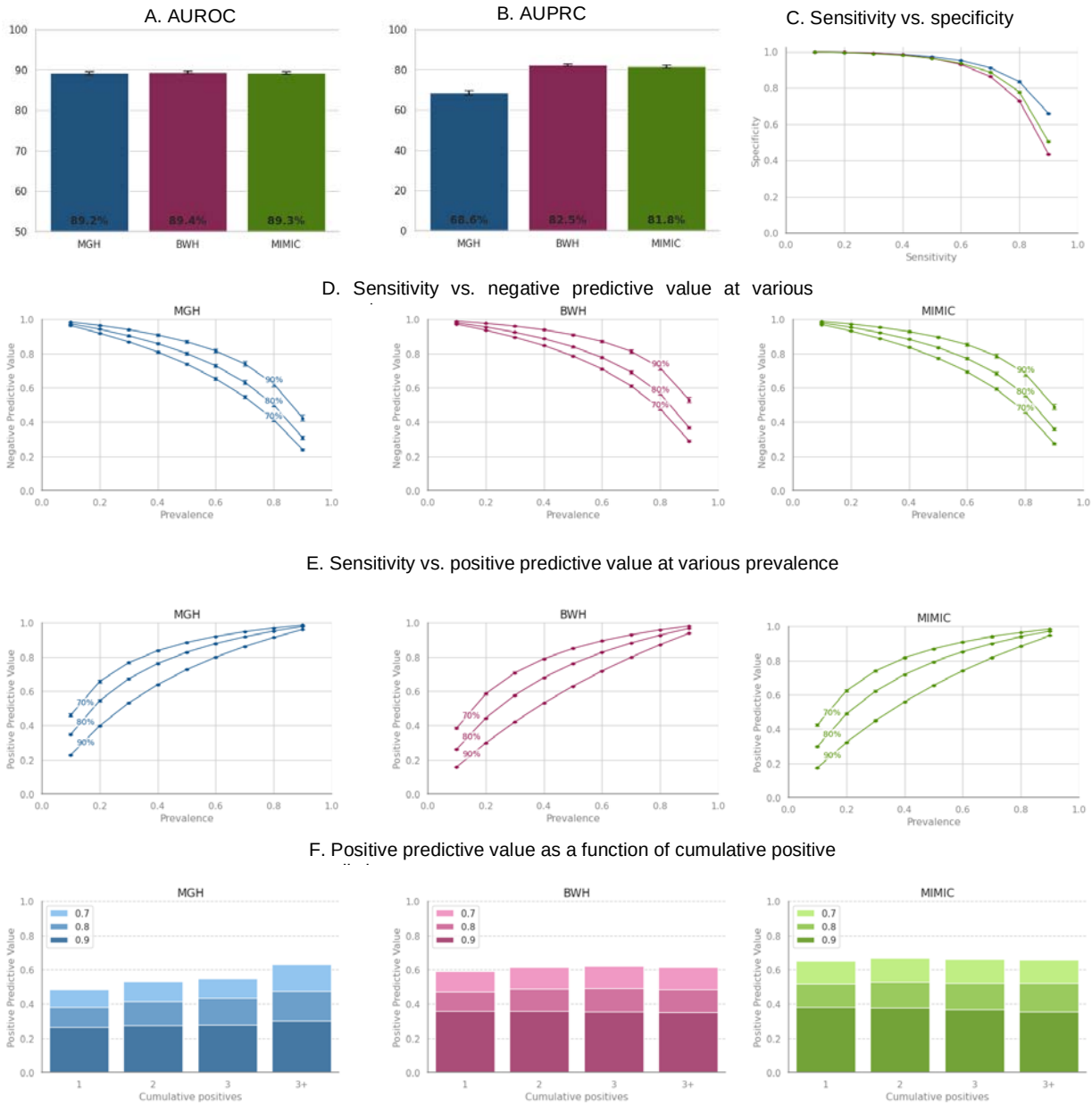

Figure 3. PULSE-HF Lead 1 model performance in patients who have a baseline LVEF above 40%. Results are shown for three hospitals. (A) Area under the receiver operating characteristic curve (AUROC). (B) Area under the precision-recall curve (AUPRC). (C) Sensitivity vs. specificity trade-off curves. (D) Sensitivity vs. negative predictive value (NPV) at various assumed prevalences of one-year incidence of LVEF decline below 40%. (E) Sensitivity vs. positive predictive value (PPV) at different assumed prevalences of one-year incidence of LVEF decline below 40%. (F) PPV as a function of cumulative positive predictions, demonstrating that multiple positive predictions improve PULSE–HF's risk stratification.

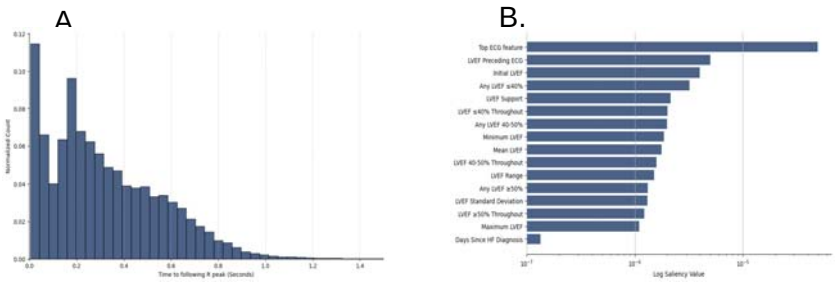

Figure 4. Determinants of PULSE–HF performance (A) Distribution of the time difference between the most salient ECG feature used by PULSE–HF and the R-peak that immediately follows this feature. (B) Mean saliency values of LVEF history features and the ECG feature with the highest saliency.

References

1. Savarese G, Becher PM, Lund LH, Seferovic P, Rosano GMC, Coats AJS. Global burden of heart failure: a comprehensive and updated review of epidemiology. Cardiovasc Res. 2023 Jan 18;118(17):3272–87.

2. Bozkurt B, Coats AJ, Tsutsui H, Abdelhamid M, Adamopoulos S, Albert N, et al. Universal Definition and Classification of Heart Failure. J Card Fail. 2021 Apr;27(4):387–413.

3. Murphy SP, Ibrahim NE, Januzzi JL. Heart Failure With Reduced Ejection Fraction: A Review. JAMA. 2020 Aug 4;324(5):488.

4. Solomon SD, Anavekar N, Skali H, McMurray JJV, Swedberg K, Yusuf S, et al. Influence of Ejection Fraction on Cardiovascular Outcomes in a Broad Spectrum of Heart Failure Patients. Circulation. 2005 Dec 13;112(24):3738–44.

5. Jiang Y, Chen X, Zhang X, Dong S, Liu Y. Dynamic trajectories of left ventricular ejection fraction in heart failure with improved ejection fraction. BMC Cardiovasc Disord. 2024 Nov 5;24(1):621.

6. Lupón J, Gavidia-Bovadilla G, Ferrer E, De Antonio M, Perera-Lluna A, López-Ayerbe J, et al. Dynamic Trajectories of Left Ventricular Ejection Fraction in Heart Failure. J Am Coll Cardiol. 2018 Aug;72(6):591–601.

7. Savarese G, Stolfo D, Sinagra G, Lund LH. Heart failure with mid-range or mildly reduced ejection fraction. Nat Rev Cardiol. 2022 Feb;19(2):100–16.

8. Breathett K, Allen LA, Udelson J, Davis G, Bristow M. Changes in Left Ventricular Ejection Fraction Predict Survival and Hospitalization in Heart Failure With Reduced Ejection Fraction. Circ Heart Fail. 2016 Oct;9(10):e002962.

9. Chen L, Huang Z, Zhao X, Liang J, Lu X, He Y, et al. Predictors and Mortality for Worsening Left Ventricular Ejection Fraction in Patients With HFpEF. Front Cardiovasc Med. 2022 Feb 24;9:820178.

10. DeVore AD, Hellkamp AS, Thomas L, Albert NM, Butler J, Patterson JH, et al. Improvement in Left Ventricular Ejection Fraction in Outpatients With Heart Failure With Reduced Ejection Fraction. Circ Heart Fail [Internet]. 2020 Jul [cited 2023 Dec 12]; Available from: https://www.ahajournals.org/doi/abs/10.1161/CIRCHEARTFAILURE.119.006833

11. Savarese G, Vedin O, D'Amario D, Uijl A, Dahlström U, Rosano G, et al. Prevalence and Prognostic Implications of Longitudinal Ejection Fraction Change in Heart Failure. JACC Heart Fail. 2019 Apr;7(4):306–17.

12. Adekkanattu P, Rasmussen LV, Pacheco JA, Kabariti J, Stone DJ, Yu Y, et al. Prediction of left ventricular ejection fraction changes in heart failure patients using machine learning and electronic health records: a multi-site study. Sci Rep. 2023 Jan 6;13(1):294.

13. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, et al. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nat Med. 2019 Jan;25(1):70–4.

14. Vaid A, Johnson KW, Badgeley MA, Somani SS, Bicak M, Landi I, et al. Using Deep-Learning Algorithms to Simultaneously Identify Right and Left Ventricular Dysfunction From the Electrocardiogram. JACC Cardiovasc Imaging. 2022 Mar;15(3):395–410.

15. Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps [Internet]. arXiv; 2014 [cited 2025 Mar 16]. Available from: http://arxiv.org/abs/1312.6034

16. Vijayarangan S, Murugesan B, R V, SP P, Joseph J, Sivaprakasam M. Interpreting Deep Neural Networks for Single-Lead ECG Arrhythmia Classification [Internet]. arXiv; 2020 [cited 2025 Mar 16]. Available from: http://arxiv.org/abs/2004.05399

17. Perez MV, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, et al. Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation. N Engl J Med. 2019 Nov 14;381(20):1909–17.

18. Bumgarner JM, Lambert CT, Hussein AA, Cantillon DJ, Baranowski B, Wolski K, et al. Smartwatch Algorithm for Automated Detection of Atrial Fibrillation. J Am Coll Cardiol. 2018 May;71(21):2381–8.

19. Heidenreich PA, Bozkurt B, Aguilar D, Allen LA, Byun JJ, Colvin MM, et al. 2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. Circulation [Internet]. 2022 May 3 [cited 2025 Mar 3];145(18). Available from: https://www.ahajournals.org/doi/10.1161/CIR.0000000000001063

20. Hage C, Shahim A. Importance of early diagnosis and implementation of guideline-directed therapies in heart failure: a message worth repeating, but we need to act! Heart. 2025 Mar 4;heartjnl-2024-325649.

21. Schlesinger DE, Stultz CM. Deep Learning for Cardiovascular Risk Stratification. Curr Treat Options Cardiovasc Med. 2020 Aug;22(8):15.