Performance of off-the-shelf machine learning architectures and biases in low left ventricular ejection fraction detection

Jake A. Bergquist, PhD,[*†‡] Brian Zenger, MD, PhD,[§] James Brundage,[§]
Rob S. MacLeod, PhD,[†‡] T. Jared Bunch, MD,[§] Rashmee Shah, MD, MS,[§]
Xiangyang Ye, PhD,[§] Ann Lyons, PhD,[||] Michael Torre, PhD,[¶] Ravi Ranjan, MD, PhD,[†‡§]
Tolga Tasdizen, PhD,[**] Benjamin A. Steinberg, MD, MHS[§]

From the [*]Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, Utah, [†]Nora Eccles Treadwell Cardiovascular Research and Training Institute, University of Utah, Salt Lake City, Utah, [‡]Department of Biomedical Engineering, University of Utah, Salt Lake City, Utah, [§]School of Medicine, University of Utah, Salt Lake City, Utah, [||]Data Science Services, University of Utah, Salt Lake City, Utah, [¶]Department of Internal Medicine, University of Utah, Salt Lake City, Utah, and [**]Department of Electrical and Computer Engineering, University of Utah, Salt Lake City, Utah.

BACKGROUND Artificial intelligence–machine learning (AI-ML) has demonstrated the ability to extract clinically useful information from electrocardiograms (ECGs) not available using traditional interpretation methods. There exists an extensive body of AI-ML research in fields outside of cardiology including several open-source AI-ML architectures that can be translated to new problems in an "off-the-shelf" manner.

OBJECTIVE We sought to address the limited investigation of which if any of these off-the-shelf architectures could be useful in ECG analysis as well as how and when these AI-ML approaches fail.

METHODS We applied 6 off-the-shelf AI-ML architectures to detect low left ventricular ejection fraction (LVEF) in a cohort of ECGs from 24,868 patients. We assessed LVEF classification and explored patient characteristics associated with inaccurate (false positive or false negative) LVEF prediction.

RESULTS We found that all of these network architectures produced LVEF detection area under the receiver-operating characteristic curve values above 0.9 (averaged over 5 instances per network), with the ResNet 18 network performing the highest (average area under the receiver-operating characteristic curve of 0.917). We also observed that some patient-specific characteristics such as race, sex, and presence of several comorbidities were associated with lower LVEF prediction performance.

CONCLUSIONS This demonstrates the ability of off-the-shelf AI-ML architectures to detect clinically useful information from ECGs with performance matching contemporary custom-build AI-ML architectures. We also highlighted the presence of possible biases in these AI-ML approaches in the context of patient characteristics. These findings should be considered in the pursuit of efficient and equitable deployment of AI-ML technologies moving forward.

KEYWORDS Machine learning; Artificial intelligence; Explainability; Electrocardiogram; Heart failure

(Heart Rhythm O^2 2024;5:644–654) © 2024 Heart Rhythm Society. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction

Artificial intelligence–machine learning (AI-ML) is a computational technique that has been demonstrated to be able to extract meaningful clinical information from diagnostic data that are not available using either human interpretation or more simple analysis methods.[1–5] AI-ML has demonstrated remarkable successes across many clinical domains, including the 12-lead electrocardiogram (ECG).[4,6,7] Recent developments have shown that AI-ML approaches applied to ECGs can accurately predict different patient characteristics and pathologies not detectable by expert physician readers, including age, sex, and low left ventricular ejection fraction (LVEF).[4,6–8] Identification of such indicators of heart health as low LVEF from the ECG has been a target of research for years, with traditional methods seeking to identify specific ECG wave changes associated with LVEF changes.[9–11] As AI-ML tools emerge as promising ECG analysis methods, many researchers and institutions are seeking to apply them to a myriad of clinical and research tasks. Some AI-ML tools pending Food and Drug

Address reprint requests and correspondence: Dr Jake A. Berquist, 72 Central Campus Drive, Salt Lake City, Utah 84112. E-mail address: jbergquist@sci.utah.edu.

https://doi.org/10.1016/j.hroo.2024.07.009

KEY FINDINGS

- Off-the-shelf machine learning architectures designed for image analysis can be readily applied to electrocardiography analysis with favorable results to contemporary electrocardiography–machine learning studies.
- Electrocardiography–machine learning tools exhibited biases in disease classification related to patient comorbidities and patient demographics.
- Grad-CAM analysis showed high variability between machine learning architectures, revealing no clear patterns.

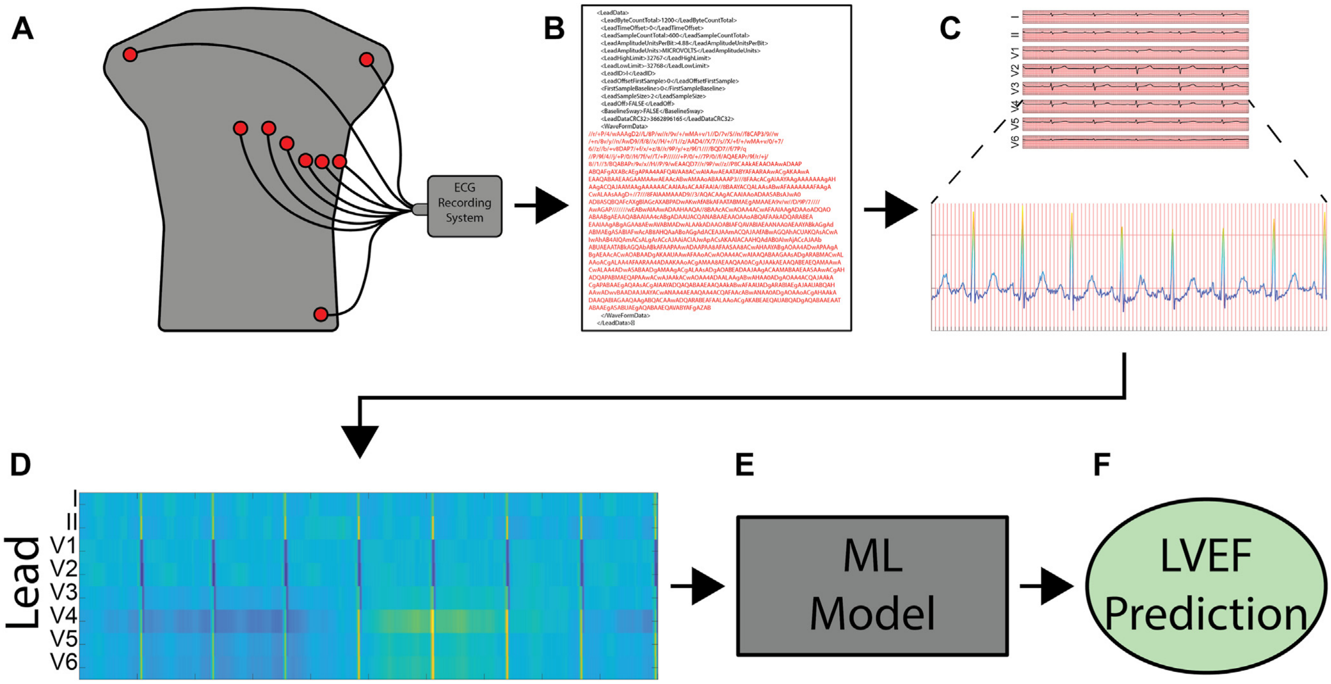
Administration authorization are being implemented in medical systems as diagnostic tests that can run on collected ECGs and provide additional diagnostic information.[4,6,7]

AI-ML, while relatively new in the healthcare space, has been around for several decades. Many robust AI-ML tools have been designed and applied across various problems, including image analysis, text prediction, and "chatbots" such as ChatGPT.[12] The progression of these tools has gone through revolutionary development with a thriving community creating open-source or freely available, predesigned AI-ML architectures that can be easily trained on similar classification problems, "off the shelf."[13] However, these architectures have not been robustly applied to ECG analysis. Many of the AI-ML techniques currently used in ECG analysis use custom implementations, which limits the trust and portability of these tools applied to other ECG datasets from different patient populations.[4,14–17] Custom and proprietary AI-ML architectures also inhibit the development and collaborative improvement of these approaches applied to relevant clinical problems. Applying off-the-shelf AI-ML architectures to ECG-based datasets opens the door for rapid development and identification of previously unknown disease biomarkers.

Despite the excellent opportunity, the ideal open-source AI-ML architecture for ECG-related problems is not known. Furthermore, there has been limited investigation on how and when these AI-ML approaches fail and possible bias or disparities associated with particular network architectures. In this study, we aimed to (1) determine if open-source, off-the-shelf AI-ML architectures could be trained to classify low LVEF from ECGs; (2) assess the accuracy of different AI-ML architectures compared with each other; and (3) identify which, if any, patient characteristics are associated with poor AI-ML performance.

Methods

The research reported in this study adheres to the Helsinki Declaration guidelines for human research. Our overall data acquisition and use pipeline is outlined in Figure 1. Data recorded from a standard clinical 12-lead ECG system is saved as raw ASCII text encoded data. We then convert these encoded data back into signal waveforms for the 8 unique leads of the 12-lead ECG and compile these into ML model inputs.



```
<LeadData>
  <LeadByteCountTotal>1200</LeadByteCountTotal>
  <LeadTimeOffset>0</LeadTimeOffset>
  <LeadSampleCountTotal>600</LeadSampleCountTotal>
  <LeadAmplitudeUnitsPerBit>4.88</LeadAmplitudeUnitsPerBit>
  <LeadAmplitudeUnits>MICROVOLTS</LeadAmplitudeUnits>
  <LeadHighLimit>32767</LeadHighLimit>
  <LeadLowLimit>-32768</LeadLowLimit>
  <LeadID>I</LeadID>
  <LeadOffsetFirstSample>0</LeadOffsetFirstSample>
  <FirstSampleBaseline>0</FirstSampleBaseline>
  <LeadSampleSize>2</LeadSampleSize>
  <LeadOff>FALSE</LeadOff>
  <BaselineSway>FALSE</BaselineSway>
  <LeadDataCRC32>3662896165</LeadDataCRC32>
  <WaveFormData>
//r/+P/4/wAAAgD2//L/8P/w//r/9v/+/wMA+v/1//D/7v/5//n//f8CAP3/9//w
/+n/8v/y//n/AwD9//f/8//x//H/+//1//z/AAD4//X/7//s//X/+f/+/wMA+v/0/+7/
6//z//b/+v8DAP7/+f/x/+z/8//r/9P/y/+z/9f/1////BQD7//f/7P/q
//P/9f/4//j/+P/0//H/7f/v//T/+P//////+P/0/+//7P/0//f/AQAEAPr/9f/r/+j/
8//1//3/BQABAPr/9v/x//H//P/9/wEAAQD7//r/9P/w//z//P8CAAkAEAAOAAwADAAP
ABQAFgAXABcAEgAPAA4AAFQAVAA8ACwAIAAwAEAATABYAFAARAAwACgAKAAwA
EAAQABAAEAAGAAMAAwAEAAcABwAMAAoABAAAAP3///8FAAcACgAIAAYAAgAAAAAAAgAH
AAgACQAJAAMAAgAAAAAACAAIAAsACAAFAAIA//8BAAYACQALAAsABwAFAAAAAAAAFAAgA
CwALAAsAAgD+//7///8FAIAAMAAAD9//3/AQACAAgACAAIAAoADAASABsAJwA0
AD8ASQBQAFcAXgBIAGcAXABPADwAKwAfABkAFAATABMAEgAMAAEA9v/w//D/9P/7////
AwAGAP///////wEABwAIAAwADAAHAAQA//8BAAcACwAOAA4ACwAFAAIAAgADAAoADQAO
ABAABgAEAAQABAAIAA4cABgADAAUACQANABAAEAAOAAoABQAFAAkADQARABEA
EAAIAAgABgAGAA8AEwAVABMADwALAAkADAAOABIAFQAVABIAEAANAA0AEAAYABkAGgAd
ABMAEgASABIAFwAcAB8AHQAaABoAGgAdACEAJAAmACQAJAAfABwAGQAhACUAKQAsACwA
IwAhAB4AIQAmACsALgArACcAJAAiACIAJwApACsAKAAIACAAHQAdAB0AIwAjACcAJAAb
ABUAEAATABkAGQAbABkAFAAPAAwADAAPAA8AFAASAA8ACwAHAAYABgAOAA4ADwAPAAgA
BgAEAAcACwAOABAADgAKAAUAAwAFAAoACwAOAA4ACwAIAAQABAAGAAsADgARABMACwAL
AAoACgALAA4AFAARAA4ADAAKAAoACgAMAA8AEAAQAA0ACgAJAAkAEAAQABEAEQAMAAwA
CwALAA4ADwASABAADgAMAAgACgALAAsADgAOABEADAAJAAgACAAMABAAEAASAAwACgAH
ADQAPABMAEQAPAAwACwAJAAkACwAOAA4ADAALAAgABwAHAA0ADgAOAA4ACQAJAAkA
CgAPABAAEgAQAAsACgAIAAYADQAQABAAEAAQAAkABwAFAAUADgARABIAEgAJAAUABQAH
AAwADwvBAADAAJAAYACwANAA4AEAAQAA4ACQAFAAcABwANAA0ADgAOAAoACgAHAAkA
DAAQABIAGAAQAAgABQACAAwADQARABEAFAALAAoACgAKABEAEQAUABQADgAQABAAEAAT
ABAAEgASABUAEgAQABAAEQAVABYAFgAZAB
  </WaveFormData>
</LeadData>
```

Figure 1 Overall data acquisition and analysis pipeline. Recorded 12-lead electrocardiograms (ECGs) (A) were extracted in their raw data format (B) and converted into digital signals for the 8 unique leads of the 12-lead ECG (C): leads I, II, V1, V2, V3, V4, V5, and V6. D: These signals were then arranged into an input matrix of size leads by time, depicted as an input image in which amplitude of the lead is encoded with color. The ECGs were then passed into the machine learning model (E), which then predicted a classification for the presence or absence of low left ventricular ejection fraction (LVEF) (F).

These are then passed into the various ML models, and the output is used to predict low LEF.

ECG dataset

As part of data associated with routine clinical care, ECGs with and LVEF measurement data were extracted from the University of Utah Electronic Data Warehouse from 2012 to 2021, resulting in 24,868 unique patient-ECG pairs. For this study, patients with an ECG recording within 30 days (average of 4.4 ± 7.3 days) of an LVEF measurement via echocardiography were selected. LVEF was calculated using echocardiography measurements verified by board-certified cardiologists. Low LVEF was defined as below 40%. The ECG recordings, performed on a GE HealthCare Marquette ECG Machine, included leads I, II, and V1 to V6. Leads III, aVF, aVR, and aVL can each be derived from the remaining leads, and thus are not used in such analyses. As is standard, recordings were 10 seconds long and sampled at 500 Hz, resulting in an 8 × 5000 point array for each ECG. Patients were split into a 90% training set (22,382 patients) and 10% testing set (2486 patients). The same training and testing sets was used for all analyses. A summary of the patient characteristics for the training and testing set is presented in Table 1.

Patient characteristics

Based on our previously described methodology, clinical data were derived from the healthcare system's enterprise data warehouse and include all administrative billing encounters with diagnosis codes (inpatient, outpatient, procedural), medication orders, and laboratory results.[18,19] Clinical comorbidities were measured using previously validated algorithms in administrative data analyses of cardiovascular disease and included all healthcare system encounters up to and including the index visit. Comorbidity rates were calculated based on International Classification of Diseases codes as part of clinical billing encounters, as previously described.[18,20] The index visit was defined as the date of echocardiogram acquisition.

Machine learning architectures

Open-source ML architectures from the PyTorch Python-based machine learning package were adapted with ECG inputs. Specifically, we implemented untrained versions of ResNet 18, ResNet 50, AlexNet, DenseNet 121, SqueezeNet 1.0, and VGG 11 (https://pytorch.org/vision/stable/models.html).[13] Each network architecture was developed for use with images and by default required a 3 × m × n input tensor (channels × height × width) and produced

Table 1 Patient demographics in the training and testing sets

Variable	Test (n = 2486)	Train (n = 22,382)	*P* value	Missing
Female	1073 (45.31)	9694 (45.43)	.91	1162
Race or Ethnicity				
White or Caucasian	1948 (82.65)	17,569 (82.59)	.84	1238
Black or African American	54 (2.29)	468 (2.2)	—	—
Asian	44 (1.87)	376 (1.77)	—	—
American Indian and Alaska Native	45 (1.91)	356 (1.67)	—	—
Other Pacific Islander	27 (1.15)	298 (1.4)	—	—
Unknown	37 (1.57)	281 (1.32)	—	—
Other	184 (7.81)	1751 (8.23)	—	—
Choose not to disclose	18 (0.76)	174 (0.82)	—	—
Hypertension	583 (24.66)	5139 (24.14)	.57	1216
Diabetes	627 (26.52)	5709 (26.82)	.76	1216
Obstructive sleep apnea	323 (13.66)	3071 (14.43)	.32	1216
Cancer	384 (16.24)	3775 (17.73)	.07	1216
Chronic kidney disease	389 (16.46)	3331 (15.65)	.31	1216
Liver disease	271 (11.46)	2488 (11.69)	.75	1216
Chronic obstructive pulmonary disease	595 (25.17)	5341 (25.09)	.93	1216
Dementia	62 (2.62)	558 (2.62)	1.00	1216
Depression	659 (27.88)	5863 (27.54)	.73	1216
Peripheral artery disease	305 (12.9)	2584 (12.14)	.28	1216
Cardiac valve disease	394 (16.67)	3627 (17.04)	.65	1216
Coronary artery disease	816 (34.52)	7250 (34.06)	.65	1216
Myocardial Infarction	531 (22.46)	4672 (21.95)	.57	1216
Congestive heart failure	512 (21.66)	4555 (21.4)	.77	1216
Cerebrovascular disease	463 (19.59)	4183 (19.65)	.94	1216
Stroke or transient ischemic attack	364 (15.4)	3239 (15.22)	.81	1216
Atrial fibrillation	318 (12.79)	3031 (13.54)	.30	0
Body mass index, kg/m^2	30.00 ± 7.64	29.85 ± 7.75	.41	4379
LVEF, %	57.70 ± 13.96	57.65 ± 14.05	.93	18,136

Values are n (%) or mean ± SD. Statistical significance indicates a difference between training and testing patients.

LVEF = left ventricular ejection fraction.

a 1000-feature vector output. To minimally adapt these architectures to ECG signals, which consist of only 1 channel, 8 signals, and 5000 time instances (1 × 8 × 5000), we preprocessed the input ECGs by adding a 2-dimensional convolutional layer to the beginning of each network with 3 output channels. To process the output of 1000 features, we then appended a final fully connected layer with a single feature output followed by a ReLU (rectified linear unit) layer as the network output. Because of the architectures of the open-source implementations, in some cases it was necessary to either zero-pad ECGs or restructure the input ECG to prevent a collapse in the lead dimension. The overall network design for adapting the open-source architectures for use with ECG data is depicted in Figure 2. Table 2 details the specific preprocessing steps used for each AI-ML architecture. No augmentations or transformations were performed on the input ECGs other than those listed in Table 2.

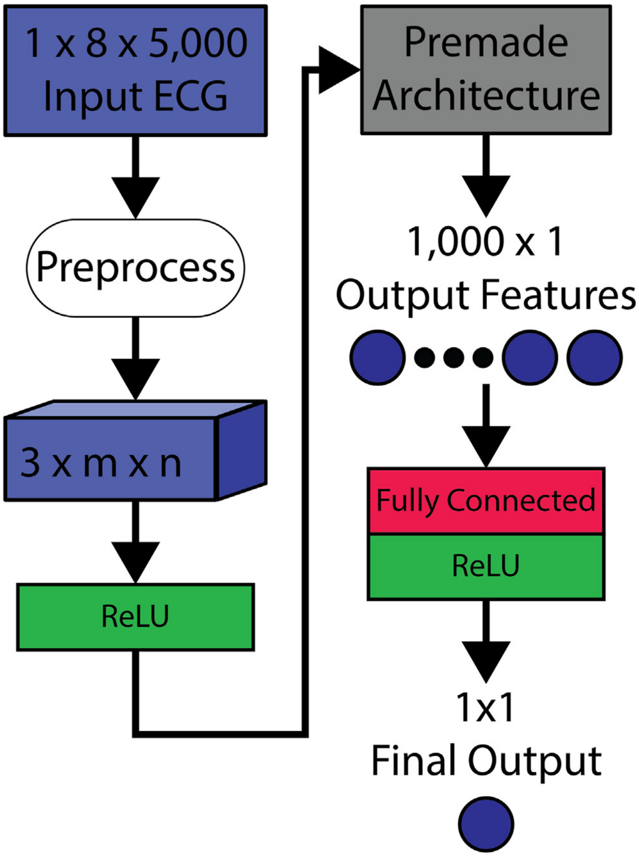


Figure 2 General infrastructure to adapt off-the-shelf machine learning architectures to operate on electrocardiogram (ECG) data. Input ECGs are considered as a 1-channel, 8-lead, 5000-time instant tensor (1 × 8 × 5000 input). The input ECGs are then preprocessed through a combination of reshaping (if needed), padding (if needed), and an initial convolutional layer to produce a 3-channel preprocessed tensor (3 × *m* × *n*). The prepros-sessing steps for each premade architecture are detailed in Table 2. The data are then passed through a rectified linear unit (ReLU) before entering the premade architecture. The output of the premade architecture is a 1 × 1000 output feature vector. These features are passed through a fully connected layer and another ReLU to produce a single output value.

Each network was trained using an Adam optimizer and binary cross-entropy loss between the network output and target LVEF classification.[5] The area under the receiver-operating characteristic curve (AUROC) for the test dataset was monitored throughout training, and the network weights that produced the highest test AUROC were saved to prevent overfitting to the training set. The training was continued for 50 iterations, and the time to complete all iterations was recorded. Weights and biases were initialized randomly for each network. For each AI-ML architecture, 5 separate instances were trained to account for differences caused by random initialization of the network weights and biases.

AI-ML performance analysis

Each trained network (5 instances per network architecture for a total of 30 networks) was evaluated on the testing dataset according to a range of standard metrics, including AUROC, F1 score, sensitivity, and specificity. The output of each architecture is a continuous variable between 0 and 1 that must be thresholded to produce a binary classification of low LVEF. To identify a robust threshold for each network, we selected a threshold that produced the highest F1 score in each network architecture. This threshold was used for the calculation of specificity and sensitivity for each network. Next, using the best-performing instance (highest AUROC) per network architecture, we grouped patients into incorrect prediction (false negative or false positive) or correct prediction (true negative or true positive) groups for each architecture. These groups were then used in subsequent demographic and comorbidity analysis.

Clinical comorbidity analysis

We computed descriptive statistics and summarized the distribution of patient demographic characteristics and medical conditions for numeric and categorical variables. Univariate comparisons across all patient characteristics between correct vs incorrect LVEF classifications were performed for each network architecture. For these comparisons, the best-performing instance of each AI-ML architecture was used to group correctly vs incorrectly classified patients.

Data processing was performed using R (version 3.6.3; R Foundation for Statistical Computing), and RStudio (version 1.2.5033), with appropriate packages. Statistical analysis was performed using R (version 4.1.0) and RStudio (version 1.0.153).[20,21] Analysis of the data collected as part of routine clinical care, and subsequent reporting of anonymized, aggregate data, was approved by the University of Utah Institutional Review Board. The Institutional Review Board waived consent because the study is a retrospective analysis with minimal patient risk. The research reported adheres to the Helsinki Declaration guidelines on human research.

Explainability analysis

To gain additional insight into the features each network was using in prediction of LVEF, we performed Grad-CAM,[22] pooling the activations and gradients for the last

Table 2 Preprocessing of ECG signals for each off-the-shelf network

Architecture	Preprocessing method	Preprocessing output size
ResNet 18	Conv2d f: 7 × 7, p: 3 × 3, och: 3; ReLu	3 × 8 × 5000
ResNet 50	Conv2d f: 7 × 7, p: 3 × 3, och: 3; ReLu	3 × 8 × 5000
AlexNet	Conv2d f: 7 × 7, p: 31 × 0, och: 3; ReLu	3 × 64 × 4994
DenseNet 121	Reshape to 1 × 64 × 625; Conv2d f: 3 × 3, p: 3 × 3, och: 3; ReLu	3 × 68 × 5629
SqueezeNet	Conv2d f: 7 × 7, p: 31 × 0, och: 3; ReLu	3 × 64 × 4994
VGG 11	Reshape to 1 × 64 × 625; Conv2d f: 3 × 3, p: 3 × 3, och: 3; ReLu	3 × 68 × 5629

Reprocessing consisted of a combination of reshaping (if necessary) of the 1×8×5000 (channels by electrodes by time) ECG input, a 2-dimensional convolutional layer (filter size f: $a \times b$, padding p: $c \times d$, and 3 och) followed by a ReLu layer.

ECG = electrocardiography; och = output channels; ReLu = rectified linear unit.

convolutional layer of the best (highest AUROC) network for each architecture. For each network, we selected a total of 8 ECGs for Grad-CAM analysis from the test dataset: 2 that the network identified as false positives, 2 false negatives, 2 true positives, and 2 true negatives. In each case, we selected 2 random examples. We then plotted the normalized (from 0 to 1) Grad-CAM weight signal interpolated to the same dimensions as the input ECG signals.

Computational implementation

All ML architectures were implemented in PyTorch, an open-source AI-ML library.[13] Models were trained and evaluated on a system consisting of 1 NVIDIA TITAN RTX graphic card (24 GB video ram, CUDA 11.4), 2 Intel Xeon Silver 4114 CPUs at 2.20 GHz (20 cores total), 256 GB DDR4 memory, and openSUSE Leap version 15.0.

Results

Network performance

All networks were trained using 22,382 combined ECG-LVEF pairs. Network performance was tested using 2486 ECG-LVEF pairs. Baseline patient characteristics and comorbidities can be found in Table 1.

The performance of each AI-ML architecture is shown in figure 3. ResNet 18 was, on average, the highest-performing architecture with a mean AUROC of 0.917 ± 0.001. VGG 11 was the worst-performing architecture with an AUROC of 0.902 ± 0.004. The F1 score was also used to assess network accuracy. The maximum F1 score was computed for each network by testing a range of network output thresholds. ResNet 18 showed the highest performance with an average maximum F1 score of 0.586 ± 0.010, and VGG 11 had the lowest with 0.520 ± 0.010. The threshold corresponding to the maximum F1 score was used to compute the sensitivity and specificity. At the threshold corresponding to peak F1 score, ResNet 18 had the highest sensitivity (0.638 ± 0.048) and specificity (0.950 ± 0.013), while VGG 11 had the lowest (0.576 ± 0.040 sensitivity, 0.941 ± 0.012 specificity). These results are summarized in Table 3 and the average rates of false negative, false positive, true negative, and true positive for each architecture are summarized in Table 4.

Effects of baseline patient characteristics

For each AI-ML architecture (ResNet 18, ResNet 50, AlexNet, DenseNet 121, SqueezeNet, and VGG 11), the highest-performing instance of the 5 trained instances was selected and thresholded based on the maximum F1 criterion as described previously. Corresponding demographic and comorbidity data were compared for each AI-ML architecture between the correct and incorrect LVEF classifications. Table 5 shows the patient characteristics in the correct vs incorrect for the best performing ResNet 18 architecture. The comparisons for each of the other networks can be found in the supplemental material (Supplemental Tables 1–5). Statistical significance indicates a difference between comorbidity or demographic frequencies in correct vs incorrect groups.

The P values for variable comparisons between correct and incorrect LVEF classification groups are shown as a heatmap in Table 6, with superscript symbols indicating a statistically significantly larger value (percentage for binary variables or mean for scalar variables) in the correct and incorrect prediction groups.

Grad-CAM analysis

Grad-CAM analysis allowed us to explore the relative importance of features in the ECG signals used by each ML network. We found that the distribution of higher normalized Grad-CAM weights was different for each network, demonstrating that each network interpreted the ECG signals in different ways to predict LVEF. Figure 4 shows the Grad-CAM visualization for the ResNet 18 architecture. We see in all positive cases (in which the network determined that LVEF was low) that there is consistently high attention (as indicated by the high normalized Grad-CAM amplitude) across most of the signal, whereas the negative cases showed reduced Grad-CAM amplitude throughout the signals. On the other hand, the SqueezeNet architecture Grad-CAM in Figure 5 showed a more varied signal that often showed higher amplitude in the early QRS complex and lower amplitude in the P-wave and T-Q segments.

Discussion

In summary, we report the first implementation and training of 6 off-the-shelf open-source AI-ML algorithms for use on

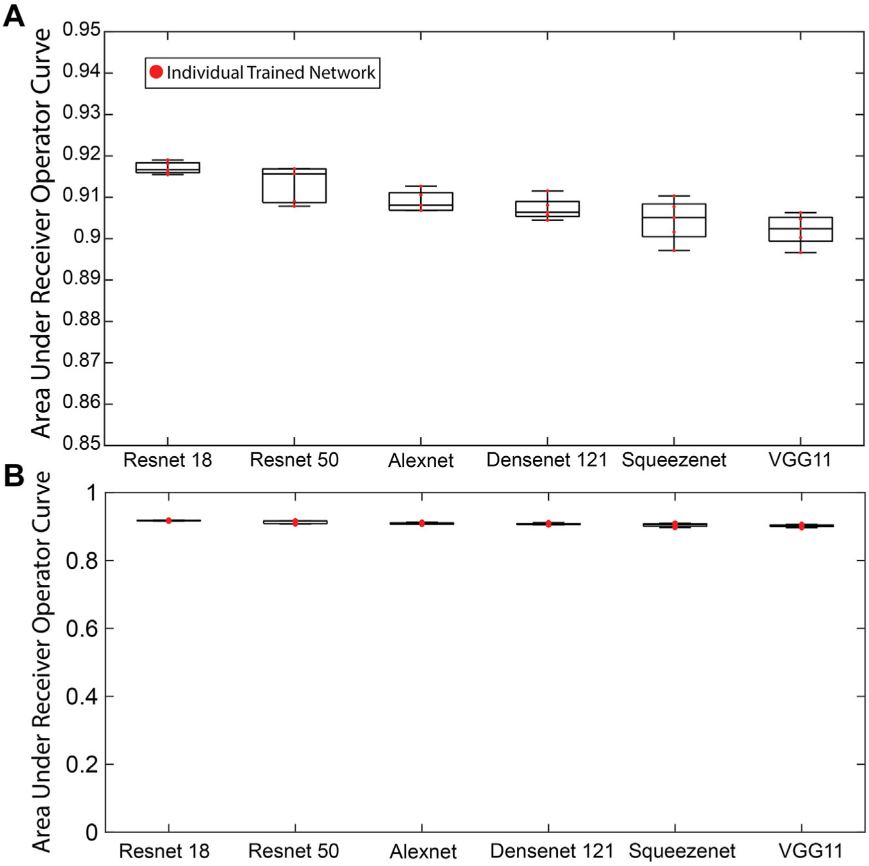


Figure 3 A: Area under the receiver-operating characteristic curve metrics for each network tested. The area under the receiver-operating characteristic curve was calculated for the 5 training tests performed with different testing and training data. B: A zoomed-in Y scale version of panel A.

ECG data to predict low or normal LVEF. There are 3 major findings from these implementations: (1) implementing these architectures was relatively simple, (2) these architectures performed favorably compared with custom-built task-specific algorithms, and (3) despite excellent overall performance, we found that some patient characteristics were more associated with AI-ML LVEF misclassification and may have implications for downstream bias.

We showed that these open-source AI-ML architectures could be rapidly adapted and trained on real-world ECG data to perform a clinically valuable task previously demonstrated primarily by custom-built networks. We implemented AI-ML architectures using open-source packages available in the Python computing language, with minimal overhead and data manipulation. Furthermore, the computational resources necessary to train and test these networks were modest (single graphics card with 24 GB video memory, 20 CPU processor cores, 256 GB RAM). Additionally, minimal data manipulation and preprocessing were required to fit the AI-ML architecture prespecified input and output parameters. The application of AI-ML technology to ECG data may provide important diagnostic value to low healthcare resource environments—the use of widely available, open-source architectures that do not require substantial computing power is an important component for such a deployment.

Table 3 AUC, optimal F1 score, sensitivity, and specificity for each network architecture in the testing set

Architecture	AUC	F1 score	Sensitivity	Specificity
ResNet 18	0.917 ± 0.001	0.586 ± 0.010	0.638 ± 0.048	0.950 ± 0.013
ResNet 50	0.913 ± 0.004	0.560 ± 0.013	0.624 ± 0.024	0.944 ± 0.006
AlexNet	0.909 ± 0.003	0.569 ± 0.021	0.614 ± 0.060	0.949 ± 0.019
DenseNet 121	0.907 ± 0.003	0.535 ± 0.016	0.597 ± 0.038	0.942 ± 0.004
SqueezeNet	0.904 ± 0.005	0.546 ± 0.019	0.602 ± 0.042	0.944 ± 0.010
VGG 11	0.902 ± 0.004	0.520 ± 0.010	0.576 ± 0.040	0.941 ± 0.012

Values are mean ± SD. Metrics are reported over the 5 trained networks per architecture. Sensitivity and specificity are calculated at the threshold corresponding to the peak F1 score.

AUC = area under the curve.

Table 4 Counts for false positive, false negative, true positive, and true negative for each network architecture

Class	False negative	False positive	True negative	True positive
ResNet 18	68 (2.7)	129 (5.1)	2147 (86.4)	142 (5.7)
ResNet 50	81 (3.3)	112 (4.5)	2164 (87.0)	129 (5.2)
AlexNet	90 (3.6)	88 (3.5)	2188 (88.0)	120 (4.8)
DenseNet121	90 (3.6)	131 (5.3)	2145 (86.3)	120 (4.8)
SqueezeNet1 0	93 (3.7)	116 (4.7)	2160 (86.9)	117 (4.7)
vgg11	90 (3.6)	126 (5.1)	2150 (86.5)	120 (4.8)

Values are n (%).

We also showed that open-source AI-ML approaches are consistent with the performance of many of the custom-built approaches on an identical task when comparing routine AI-ML metrics. ResNet 18 performed the best with a mean AUROC of 0.917 ± 0.001 over 5 instances, with VGG 11 performing least accurately but still highly successful with an average AUROC of 0.902 ± 0.004. These values are comparable to other published architectures.[15,23] Furthermore, the maximum F1 score was also relatively high across networks, with the mean F1 score ranging from 0.586 to 0.520 for the ResNet 18 and VGG 11 networks, respectively. Interestingly, the sensitivity ranged from 58% to 63% and specificity from 94% to 95% at the maximum F1 score threshold per network. The clinical implications of the relatively low sensitivity and high specificity indicate that our current implementation is not an ideal screening test but could be used to rule in low LVEF. Adjustments to the selected threshold would allow for tuning of the network toward higher sensitivity or specificity.

The results of this study demonstrate prediction of low LVEF with AUROC comparable to that seen in other studies.[15,23] However, these other studies leveraged much larger datasets (ranging from 44,000 to 97,000 ECG-LVEF pairs) than the one presented in our study (24,868 ECG-LVEF pairs). Our results demonstrate that state of the art LVEF classification by AI-ML can be achieved using substantially smaller training datasets than has been previously published. Furthermore, our results reveal that existing AI-ML network architectures can be successfully adapted to ECG data, with results

Table 5 Demographic and comorbidity comparison between correct predictions (true positive or true negative) vs incorrect predictions (false positive or false negative) for the best-performing ResNet 18 implementation

Variable	Incorrect prediction (n = 197)	Correct prediction (n = 2289)	*P* value	Missing
Ejection fraction, %	47.45 ± 13.59	60.09 ± 11.55	<.001	0
Ejection fraction category				
Low	68 (34.52)	142 (6.2)	<.001	0
Normal	129 (65.48)	2147 (93.8)	—	—
Age, y	64.91 ± 15.43	58.54 ± 17.60	<.001	118
Female	56 (30.94)	1018 (46.55)	<.001	118
Race				
White or Caucasian	153 (84.53)	1794 (82.44)	<.001	129
Black or African American	7 (3.87)	47 (2.16)	—	—
American Indian and Alaska Native	4 (2.21)	41 (1.88)	—	—
Asian	0 (0)	44 (2.02)	—	—
Other Pacific Islander	0 (0)	27 (1.24)	—	—
Other	6 (3.31)	180 (8.27)	—	—
Unknown	10 (5.52)	26 (1.19)	—	—
Choose not to disclose	1 (0.55)	17 (0.78)	—	—
Hypertension	83 (45.86)	505 (23.13)	<.001	122
Diabetes	66 (36.46)	565 (25.88)	.002	122
Obstructive sleep apnea	34 (18.78)	291 (13.33)	.041	122
Cancer	30 (16.57)	353 (16.17)	.89	122
Chronic kidney disease	37 (20.44)	355 (16.26)	.15	122
Liver disease	34 (18.78)	239 (10.95)	.002	122
Chronic obstructive pulmonary disease	58 (32.04)	536 (24.55)	.026	122
Dementia	5 (2.76)	57 (2.61)	.81	122
Depression	39 (21.55)	620 (28.4)	.048	122
Peripheral artery disease	33 (18.23)	272 (12.46)	.026	122
Cardiac valve disease	49 (27.07)	346 (15.85)	<.001	122
Coronary artery disease	117 (64.64)	701 (32.11)	<.001	122
Myocardial infarction	76 (41.99)	456 (20.89)	<.001	122
Congestive heart failure	107 (59.12)	406 (18.6)	<.001	122
Cerebrovascular disease	34 (18.78)	429 (19.65)	.78	122
Stroke or transient ischemic attack	25 (13.81)	338 (15.48)	.55	122
Atrial fibrillation	44 (22.34)	274 (11.97)	<.001	0
Body mass index, kg/m^2	29.90 ± 9.37	30.02 ± 7.47	.87	429

Values are mean ± SD or n (%). Statistical significance indicates a difference between correct and incorrect.

Table 6 *P* values for each variable across each architecture comparing correct predictions (true positive or true negative) vs incorrect predictions (false positive or false negative)

Variable	ResNet 18	ResNet 50	AlexNet	DenseNet121	SqueezeNet1 0	vgg11
Ejection fraction	<.001†	<.001†	<.001†	<.001†	<.001†	<.001†
Ejection fraction category: low	<.001*	<.001*	<.001*	<.001*	<.001*	<.001*
Age	<.001*	<.001*	<.001*	<.001*	<.001*	<.001*
Female	<.001†	.002†	<.001†	<.001†	<.001†	<.001†
Race						
White or Caucasian	<.001*	<.001*	.009*	.003*	.024*	<.001*
Hypertension	<.001*	<.001*	<.001*	<.001*	<.001*	<.001*
Diabetes	.002*	.13	.014*	.023*	.06	.043*
Obstructive sleep apnea	.041*	.63	.06	.45	.49	.015*
Cancer	.89	.50	.81	.36	.09	.93
Chronic kidney disease	.15	.003*	.06	<.001*	<.001*	.019*
Liver disease	.002*	.15	.054	.012*	.047*	.13
Chronic obstructive pulmonary disease	.026*	.026*	.042*	.030*	.006*	<.001*
Dementia	.81	.80	1.00	.73	.68	.18
Depression	.048†	.06	.053	.43	.08	.47
Peripheral artery disease	.026*	.003*	.07	.025*	.13	.19
Cardiac valve disease	<.001*	<.001*	<.001*	<.001*	<.001*	<.001*
Coronary artery disease	<.001*	<.001*	<.001*	<.001*	<.001*	<.001*
Myocardial infarction	<.001*	<.001*	<.001*	<.001*	<.001*	<.001*
Congestive heart failure	<.001*	<.001*	<.001*	<.001*	<.001*	<.001*
Cerebrovascular disease	.78	.56	.73	.73	.60	.10
Stroke or transient ischemic attack	.55	.62	.58	.95	.40	.14
Atrial fibrillation	<.001*	.001*	<.001*	<.001*	<.001*	<.001*
Body mass index	.87	.53	.54	.67	.55	.96

P values were significant at <.05.

*Either higher percent or higher mean in the incorrect predictions.

†Higher percent or higher mean in the correct predictions.

that rival custom made architectures, even when trained using smaller datasets. We anticipate that future studies will leverage the designs of the off-the-shelf network architectures explored in this study as well as the growing availability of larger and larger training datasets to further improve and refine ECG AI-ML tools.

The metrics of AUROC, F1 score, sensitivity, and specificity are commonly used to evaluate AI-ML algorithms, particularly in nonclinical settings. However, in our results there is little absolute variability in these measures among the tested architectures: <0.03 in AUROC, <0.07 in F1 score, <5% in sensitivity, and <1% in specificity. Therefore, selecting an ideal clinical AI-ML approach based on conventional metrics alone may not be helpful—at least when they are each so close. Other criteria, such as resource utilization and portability may become important. Furthermore, these metrics convey minimal clinically relevant information. To address this, we also sought to understand what features of the ECGs were being leveraged by each network architecture to make their classifications. By performing a Grad-CAM analysis, in which we visualize the gradient and activation weights in the trained networks given a particular sample, we were able not only to observe where in the ECG signals each network focused when making their decisions, but also to compare these distributions across network architectures. As shown in Figures 4 and 5, as well as Supplemental Figures 1 to 4, the attention maps for each network architecture vary. A common theme we observed in several architectures was that there was increased attention at or preceding the QRS complex, and lower attention in the P-wave segments or the T-R interval of the ECG signals; however, these results varied substantially. This Grad-CAM analysis represents a first step in attempting to understand both what features the ECG-ML tools use for diagnosis as well as a possible avenue to explore when and why these networks fail on particular signals.

Despite the high level of detail of such Grad-CAM analyses, such measures do not provide information about how the AI-ML approaches perform with different cohorts of patients and if particular patient characteristics are associated with worse AI-ML performance—potential contributors to bias and worsening disparities of care. To provide a framework to understand these possibilities, we investigated whether baseline characteristics differed in patients with correct vs incorrect AI-ML LVEF classification.

We found several baseline comorbidities that were associated with were statistically different in patients with correct vs incorrect AI-ML LVEF classification. These include hypertension, chronic obstructive pulmonary disease (COPD), coronary artery disease (CAD), myocardial infarction (MI), valve disease, congestive heart failure (CHF), and atrial fibrillation (AF). Physiologic reasons for this poor performance can be postulated for patients with an intrinsic cardiac pathology such as hypertension, CHF, MI, CAD, valvular

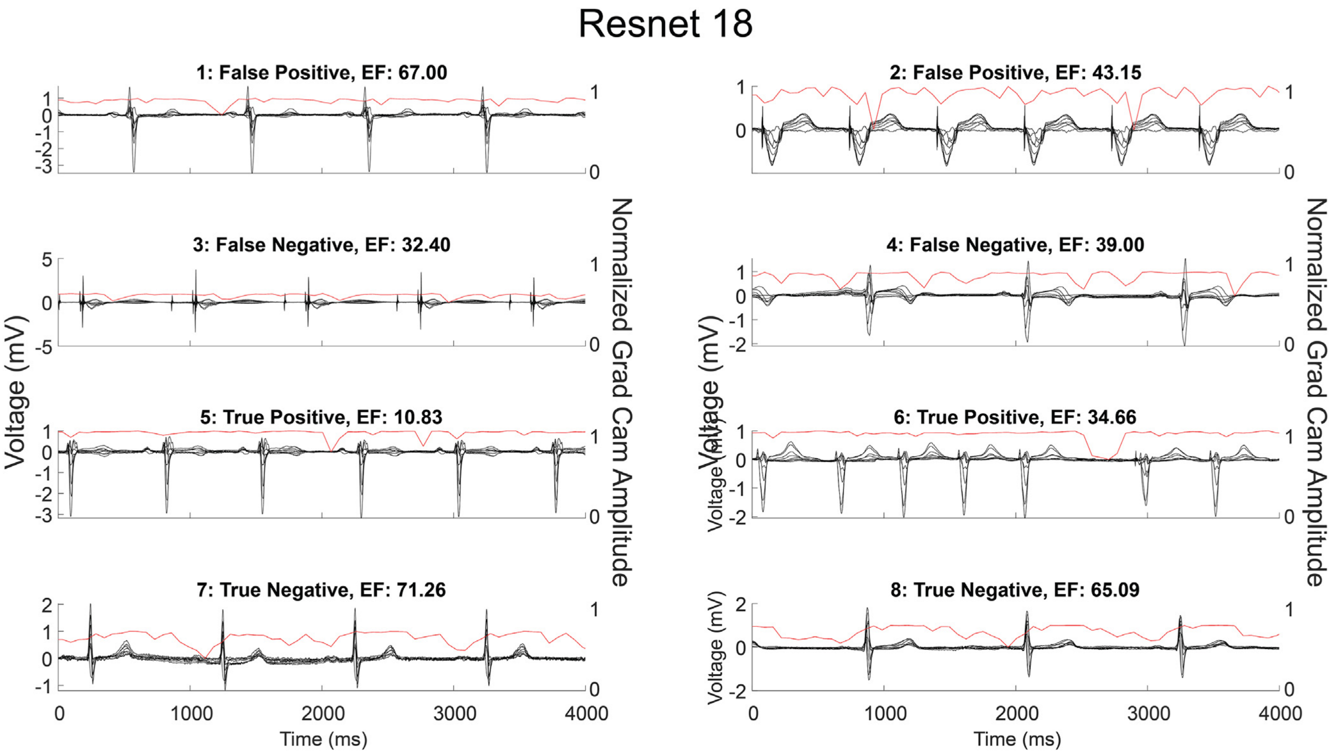


Figure 4 Grad-CAM analysis for the best performing ResNet 18 network. Each plot shows the first 4 seconds of electrocardiogram (ECG) signal for all 8 leads of the ECG overlaid in black. ECG signals use the left-side y-axis scale of voltage in mV. Each plot also shows the Grad-CAM amplitudes in red interpolated to the same dimensions as the input ECG. The Grad-CAM amplitudes use the right-side y-axis ranging from 0 to 1. A higher-amplitude Grad-CAM indicates higher importance of this region of the input signal in the final classification decision. Plots 1 and 2 show 2 false positive cases in which the network identified these ECGs as belonging to a patient with low left ventricular ejection fraction (EF), despite their high EF. Plots 3 and 4 show 2 examples for false negative cases in which this network failed to identify the presence of low left ventricular EF. Plots 5 and 6 show 2 examples in which this network correctly identified low left ventricular EF. Plots 7 and 8 show 2 examples in which this network correctly identified a lack of low left ventricular EF.

disease, or AF because each can significantly affect the cardiac electrical conduction system and alter the 12-lead ECG. However, COPD is not an intrinsic cardiac pathology but could be related to changes in cardiac electrical signals. Patients with COPD often have higher lung volumes, which can be an excellent electrical insulator. These hypotheses are preliminary and not explicitly confirmed by the data presented, but rather simply demonstrate the possible links between changes in ECG signals and baseline patient comorbidities.

Importantly, we also found constitutional patient characteristics, such as sex and race, in which performance differed. Specifically, patients who are older and White are more likely to have a false positive or false negative AI-ML classified low LVEF. However, this result must be interpreted in context of the small number of non-White patients in the dataset (17.41%). We feel that the representation of non-White patients in this dataset is insufficient to draw definitive conclusions about the ML performance in these groups. We hypothesize that additional confounding variables not available in the present study (such as socioeconomic status, access to medical care, and other systemic oppression and prejudice) may play a role. Further studies are needed to carefully interrogate these results accounting for other social determinants of health, which address the challenges of such imbalances in the data. While we realize that the results presented in this study are not definitive performance metrics, we believe they demonstrate the significant potential for AI-ML approaches to impact disparities of care both positively and/or negatively. AI-ML approaches used in other fields have begun to grapple with realities of bias. We find it prudent to highlight that AI-ML applied to ECG-related problems could also experience similar inherent biases. Further work is needed to recognize, describe, and correct for the disparities that develop from these approaches.

Additionally, our results will contribute to explanatory AI-ML. Traditional, human-based ECG interpretation has been refined over decades, to describe patterns associated with disease commonly via pathophysiologic links. In contrast, modern ML-ECG algorithms remain more black-box technologies that generate predictive output with little explanation as to why the algorithm has linked a specific ECG to a target. And while some algorithms intuitively link the ECG to a related cardiovascular outcome (eg, future arrhythmia),[16,24] others have linked the ECG waveform to seemingly unrelated conditions such as liver disease—a less clear pathophysiologic link.[25] In the present study, we

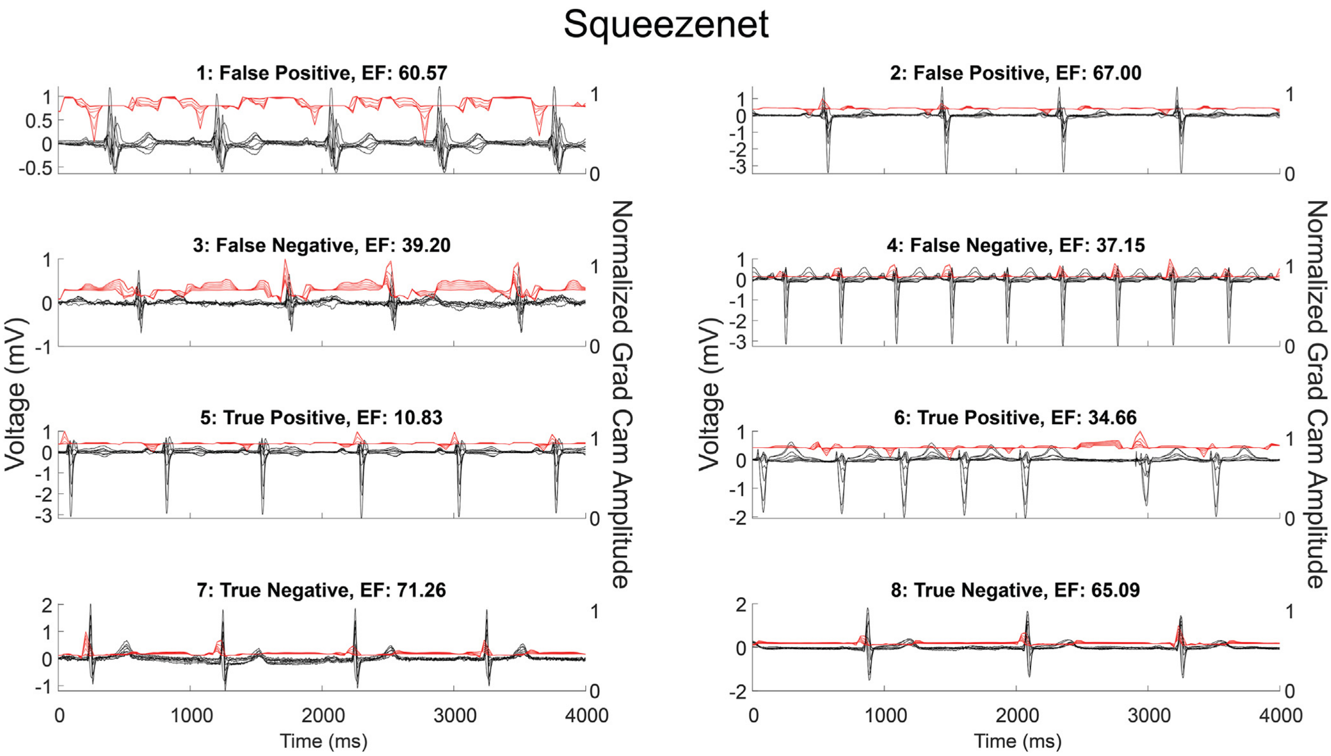


Figure 5 Grad-CAM analysis for the best performing SqueezeNet network. Each plot shows the first 4 seconds of electrocardiogram (ECG) signal for all 8 leads of the ECG overlaid in black. ECG signals use the left-side y-axis scale of voltage in mV. Each plot also shows the Grad-CAM amplitudes in red interpolated to the same dimensions as the input ECG. The Grad-CAM amplitudes use the right-side y-axis ranging from 0 to 1. A higher amplitude Grad-CAM indicates higher importance of this region of the input signal in the final classification decision. Plots 1 and 2 show 2 false positive cases in which the network identified these ECGs as belonging to a patient with low left ventricular ejection fraction (EF), despite their high EF. Plots 3 and 4 show 2 examples for false negative cases in which this network failed to identify the presence of low left ventricular EF. Plots 5 and 6 show 2 examples in which this network correctly identified low left ventricular EF. Plots 7 and 8 show 2 examples in which this network correctly identified a lack of low left ventricular EF.

also begin the important work of assessing how each architecture interprets the ECGs to make decisions using Grad-CAM. However, interpretation of such Grad-CAM results is difficult, especially when seeking clinically meaningful features. Grad-CAM analysis alone does not elucidate how the AI-ML algorithms use the features of interest. Future studies could expand on these explanatory machine learning techniques to further explore what areas of the ECG are identified by each architecture and if changes in those regions of the ECG are related to an underlying diagnosis.[23,26,27]

One crucial factor that could be driving these results is the frequency of patients with specific underlying characteristics or comorbidities in the training data. However, our patient population has average or higher rates of comorbidities than the general U.S. population.[28–30] Our data also have higher rates of cardiac diagnoses, including CAD, MI, AF, and others. Furthermore, our training vs testing dataset had minimal differences in patient baseline characteristics and comorbidities.

Limitations

There were several limitations to this study. First, these ML architectures we implemented were designed for use with images, and thus in some cases we were required to reshape our input ECGs to allow for application of these architectures (DenseNet 121 and VGG 11). Such reshaping may be deleterious for the spatially coherent information present in an ECG signal, and thus may negatively impact network performance. Furthermore, such open-source AI-ML architectures are not designed to leverage the unique features of ECG data such as its temporal coherence, as is seen in other ECG-specific approaches.[1,2] Our dataset was also limited to a single center with a relatively socially homogenous population. Finally, our dataset was biased to have more individuals with a normal LVEF.

Conclusion

We found that several off-the-shelf, open-source AI-ML architectures could be used to predict low LVEF from ECGs. Specifically, we found that these approaches were easy to implement and performed comparably to previously reported custom-built networks. Furthermore, we found baseline patient characteristics differed substantially between patients with correct vs incorrect AI-ML LVEF classification. These findings should be considered in the pursuit of efficient and equitable deployment of AI-ML technologies moving forward.

Funding Sources: Support for this research came from the Center for Integrative Biomedical Computing (www.sci.utah.edu/cibc); National Institutes of Health/National Institute of General Medical Sciences grants P41 GM103545 and R24 GM136986; National Institutes of Health/National Institute of Biomedical Imaging and Bioengineering grant U24EB029012; the National Heart, Lung, and Blood Institute of the National Institutes of Health under Award Number K23HL143156 (to Benjamin A. Steinberg) award number R21 HL172288-01 (to Benjamin A. Steinberg and Tolga Tasdizen) and Award Number F30HL149327 (to Brian Zenger); the University of Utah Study Design and Biostatistics Center, with funding in part from the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through grant 8UL1TR000105 (formerly UL1RR025764); and the Nora Eccles Harrison Foundation for Cardiovascular Research.

Disclosures: Benjamin A. Steinberg has received salary support from the National Institutes of Health/National Heart, Lung, and Blood Institute (K23HL143156, R56HL168264, R21HL172288) and American Heart Association/Patient-Centered Outcomes Research Institute (18SFRN34110489); received research support from Abbott, Cardiva, Sanofi, and AltaThera; served as a consultant for Sanofi, InCarda, Milestone, Pfizer, and AltaThera. T. Jared Bunch has received research support from Altathera, Boehinger Ingelheim, Boston Scientific, and the Heartline Study Steering Committee for Janssen Scientific Affairs/Johnson & Johnson. Ravi Ranjan has received support for research from Biosense Webster and National Institutes of Health grant R01 HL162353; and served as a consultant for Abbott. All other authors did not report any relevant disclosures.

Authorship: All authors attest they meet the current ICMJE criteria for authorship.

Patient Consent: The Institutional Review Board waived consent because the study is a retrospective analysis with minimal patient risk.

Ethics Statement: All studies described were approved by the University of Utah Institutional Review Board. The research reported adheres to the Helsinki Declaration guidelines on human research.

References

1. Xue J, Yu L. Applications of machine learning in ambulatory ECG. Hearts 2021; 2:472–494.
2. Natarajan A, Chang Y, Mariani S, et al. A wide and deep transformer neural network for 12-lead ECG classification. In: 2020 Computing in Cardiology. Piscataway, NJ: IEEE; 2020. p. 1–4.
3. Bergquist JA, Rupp L, Zenger B, Brundage J, Busatto A, MacLeod R. Body surface potential mapping: contemporary applications and future perspectives. Hearts 2021;2:514–542.
4. Yao X, McCoy RG, Friedman PA, et al. ECG AI-guided screening for low ejection fraction (EAGLE): rationale and design of a pragmatic cluster randomized trial. Am Heart J 2020;219:31–36.
5. Jentzer JC, Kashou AH, Attia ZI, et al. Left ventricular systolic dysfunction identification using artificial intelligence-augmented electrocardiogram in cardiac intensive care unit patients. Int J Cardiol 2021;326:114–123.
6. Aro AL, Reinier K, Rusinaru C, et al. Electrical risk score beyond the left ventricular ejection fraction: prediction of sudden cardiac death in the Oregon Sudden Unexpected Death Study and the Atherosclerosis Risk in Communities Study. Eur Heart J 2017;38:3017–3025.
7. Pour-Ghaz I, Heckle M, Ifedili I, et al. Beyond ejection fraction: novel clinical approaches towards sudden cardiac death risk stratification in patients with dilated cardiomyopathy. Curr Cardiol Rev 2022;18:e040821195265.
8. Al-Khatib SM, LaPointe NN, Kramer JM, Califf RM. What clinicians should know about the QT interval. JAMA 2003;289:2120–2127.
9. Kataoka H, Madias JE. Changes in the amplitude of electrocardiogram QRS complexes during follow-up of heart failure patients. J Electrocardiol 2011; 44:394.e1–394.e9.
10. Magnani JW, Wang N, Nelson KP, et al. Electrocardiographic PR interval and adverse outcomes in older adults. Circ Arrhythm Electrophysiol 2013;6:84–90.
11. Dhingra R, Pencina MJ, Wang TJ, et al. Electrocardiographic QRS duration and the risk of congestive heart failure. Hypertension 2006;47:861–867.
12. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol 2022;23:40–55.
13. Paszke A, Gross S, Massa F, et al. Pytorch: an imperative style, high-performance deep learning library. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates; 2019. p. 8026–8037.
14. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. Nat Rev Cardiol 2021;18:465–478.
15. Harmon DM, Carter RE, Cohen-Shelly M, et al. Real-world performance, long-term efficacy, and absence of bias in the artificial intelligence enhanced electrocardiogram to detect left ventricular systolic dysfunction. Eur Heart J Digit Health 2022;3:238–244.
16. Christopoulos G, Attia ZI, Van Houten HK, et al. Artificial intelligence-electrocardiography to detect atrial fibrillation: trend of probability before and after the first episode. Eur Heart J Digit Health 2022;3:228–235.
17. Noseworthy PA, Attia ZI, Behnken EM, et al. Artificial intelligence-guided screening for atrial fibrillation using electrocardiogram during sinus rhythm: a prospective non-randomised interventional trial. Lancet 2022;400:1206–1212.
18. Steinberg BA, Turner J, Lyons A, et al. Systematic collection of patient-reported outcomes in atrial fibrillation: feasibility and initial results of the Utah Meval AF Programme. Europace 2020;22:368–374.
19. Zenger B, Zhang M, Lyons A, et al. Patient-reported outcomes and subsequent management in atrial fibrillation clinical practice: results from the Utah Meval AF Programme. J Cardiovasc Electrophysiol 2020;31:3187–3195.
20. Wasey J. ICD: comorbidity calculations and tools for ICD-9 and ICD-10 codes (R package v3.3). 2018. Available at: https://cran.r-project.org/web/packages/icd/index.html. Accessed January 1, 2020.
21. Yoshida K, Bohn J. Create table 1 to describe baseline characteristics (R package). 2018. Available at: https://github.com/kaz-yos/tableone. Accessed January 1, 2020.
22. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE; 2017. p. 618–626.
23. Noseworthy PA, Attia ZI, Brewer LC, et al. Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ECG analysis. Circ Arrhythm Electrophysiol 2020;13:e007988.
24. Siontis KC, Yao X, Pirruccello JP, Philippakis AA, Noseworthy PA. How will machine learning inform the clinical care of atrial fibrillation? Circ Res 2020; 127:155–169.
25. Ahn JC, Attia ZI, Rattan P, et al. Development of the AI-Cirrhosis-ECG score: an electrocardiogram-based deep learning model in cirrhosis. Am J Gastroenterol 2022;117:424–432.
26. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018;178:1544–1547.
27. Food and Drug Administration. Executive Summary for the Patient Engagement Advisory Committee Meeting: Artificial Intelligence (AI) and Machine Learning (ML) in Medical Devices. Available at: https://www.fda.gov/media/151482/download. Accessed January 1, 2022.
28. Sharma A, Zhao X, Hammill BG, et al. Trends in noncardiovascular comorbidities among patients hospitalized for heart failure. Circ Heart Fail 2018; 11:e004646.
29. Charlson M, Szatrowski TP, Peterson J, Gold J. Validation of a combined comorbidity index. J Clin Epidemiol 1994;47:1245–1251.
30. Charlson ME, Pompei P, Ales KL, MacKenzie C. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chron Dis 1987;40:373–383.