A comprehensive artificial intelligence–enabled electrocardiogram interpretation program

Anthony H. Kashou, MD,*[1] Wei-Yin Ko, MS,†[1] Zachi I. Attia, PhD,† Michal S. Cohen, MS,† Paul A. Friedman, MD, FHRS,† Peter A. Noseworthy, MD, FHRS†

*From the *Department of Medicine, Mayo Clinic, Rochester, Minnesota, and †Department of Cardiovascular Medicine, Mayo Clinic, Rochester, Minnesota.*

BACKGROUND Automated computerized electrocardiogram (ECG) interpretation algorithms are designed to enhance physician ECG interpretation, minimize medical error, and expedite clinical workflow. However, the performance of current computer algorithms is notoriously inconsistent. We aimed to develop and validate an artificial intelligence–enabled ECG (AI-ECG) algorithm capable of comprehensive 12-lead ECG interpretation with accuracy comparable to practicing cardiologists.

METHODS We developed an AI-ECG algorithm using a convolutional neural network as a multilabel classifier capable of assessing 66 discrete, structured diagnostic ECG codes using the cardiologist's final annotation as the gold-standard interpretation. We included 2,499,522 ECGs from 720,978 patients ≥18 years of age with a standard 12-lead ECG obtained at the Mayo Clinic ECG laboratory between 1993 and 2017. The total sample was randomly divided into training (n = 1,749,654), validation (n = 249,951), and testing (n = 499,917) datasets with a similar distribution of codes. We compared the AI-ECG algorithm's performance to the cardiologist's interpretation in the testing dataset using receiver operating characteristic (ROC) and precision recall (PR) curves.

RESULTS The model performed well for various rhythm, conduction, ischemia, waveform morphology, and secondary diagnoses codes with an area under the ROC curve of ≥0.98 for 62 of the 66 codes. PR metrics were used to assess model performance accounting for category imbalance and demonstrated a sensitivity ≥95% for all codes.

CONCLUSIONS An AI-ECG algorithm demonstrates high diagnostic performance in comparison to reference cardiologist interpretation of a standard 12-lead ECG. The use of AI-ECG reading tools may permit scalability as ECG acquisition becomes more ubiquitous.

KEYWORDS Artificial intelligence; Convolutional neural network; Deep learning; ECG; Electrocardiogram; Electrocardiography; Machine learning

(Cardiovascular Digital Health Journal 2020;1:62–70) © 2020 The Author(s). Published by Elsevier Inc. on behalf of Heart Rhythm Society. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction

The electrocardiogram (ECG) is one of the most widely available and routinely performed diagnostic tests in modern medicine. The first attempts for a computer program to automatically extract clinically relevant measurements from an ECG without human intervention occurred more than 50 years ago.[1,2] From its inception, the overarching aims of automated ECG analysis were to enhance physician interpretation, reduce health care costs, minimize medical error, optimize clinical workflow, and facilitate clinical decision making.[3] It has since become a mainstay in modern medical practice, allowing for rapid ECG interpretation.

Most existing automated computerized algorithms were developed using expert physicians as the gold standard. However, the performance of current computer algorithms is inconsistent,[3–6] particularly for important conditions such as myocardial infarction.[7,8] Guglin and Thatai[5] attempted to examine the faulty nature of computer ECG reading and reported that the most frequent errors were related to arrhythmias, conduction disorders, and electronic pacemakers. They noted that interpretation challenges occurred with non–sinus rhythms, in which there were difficulties in accurately detecting P waves with low amplitude, varying morphology, or those masked by artifact or other aspects of the cardiac complex. Similarly, atrial fibrillation was often difficult to detect owing to artifact or sinus rhythm with associated premature beats. Low-voltage pacemaker spikes were also frequently missed, leading to multiple interpretation errors (eg, myocardial infarction, left bundle branch block, left ventricular hypertrophy, and intraventricular conduction delay). Moreover, currently implemented computer algorithms were developed on ECG databases of selected populations that very likely do not accurately reflect all population subsets and do not include all possible clinical diagnoses, which undeniably play roles in their inaccuracies. Apart from these issues, automated ECG interpretations significantly influence

[1] The first 2 authors contributed equally to this work. **Address reprint requests and correspondence:** Dr Peter A. Noseworthy, Department of Cardiovascular Diseases, Electrophysiology, Mayo Clinic, 200 First St SW, Rochester, MN 55905. E-mail address: noseworthy.peter@mayo.edu.

https://doi.org/10.1016/j.cvdhj.2020.08.005

KEY FINDINGS

An artificial intelligence–enabled electrocardiogram (AI-ECG) algorithm demonstrates high diagnostic performance in comparison to reference cardiologist interpretation of a standard 12-lead ECG.

the over-reading clinician's final interpretation[9–12] and errors are often not corrected.[11,12] Therefore, interpretation errors are simply propagated forward for clinical implementation. It is clear that a more robust and accurate automated ECG interpretation system is needed.

In recent years, significant advances have been made in the field of machine learning.[13] Convolutional neural networks (CNNs) represent a specific class of models that have proven effective in image and speech recognition.[14,15] Though these computational models require some guidance, they consist of multiple processing layers that can be continuously trained with input data to fine tune their final output to perform a specific task. The enormous amount of available medical data has allowed for researchers to harness the capabilities of machine learning to improve patient care.[16,17]

With emergence of ECG-enabled smart phone and clothing technology, the ability to capture ECG signals is becoming ubiquitous. This sheer volume of often time-critical ECG signals requiring interpretation exceeds the capacity of current health care systems. Fortunately, ECG data are amenable to analysis with a CNN. The use of raw digital ECG data allows investigators to train models capable of ECG interpretation. These ECG signals can provide critical diagnostic information and clues for more urgent medical therapy. A deep learning approach has been shown to identify a host of distinct arrhythmias from single-lead ECGs,[18] detect ventricular dysfunction,[19] and predict the likelihood of developing of atrial fibrillation in sinus rhythm.[20]

Given the advances in deep learning in electrocardiography, we hypothesized that a CNN would permit comprehensive ECG analysis with similar accuracy to that of board-certified, practicing cardiologists. To test this hypothesis, we used a large unselected dataset of standard 12-lead ECGs paired with labels applied by board-certified practicing cardiologists to develop, validate, and test a CNN capable of detecting 66 diagnostic ECG codes on a total of 2,499,522 standard ECGs from 720,978 patients.

Methods

Disclosure statement

The authors are unable to make the data publicly available as it originates from Mayo Clinic's ECG database, which has patient identifying information. However, the authors have made the methods available to the reader. The research reported in this paper adhered to Helsinki Declaration as revised in 2013.

Funding

This project was conceived, executed, and funded by Mayo Clinic without industry support. Data from the Mayo Clinic digital vault was only used from individuals that provided consent to use of their anonymized records for research. The Mayo Clinic Foundational Institutional Review Board approved this study.

Study cohort and ECGs

Our study population included 720,978 adult patients aged 18 years or older with a standard 12-lead ECG performed at the Mayo Clinic ECG laboratory between 1993 and 2017. From the total patient cohort, a total of 2,499,522 fully de-identified ECGs were obtained. No patients or ECGs were excluded from the study.

ECGs were performed with a Marquette ECG machine (GE Healthcare, Chicago, IL) and then stored with MUSE data management for retrieval. All ECGs included a final expert annotation by board-certified practicing cardiologists. The annotations included primary- and secondary-class diagnoses for a total of 66 discrete diagnostic ECG codes attributed by expert readers, who served as the comparative gold standard. The diagnostic ECG codes included primary and secondary rhythms, axis deviation, chamber enlargement, atrioventricular (AV) and intraventricular conduction delays, myocardial ischemia, waveform abnormalities, clinical disorders, and pacemaker activity. The total sample of ECGs was randomly divided into a training dataset (n = 1,749,654), validation dataset (n = 249,951), and testing dataset (n = 499,917)—all with a similar distribution of ECG codes (Supplemental Table 1). Supplemental Figure 1 demonstrates the age and gender distribution of the total population and its subgroups (ie, testing, validation, and testing populations).

Table 1 Convolutional neural network design

ResNet Bottleneck Block, input channel = 1, output channel=16, stride=2
ResNet Bottleneck Block, input channel = 16, output channel = 32, stride = 2
ResNet Bottleneck Block, input channel = 32, output channel = 32, stride = 2
ResNet Bottleneck Block, input channel = 32, output channel = 64, stride = 2
ResNet Bottleneck Block, input channel = 64, output channel = 64, stride = 2
ResNet Bottleneck Block, input channel = 64, output channel = 128, stride = 2
ResNet Bottleneck Block, input channel = 128, output channel = 128, stride = 2
ResNet Bottleneck Block, input channel = 128, output channel = 256, stride = 2
ResNet Bottleneck Block, input channel = 256, output channel = 512, stride = 2
ResNet Bottleneck Block, input channel = 512, output channel = 1024, stride = 2
ResNet Bottleneck Block, input channel = 1024, output channel = 2048, stride=2
Linear Layer, input channel = 6144, output channel = 66

Algorithm development

We developed an artificial intelligence–enabled ECG (AI-ECG) algorithm using a CNN as a multilabel classifier capable of detecting 66 discrete, structured ECG codes from the standard 12-lead ECG. The CNN takes the raw ECG data tracings as input and outputs discrete diagnostic ECG labels. The task was considered a multilabel problem as each ECG can have multiple codes. As such, the model created a binary evaluation of whether or not the code was present on the ECG for each of the 66 codes in parallel. The CNN used for predicting the 66 diagnostic ECG code groups was a custom-built model using the PyTorch deep learning library that uses bottleneck ResNet blocks (Table 1). The network architecture contains a total of 11 bottleneck ResNet blocks, which consists of 33 convolutional layers. The number of dimensions was 3. The convolutional kernels are either 3 × 3 in size or 1 × 1 (ie, the bottlenecks) with a stride of 2 for each block. The final block has a channel output of 2048. The average length of the ECG record was 10 seconds at 500 Hz, with the standard 12-lead ECG representing a 5000 × 12 × 1 matrix. We found that the deep convolution was able to counterbalance the fact that the different ECG codes use different relevant features while leveraging the similar features used in similar codes. Therefore, the convolutional output was simply flattened, and a linear layer was applied get the output results for 66 classes.

The network was trained on a training set of ECGs (n = 1,749,654). The model is initialized using Glorot initialization and optimized with the Adam optimizer.[21] The learning rate started at 0.0001 and was manually decreased over time. For each trial, we trained the model for as long as it took for clear overfitting of the training set to occur. We monitored the validation loss and stopped the training process if validated loss stopped decreasing for 10 epochs. Thereafter, we selected the best-performing model checkpoint on the validation set and applied it to the testing set to get the final result. The optimal network was selected by convoluting between the leads rather than treating all leads independently. This confirmed that there was information between leads that can only be captured when examined together. This suggested that like cardiologists, the model is evaluating the ECGs based on the relationships of the raw voltage values for code classification.

Algorithm validation, testing, and statistical analysis

Receiver operator characteristic (ROC) and complementary precision recall (PR) curves were created using a validation set of ECGs (n = 249,951) with area under the curve (AUC) and PR as primary assessments of the network's strength for each code, respectively. This allowed for evaluation of the model's ability to discriminate each code. The ROC curve summarized the trade-off between the true-positive rate and false-positive rate of the model, while the PR curve summarized the relationship between positive predictive value (precision) and sensitivity (recall). ROC curves allowed for observations balanced between each class, while PR curves enabled evaluation of class-imbalanced datasets.

Sensitivity and specificity for every code were calculated at binary decision thresholds to allow for equal weighting with both ROC and PR curves. The probability thresholds were applied post-prediction as a straightforward gauge of binary decision performance on the testing ECG dataset (n = 499,917) given equal weighting of true-positive rate/false-positive rate for ROC curve and precision/recall for the PR curve. We then assessed the model's diagnostic performance (ie, ability to interpret a 12-lead ECG compared to that of a cardiologist) by calculating the AUC, average precision score (AP), and the sensitivity and specificity for the ROC and PR curves. We also computed the F_1 score, which is the harmonic mean of PPV and sensitivity based on the selected threshold, to compare the diagnostic performance of the model to the expert annotation. F_1 scores closer to 1 maximize both PPV and sensitivity instead of favoring 1 over the other. The F_1 score is less sensitive than AUC when class imbalance is present. None of the ECGs used in testing set were used to derive or validate the CNN.

Results

The model's diagnostic performance, including AUC and AP along with their respective sensitivity and specificity, for each individual code is recorded in Tables 2–6. The F_1 score for each code is also reported. Supplemental Figures 2 and 3 demonstrate ROC and PR curves for each ECG code.

In general, in terms of sensitivity, an AUC threshold performed better than a PR threshold for more balanced classes (ie, more prevalent codes), while the PR threshold outperformed the AUC threshold for heavily imbalanced classes (ie, less prevalent codes). For instance, sinus bradycardia represented the third most common ECG code (prevalence 14.5%, n = 72,286; AUC and AP > 0.99) in the testing set with a specificity measure for AUC of 0.97 but only 0.55 with PR. In comparison, a relatively uncommon code like dextrocardia (prevalence 0.03%, n = 129; AUC > 0.99, AP > 0.60) in the testing set had an AUC and PR specificity of 42% and 91%, respectively. Using the thresholds, which were selected not by clinical needs but simply for benchmarking purposes between the 2 curves, these findings suggest that AUC curve may be a better measure than the PR curve for more common diagnoses, and vice versa for less common diagnoses.

Primary and secondary rhythm interpretation

Table 2 summarizes the model's diagnostic performance for determining 23 different primary and secondary rhythms. For detection of a normal ECG (prevalence 19.3%, n = 96,500), the AUC was 0.983 (sensitivity 87.7%, specificity 95.7%, F_1 score 0.853). The most common code was normal sinus rhythm, with a prevalence of 63.4% (n = 316,500) in the testing dataset.

Table 2 Diagnostic performance of the model for the determination of primary and secondary rhythms.

		Preferred ROC metrics for common codes				Preferred PR metrics for uncommon codes			
ECG code	Prevalence, % (n)	AUC (CI)	Sensitivity	Specificity	F_1	AP (CI)	Sensitivity	Specificity	F_1
Normal ECG	19.3 (96,500)	0.983 (0.982–0.983)	0.877	0.957	0.853	0.924 (0.924–0.925)	0.999	0.669	0.591
Primary rhythms									
Normal sinus rhythm	63.4 (316,722)	0.998 (0.998–0.999)	0.887	0.998	0.939	0.999 (0.999–0.999)	1.00	0.673	0.914
Sinus bradycardia	14.5 (72,286)	0.999 (0.999–0.999)	0.994	0.968	0.911	0.993 (0.993–0.993)	1.00	0.554	0.431
Atrial fibrillation	8.5 (42,736)	0.999 (0.999–0.999)	1.00	0.694	0.380	0.988 (0.988–0.989)	1.00	0.628	0.334
Sinus tachycardia	7.2 (35,827)	0.999 (0.999–0.999)	1.00	0.643	0.302	0.990 (0.990–0.991)	1.00	0.582	0.270
Atrial flutter	1.9 (9699)	0.995 (0.995–0.996)	1.00	0.510	0.075	0.864 (0.862–0.866)	1.00	0.704	0.118
Ectopic atrial rhythm	0.47 (2340)	0.992 (0.991–0.994)	1.00	0.497	0.018	0.649 (0.643–0.655)	1.00	0.510	0.012
Junctional rhythm	0.35 (1735)	0.997 (0.996–0.998)	1.00	0.470	0.013	0.719 (0.712–0.726)	0.997	0.913	0.074
Ectopic atrial tachycardia	0.30 (1514)	0.987 (0.985–0.989)	1.00	0.455	0.011	0.385 (0.378–0.391)	0.997	0.733	0.022
Supraventricular tachycardia	0.28 (1387)	0.997 (0.996–0.998)	1.00	0.493	0.011	0.697 (0.689–0.704)	0.998	0.875	0.043
Ectopic atrial bradycardia	0.14 (688)	0.998 (0.997–0.999)	1.00	0.434	0.005	0.633 (0.622–0.645)	0.991	0.954	0.056
Wandering atrial pacemaker	0.09 (467)	0.975 (0.969–0.980)	0.994	0.464	0.003	0.090 (0.866–0.942)	0.989	0.629	0.005
Multifocal atrial tachycardia	0.07 (357)	0.997 (0.995–0.999)	1.00	0.414	0.002	0.283 (0.272–0.294)	0.975	0.972	0.048
Junctional bradycardia	0.05 (234)	0.999 (0.998–1.00)	1.00	0.416	0.002	0.608 (0.588–0.628)	0.996	0.990	0.082
Ventricular tachycardia	0.04 (185)	0.998 (0.996–1.00)	1.00	0.392	0.001	0.623 (0.601–0.646)	0.995	0.932	0.011
Junctional tachycardia	0.03 (167)	0.998 (0.995–1.00)	1.00	0.441	0.001	0.269 (0.253–0.284)	0.988	0.971	0.023
Idioventricular rhythm	0.03 (136)	0.997 (0.994–1.00)	1.00	0.391	0.001	0.230 (0.215–0.246)	0.956	0.985	0.033
Secondary rhythms									
Premature atrial complexes	6.4 (32,173)	0.993 (0.993–0.994)	0.999	0.584	0.248	0.922 (0.921–0.923)	0.999	0.543	0.231
Premature ventricular complexes	6.3 (31,277)	0.997 (0.997–0.997)	1.00	0.526	0.220	0.952 (0.951–0.953)	1.00	0.548	0.228
Sinus arrhythmia	4.4 (21,830)	0.982 (0.981–0.983)	0.995	0.679	0.221	0.802 (0.800–0.804)	0.998	0.539	0.165
Junctional escape beats	0.14 (708)	0.979 (0.975–0.983)	1.00	0.462	0.005	0.141 (0.136–0.146)	0.993	0.779	0.013
Ventricular escape beats	0.02 (86)	0.985 (0.976–0.994)	1.00	0.388	0.001	0.027 (0.024–0.030)	0.953	0.932	0.005
Premature junctional complexes	0.01 (58)	0.960 (0.942–0.978)	0.983	0.464	0.000	0.011 (0.974–1.22)	0.966	0.701	0.001

AP = average precision score; AUC = area under the curve; CI = confidence interval; PR = precision recall curve; ROC = receiver operator characteristic curve.

Table 3 Diagnostic performance of the model for the detection of axis deviation and chamber enlargement

ECG code	Prevalence, % (n)	Preferred ROC metrics for common codes				Preferred PR metrics for uncommon codes			
		AUC (CI)	Sensitivity	Specificity	F_1	AP (CI)	Sensitivity	Specificity	F_1
Axis deviation									
Right axis deviation	0.58 (2923)	0.993 (0.992–0.994)	1.00	0.458	0.021	0.469 (0.464–0.474)	0.997	0.934	0.151
Left axis deviation	0.55 (2761)	0.977 (0.975–0.979)	1.00	0.555	0.024	0.236 (0.232–0.239)	0.989	0.793	0.050
Right superior axis deviation	0.29 (1458)	0.997 (0.996–0.998)	1.00	0.431	0.010	0.568 (0.561–0.576)	0.997	0.960	0.127
Atrial enlargement									
Left atrial enlargement	3.8 (19,220)	0.980 (0.979–0.980)	0.987	0.710	0.214	0.757 (0.755–0.759)	0.996	0.525	0.143
Right atrial enlargement	0.46 (2275)	0.992 (0.991–0.994)	0.999	0.511	0.018	0.714 (0.708–0.720)	0.998	0.628	0.024
Biatrial enlargement	0.23 (1127)	0.995 (0.993–0.996)	1.00	0.415	0.008	0.607 (0.598–0.616)	0.998	0.806	0.023
Ventricular hypertrophy									
Left ventricular hypertrophy	7.3 (36,526)	0.988 (0.988–0.989)	0.964	0.936	0.695	0.876 (0.875–0.878)	1.00	0.544	0.257
Right ventricular hypertrophy	0.26 (1312)	0.996 (0.995–0.998)	1.00	0.412	0.009	0.615 (0.607–0.624)	0.998	0.890	0.045
Biventricular hypertrophy	0.005 (24)	0.994 (0.982–1.01)	1.00	0.468	0.000	0.164 (0.138–0.190)	0.958	0.953	0.002

AP = average precision score; AUC = area under the curve; CI = confidence interval; PR = precision recall curve; ROC = receiver operator characteristic curve.

For more common secondary rhythm codes such as premature atrial complexes (prevalence 6.4%, n = 32,173), premature ventricular complexes (prevalence 6.3%, n = 31,277), and sinus arrhythmia (prevalence 4.4%, n = 21,830), ROC metrics were favored. For less common primary and secondary rhythm codes, the specificity was greatly improved using PR metrics compared to the ROC metrics.

Axis deviation and chamber enlargement detection

Table 3 summarizes the model's diagnostic performance for detecting axis deviation and chamber enlargement. The sensitivity for detecting all forms of axis deviation (prevalence ≤0.58% for each) was 100% using ROC metrics and ≥98.9% using PR metrics. Again, the PR curve was a better indicator of specificity than the ROC curve for these relatively rare codes.

Left atrial enlargement (prevalence 7.3%, n = 36,526) and left ventricular hypertrophy (prevalence 7.3%, n = 36,526) were the most common forms of chamber enlargement. The ROC specificity for these relatively more prevalent codes was much better than the PR specificity. PR specificity was better than ROC specificity for less common forms of chamber enlargement, such as right atrial enlargement and right ventricular hypertrophy.

Atrioventricular and intraventricular conduction delay detection

Table 4 summarizes the model's diagnostic performance for detecting AV and intraventricular conduction delay. Similarly, PR specificity (96.1%) proved better than ROC specificity (36.5%) for detecting third-degree AV block (prevalence 0.14%, n = 714). The PR specificities appeared to be better metrics than ROC specificities for all forms of intraventricular conduction delay.

Myocardial ischemia detection

Table 5 summarizes the model's diagnostic performance for detecting myocardial ischemia. Anterolateral infarct was the most common of these codes, with a prevalence of 7.4% (n = 36,993). The ROC and PR sensitivities for all other infarct codes (ie, anteroseptal, lateral, posterior, anterior, and inferior infarct) were nearly 100%. The PR specificity (86.6%) was greater than the ROC specificity (49.6%) for the less prevalent anterior infarct code (prevalence 1.2%, n = 6248).

In general, myocardial injury detection was much less common compared to infarct detection. The ROC sensitivity was 100% for every myocardial injury code. The most common myocardial injury code reported in the testing dataset was inferior injury (prevalence 0.10%, n =486).

Waveform abnormality, clinical disorder, and pacemaker activity detection

Table 6 summarizes the model's diagnostic performance for detecting waveform abnormalities, clinical disorders, and pacemaker activity. Clinical disorders and pacemaker codes were relatively uncommon. The PR specificity for detection

Table 4 Diagnostic performance of the model for the detection of atrioventricular and intraventricular conduction delay

		Preferred ROC metrics for common codes				Preferred PR metrics for uncommon codes			
ECG code	Prevalence, % (n)	AUC (CI)	Sensitivity	Specificity	F_1	AP (CI)	Sensitivity	Specificity	F_1
AV conduction delay									
First-degree AV block	11.4 (56,867)	0.989 (0.989–0.989)	0.937	0.958	0.826	0.934 (0.934–0.935)	0.999	0.580	0.379
Variable AV block	1.2 (6218)	0.995 (0.995–0.996)	1.00	0.410	0.409	0.808 (0.805–0.812)	0.995	0.914	0.225
2:1 AV block	0.49 (2442)	0.996 (0.995–0.997)	1.00	0.493	0.019	0.759 (0.753–0.765)	0.999	0.701	0.032
Second-degree AV block, type I	0.15 (774)	0.995 (0.993–0.997)	1.00	0.433	0.005	0.609 (0.599–0.620)	0.996	0.879	0.025
4:1 AV block	0.13 (658)	1.00 (0.999–1.00)	1.00	0.446	0.005	0.859 (0.849–0.868)	0.998	0.967	0.075
Third-degree AV block	0.14 (714)	0.997 (0.995–0.998)	1.00	0.365	0.004	0.536 (0.525–0.547)	0.985	0.961	0.068
Second-degree AV block, type II	0.06 (296)	0.992 (0.988–0.996)	0.997	0.466	0.002	0.445 (0.429–0.461)	0.993	0.733	0.004
3:1 AV block	0.04 (223)	0.994 (0.991–0.998)	1.00	0.424	0.002	0.373 (0.356–0.390)	0.991	0.864	0.006
Intraventricular conduction delay									
Right bundle branch block	5.9 (29,333)	0.999 (0.999–0.999)	1.00	0.543	0.214	0.979 (0.978–0.979)	1.00	0.932	0.648
Left bundle branch block	3.3 (16,635)	0.999 (0.999–0.999)	1.00	0.532	0.128	0.964 (0.963–0.965)	1.00	0.912	0.437
Left anterior fascicular block	2.1 (10,280)	0.972 (0.970–0.973)	0.979	0.844	0.208	0.417 (0.415–0.420)	0.996	0.707	0.125
Left posterior fascicular block	1.6 (8154)	0.998 (0.997–0.998)	1.00	0.412	0.053	0.874 (0.872–0.877)	0.999	0.939	0.353
Nonspecific IVCD	0.77 (3836)	0.982 (0.980–0.983)	0.999	0.570	0.035	0.351 (0.347–0.355)	0.995	0.711	0.051
Bifascicular block	0.50 (2483)	0.996 (0.995–0.997)	1.00	0.389	0.016	0.548 (0.542–0.554)	1.00	0.826	0.054

AP = average precision score; AUC = area under the curve; AV = atrioventricular; CI = confidence interval; IVCD = intraventricular conduction delay; PR = precision recall curve; ROC = receiver operator characteristic curve.

Table 5 Diagnostic performance of the model for the detection of myocardial ischemia

		Preferred ROC metrics for common codes				Preferred PR metrics for uncommon codes			
ECG code	Prevalence, % (n)	AUC (CI)	Sensitivity	Specificity	F_1	AP (CI)	Sensitivity	Specificity	F_1
Myocardial infarction									
Anterolateral infarct	7.4 (36,993)	0.983 (0.982–0.983)	0.904	0.951	0.718	0.839 (0.837–0.840)	1.00	0.556	0.265
Anteroseptal infarct	2.5 (12,413)	0.976 (0.975–0.977)	0.992	0.796	0.199	0.541 (0.538–0.544)	0.998	0.685	0.139
Lateral infarct	2.5 (12,335)	0.975 (0.974–0.976)	0.995	0.747	0.166	0.502 (0.499–0.505)	0.999	0.515	0.094
Posterior infarct	2.1 (10,280)	0.987 (0.986–0.988)	0.999	0.667	0.112	0.684 (0.681–0.687)	1.00	0.584	0.092
Anterior infarct	1.2 (6248)	0.993 (0.993–0.994)	1.00	0.496	0.048	0.670 (0.666–0.674)	0.999	0.866	0.158
Inferior infarct	0.70 (3494)	0.978 (0.976–0.979)	0.997	0.572	0.032	0.392 (0.387–0.396)	0.993	0.684	0.042
Myocardial injury									
Inferior injury	0.10 (486)	0.993 (0.991–0.996)	1.00	0.448	0.004	0.400 (0.388–0.412)	0.992	0.848	0.013
Anterolateral injury	0.05 (256)	0.997 (0.984–0.995)	1.00	0.433	0.002	0.295 (0.138–0.157)	0.992	0.956	0.023
Inferolateral injury	0.04 (189)	0.996 (0.992–0.999)	1.00	0.438	0.001	0.262 (0.248–0.276)	0.984	0.903	0.008
Anterior injury	0.03 (163)	0.989 (0.984–0.995)	1.00	0.403	0.001	0.147 (0.138–0.157)	0.982	0.861	0.005
Lateral injury	0.02 (79)	0.978 (0.967–0.990)	1.00	0.460	0.001	0.059 (0.531–0.646)	0.949	0.845	0.002

AP = average precision score; AUC = area under the curve; CI = confidence interval; PR = precision recall curve; ROC = receiver operator characteristic curve.

of ventricular pacemaker (prevalence 0.97%, n = 4825) and dual-chamber pacemaker (prevalence 0.75%, n = 3733) activity was 81.7% and 76.5%, respectively.

Discussion

We demonstrate that an AI-ECG algorithm is capable of generating codes from a retrospective dataset that are consistent with those determined by cardiologists during routine care. Overall, the model performed very well for a wide range of rhythm, conduction, ischemia, waveform morphology, and secondary diagnoses codes with an area under the ROC curve (AUC) of ≥0.98 for 62 of the 66 reported codes. For heavily imbalanced codes, PR metrics demonstrated a sensitivity of ≥95% for all codes. In an era of ever-growing ECG signal availability, these findings demonstrate of the capability of an AI-ECG algorithm to make prediction beyond rhythm abnormalities.

In this study, we demonstrate the ability of an AI-ECG algorithm to generate a wide range of ECG codes. We believe that the over-reading cardiologist's interpretation provides a clinically relevant standard for comparison as they provide the final read upon which medical decisions are made. These findings demonstrate that an AI-ECG algorithm has the potential to streamline workflow and potentially improve the consistency and accuracy of current ECG analysis. A 3-way head-to-head prospective trial (ie, computer-generated interpretation vs over-reading cardiologist interpretation vs AI-ECG algorithm) would help to assess the AI-ECG algorithm's performance and validate such claims.

While current automated approaches rely on class-specific feature extraction, CNNs allow for a fundamentally different approach in which a single algorithm can accomplish the same tasks while leveraging both similar and unique features between all 66 diagnostic ECG codes. Moreover, CNNs continually improve with additional high-quality raw data. This suggests that it has the capability to not only improve its current prediction performance but also learn important manually derived features and those yet to be recognized.

Early CNNs were constrained by the number of layers as well as algorithmic and computational limitations.[22,23] More recent employment of deeper networks and end-to-end CNN approaches has demonstrated satisfactory performance for detection of atrial fibrillation and ventricular arrhythmias,[24,25] and even single-lead rhythm determination.[18] Despite these promising efforts, it has not been evaluated whether an AI-ECG algorithm could provide a comprehensive 12-lead ECG interpretation on a large unselected cohort. Our novel approach demonstrates that a CNN as a multilabel classifier is capable of learning features from 66 discrete, structured diagnostic ECG codes and providing simultaneous prediction probabilities for standard 12-lead ECG interpretation. This includes the prediction of primary and secondary rhythms and detection of axis deviation, chamber enlargement, AV and intraventricular conduction delay, myocardial ischemia, waveform abnormalities, clinical disorders, and pacemaker activity.

Table 6 Diagnostic performance of the model for the detection of waveform abnormalities, clinical disorders, and pacemaker activity

		Preferred ROC metrics for common codes				Preferred PR metrics for uncommon codes			
ECG code	Prevalence, % (n)	AUC (CI)	Sensitivity	Specificity	F_1	AP (CI)	Sensitivity	Specificity	F_1
Waveform abnormalities									
Low QRS voltage	5.1 (25,350)	0.984 (0.983–0.984)	0.956	0.920	0.553	0.819 (0.818–0.821)	0.998	0.565	0.197
Prolonged QT interval	4.6 (23,231)	0.960 (0.960–0.961)	0.792	0.939	0.521	0.596 (0.594–0.598)	0.998	0.560	0.181
Short PR interval	0.95 (4769)	0.988 (0.987–0.989)	1.00	0.573	0.043	0.577 (0.573–0.581)	0.997	0.713	0.063
Clinical disorders									
Early repolarization	1.6 (8123)	0.965 (0.964–0.966)	0.994	0.573	0.071	0.434 (0.431–0.437)	0.995	0.532	0.066
Wolff-Parkinson-White	0.04 (199)	0.996 (0.993–0.999)	1.00	0.389	0.001	0.580 (0.559–0.601)	0.985	0.925	0.010
Dextrocardia	0.03 (129)	0.997 (0.994–1.00)	1.00	0.418	0.001	0.603 (0.577–0.630)	0.992	0.907	0.006
Acute pericarditis	0.01 (64)	0.999 (0.996–1.00)	1.00	0.442	0.000	0.077 (0.686–0.854)	0.984	0.994	0.043
Pacemaker activity									
Ventricular pacemaker	0.97 (4825)	0.998 (0.998–0.999)	1.00	0.476	0.036	0.915 (0.912–0.918)	0.999	0.817	0.096
Dual-chamber pacemaker	0.75 (3733)	0.996 (0.995–0.997)	1.00	0.503	0.029	0.834 (0.830–0.838)	0.999	0.765	0.060

AP = average precision score; AUC = area under the curve; CI = confidence interval; PR = precision recall curve; ROC = receiver operator characteristics curve.

This current work is unique because while other investigators have shown CNN algorithms capable of predicting primary rhythms from single-lead ECGs, we have demonstrated its capability on standard 12-lead ECGs used in clinical practice. Furthermore, at our institution, current workflow is highly resource intensive and difficult to scale. It consists of more than 100 trained ECG technologists staffing our ECG lab 24/7 with physician oversight. Our goal with this work was to demonstrate that the CNN interpretation performs similarly well and that such a technology could be used to scale the scope of our work and improve the consistency of results. In fact, diagnostic performance was on par with the reference cardiologist's interpretation, which includes a multistep process of automated interpretation, certified rhythm analysis technician overview, and expert practicing cardiologist finalization. Thus, the unique features of this work include its high diagnostic performance for a large number of primary and secondary ECG diagnostic codes, use of standard 12-lead ECGs obtained in clinical practice, ability to potentially scale a highly resource-intensive practice, and use of nearly 2.5 million ECGs from more than 720,000 patients to derive, validate, and test the algorithm.

Owing to the known limitations and inconsistent performance of current ECG analysis algorithms, AI-ECG has the potential to increase overall accuracy and cause a paradigm shift in standard automated preliminary ECG interpretation methods. The ability to selectively choose the inputted data also permits for training and development of accurate region-, population-, and disease-specific prediction models. While this may not replace expert provider confirmation, a CNN with accurate 12-lead interpretation prediction capabilities could expedite clinical workflow by facilitating triage in emergent settings, providing immediate warning of critical results, and prioritizing ECGs requiring provider review (eg, low-confidence predictions).

There are several areas that require further investigation and clarification. One important question to answer is why lower sensitivities were related to myocardial injury localization. It is possible that the presence of artifact and nonischemic ST-segment elevation patterns (eg, early repolarization) contribute to inaccurate interpretations of acute ST-segment elevation myocardial infarction patterns. Currently implemented algorithms have high false-positive and false-negative rates for predicting acute ST-segment elevation myocardial infarction, making it impractical to rely solely on them to activate the cardiac catheterization laboratory. You could imagine that if these rates were reduced, this could be a very important clinical tool that could aid in triaging critical ECGs that may improve door-to-balloon time and patient outcomes.

Another challenge for algorithms and our CNN-based algorithm was precise and accurate QT-interval measurement. In general, current algorithms work by identifying the longest QT interval on the standard 12-lead ECG. Unfortunately, this can be compromised by poor-quality ECGs (eg, artifact) or other aspects of the complex interfering with accurate measurement (eg, U waves occurring at the end of a T wave, thereby artificially prolonging the QT interval). Accurate QT measurement with a single lead is erroneous, as individual lead measurements typically underestimate the true QT interval. Admittedly, this will be a difficult hurdle to overcome and certainly a limitation of new and popularized single-lead ECGs.

Lastly, we note that while there were no ECGs included in both the training and testing datasets, there were some patients who contributed ECGs to each group. If the ECGs were similar within an individual, it is possible that some of the features/codes would be more easily recognized by the network based on similarity between ECGs from the same individuals in the training and testing sets. Indeed, this issue has been discussed in prior similar studies.[18,26,27] Although it does not eliminate this potential bias, we do note that the ECGs included in this study came from 720,978 patients, and therefore we anticipate that a wide range of ECG variation was well represented and a small number of easily recognizable and unique features did not excessively bias the sample.

We used expert ECG annotation provided by cardiologist over-read during routine clinical care as our "gold standard." However, a cardiologist's final interpretation is not infallible and could allow for perpetuation of erroneous diagnoses. However, its performance relies on the cardiologist's final interpretation, which is not infallible and could allow for perpetuation of erroneous diagnoses. Nevertheless, clinical decisions are made every day with these expert yet error-prone human interpretations and therefore provide for a reasonable preliminary standard. In addition, the large sample size of patients and ECGs for the derivation, validation, and testing datasets with a similar distribution may mitigate the impact of these potential labeling errors. Further study is needed to evaluate the AI-ECG algorithm's performance in real time. A head-to-head comparison trial whereby the cardiologists select between automated interpretations made by a currently implemented computer-based algorithm, over-reading cardiologist, or the AI-ECG algorithm would also be interesting to evaluate provider preference and inter-annotator agreement. Because CNN-based algorithms rely on raw ECG data tracings as input, it will be key that future algorithms be developed, validated, and tested on high-quality, artifact-free ECGs from various populations. We suspect that computerized ECG interpretations will serve as an adjunct and not a substitute for the over-reading provider. External validation of our model in diverse, population-specific datasets needs to be assessed to verify its accuracy. This suggests that an AI-ECG algorithm can serve as a useful and supportive tool for ECG interpretation.

Conclusions

We demonstrate that an AI-ECG algorithm can provide comprehensive interpretation of a standard 12-lead ECG with accuracy comparable to board-certified, practicing cardiologists. Further study is warranted to compare its performance against currently implemented algorithms as well as improve its accuracy and scalability for clinical application.

Sources of Funding: This project was conceived, executed, and funded by Mayo Clinic without industry support. Disclosures: The authors have no conflicts of interest to disclose.

Appendix

Supplementary data

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.cvdhj.2020.08.005.

References

1. Taback L, Marden E, Mason HL, Pipberger HV. Digital recording of electrocardiographic data for analysis by a digital computer. IRE Trans Med Electro 1959; 6:167–171.
2. Caceres CA, Steinberg CA, Abraham S, et al. Computer extraction of electrocardiographic parameters. Circulation 1962;25:356–362.
3. Schlapfer J, Wellens HJ. Computer-interpreted electrocardiograms: benefits and limitations. J Am Coll Cardiol 2017;70:1183–1192.
4. Shah AP, Rubin SA. Errors in the computerized electrocardiogram interpretation of cardiac rhythm. J Electrocardiol 2007;40:385–390.
5. Guglin ME, Thatai D. Common errors in computer electrocardiogram interpretation. Int J Cardiol 2006;106:232–237.
6. Poon K, Okin PM, Kligfield P. Diagnostic performance of a computer-based. ECG rhythm algorithm. J Electrocardiol 2005;38:235–238.
7. Garvey JL, Zegre-Hemsey J, Gregg RE, Studnek JR. Electrocardiographic diagnosis of ST segment elevation myocardial infarction: an evaluation of three automated interpretation algorithms. J Electrocardiol 2016;49:728–732.
8. Mawri S, Michaels A, Gibbs J, et al. The comparison of physician to computer interpreted electrocardiograms on ST-elevation myocardial infarction door-to-balloon times. Crit Pathw Cardiol 2016;15:22–25.
9. Partinez-Losas P, Higueras J, Gomez-Polo JC, et al. The influence of computerized interpretation of an electrocardiogram reading. Am J Emerg Med 2016; 34:2031–2032.
10. Novotny T, Bond R, Andrsova I, et al. The role of computerized diagnostic proposals in the interpretation of the 12-lead electrocardiogram by cardiology and non-cardiology fellows. Int J Med Inform 2017;101:85–92.
11. Bogun F, Anh D, Kalahasty G, et al. Misdiagnosis of atrial fibrillation and its clinical consequences. Am J Med 2004;117:636–642.
12. Anh D, Krishnan S, Bogun F. Accuracy of electrocardiogram interpretation by cardiologists in the setting of incorrect computer analysis. J Electrocardiol 2006;39:343–345.
13. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–444.
14. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. Proc IEEE Int Conf Comput Vis 2015;1026–1034.
15. He K, Zhang X, Sun J, et al. Deep residual learning for image recognition. IEEE Conf Comput Vis 2016;770–778.
16. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402–2410.
17. Esteva A. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–118.
18. Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019;25:65–69.
19. Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nat Med 2019;25:70–74.
20. Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. Lancet 2019; 394:861–867.
21. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. Proceedings of Machine Learning Research 2010; 9:249–256.
22. Holst H, Ohlsson M, Peterson C, Edenbrandt L. A confident decision support system for interpreting electrocardiograms. Clin Physiol 1999;19:410–418.
23. Cubanski D, Cyganski D, Antman EM, Feldman CL. A neural network system for detection of atrial fibrillation in ambulatory electrocardiograms. J Cardiovasc Electrophysiol 1994;5:602–608.
24. Xiong Z, Zhao J, Stiles MK. Robust ECG signal classification for detection of atrial fibrillation using a novel neural network. Comput Cardiol 2017; https://doi.org/10.22489/CinC.2017.066-138.
25. Acharya UR, Fujita H, Lih OS, Hagiwara Y, Tan JH, Adam M. Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network. Inf Sci (NY) 2017;405:81–90.
26. Alfaras M, Soriano MC, Ortín S. A fast machine learning model for ECG-based heartbeat classification and arrhythmia detection. Front Phys 2019;7.
27. de Chazal P, O'Dwyer M, Reilly RB. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. IEEE Trans Biomed Eng 2004; 51:1196–1206.