nature medicine

FOCUS | LETTERS

https://doi.org/10.1038/s41591-018-0268-3

Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network

Awni Y. Hannun[1,6*], Pranav Rajpurkar[1,6], Masoumeh Haghpanahi[2,6], Geoffrey H. Tison[3,6], Codie Bourn[2], Mintu P. Turakhia[4,5] and Andrew Y. Ng[1]

Computerized electrocardiogram (ECG) interpretation plays a critical role in the clinical ECG workflow[1]. Widely available digital ECG data and the algorithmic paradigm of deep learning[2] present an opportunity to substantially improve the accuracy and scalability of automated ECG analysis. However, a comprehensive evaluation of an end-to-end deep learning approach for ECG analysis across a wide variety of diagnostic classes has not been previously reported. Here, we develop a deep neural network (DNN) to classify 12 rhythm classes using 91,232 single-lead ECGs from 53,549 patients who used a single-lead ambulatory ECG monitoring device. When validated against an independent test dataset annotated by a consensus committee of board-certified practicing cardiologists, the DNN achieved an average area under the receiver operating characteristic curve (ROC) of 0.97. The average F_1 score, which is the harmonic mean of the positive predictive value and sensitivity, for the DNN (0.837) exceeded that of average cardiologists (0.780). With specificity fixed at the average specificity achieved by cardiologists, the sensitivity of the DNN exceeded the average cardiologist sensitivity for all rhythm classes. These findings demonstrate that an end-to-end deep learning approach can classify a broad range of distinct arrhythmias from single-lead ECGs with high diagnostic performance similar to that of cardiologists. If confirmed in clinical settings, this approach could reduce the rate of misdiagnosed computerized ECG interpretations and improve the efficiency of expert human ECG interpretation by accurately triaging or prioritizing the most urgent conditions.

The electrocardiogram is a fundamental tool in the everyday practice of clinical medicine, with more than 300 million ECGs obtained annually worldwide[3]. The ECG is pivotal for diagnosing a wide spectrum of abnormalities from arrhythmias to acute coronary syndrome[4]. Computer-aided interpretation has become increasingly important in the clinical ECG workflow since its introduction over 50 years ago, serving as a crucial adjunct to physician interpretation in many clinical settings[1]. However, existing commercial ECG interpretation algorithms still show substantial rates of misdiagnosis[1,5–7]. The combination of widespread digitization of ECG data and the development of algorithmic paradigms that can benefit from large-scale processing of raw data presents an opportunity to reexamine the standard approach to algorithmic ECG analysis and may provide substantial improvements to automated ECG interpretation.

Substantial algorithmic advances in the past five years have been driven largely by a specific class of models known as deep neural networks[2]. DNNs are computational models consisting of multiple processing layers, with each layer being able to learn increasingly abstract, higher-level representations of the input data relevant to perform specific tasks. They have dramatically improved the state of the art in speech recognition[8], image recognition[9], strategy games such as Go[10], and in medical applications[11,12]. The ability of DNNs to recognize patterns and learn useful features from raw input data without requiring extensive data preprocessing, feature engineering or handcrafted rules[2] makes them particularly well suited to interpret ECG data. Furthermore, since DNN performance tends to increase as the amount of training data increases[2], this approach is well positioned to take advantage of the widespread digitization of ECG data.

A comprehensive evaluation of whether an end-to-end deep learning approach can be used to analyze raw ECG data to classify a broad range of diagnoses remains lacking. Much of the previous work to employ DNNs toward ECG interpretation has focused on single aspects of the ECG processing pipeline, such as noise reduction[13,14] or feature extraction[15,16], or has approached limited diagnostic tasks, detecting only a handful of heartbeat types (normal, ventricular or supraventricular ectopic, fusion, and so on)[17–20] or rhythm diagnoses (most commonly atrial fibrillation or ventricular tachycardia)[21–25]. Lack of appropriate data has limited many efforts beyond these applications. Most prior efforts used data from the MIT-BIH Arrhythmia database (PhysioNet)[26], which is limited by the small number of patients and rhythm episodes present in the dataset.

In this study, we constructed a large, novel ECG dataset that underwent expert annotation for a broad range of ECG rhythm classes. We developed a DNN to detect 12 rhythm classes from raw single-lead ECG inputs using a training dataset consisting of 91,232 ECG records from 53,549 patients. The DNN was designed to classify 10 arrhythmias as well as sinus rhythm and noise for a total of 12 output rhythm classes (Extended Data Fig. 1). ECG data were recorded by the Zio monitor, which is a Food and Drug Administration (FDA)-cleared, single-lead, patch-based ambulatory ECG monitor[27] that continuously records data from a single vector (modified Lead II) at 200 Hz. The mean and median wear time of the Zio monitor in our dataset was 10.6 and 13.0 days, respectively. Mean age was 69 ± 16 years and 43% were women. We validated the DNN on a test dataset that consisted of 328 ECG records collected from 328 unique patients, which was annotated by a consensus committee of expert cardiologists (see Methods). Mean age on the test dataset was 70 ± 17 years and 38% were women. The mean inter-annotator agreement on the test dataset was 72.8%.

[1]Department of Computer Science, Stanford University, Stanford, CA, USA. [2]iRhythm Technologies Inc., San Francisco, CA, USA. [3]Division of Cardiology, Department of Medicine, University of California San Francisco, San Francisco, CA, USA. [4]Department of Medicine and Center for Digital Health, Stanford University School of Medicine, Stanford, CA, USA. [5]Veterans Affairs Palo Alto Health Care System, Palo Alto, CA, USA. [6]These authors contributed equally: Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison. *e-mail: awni@cs.stanford.edu

Table 1 | Diagnostic performance of the DNN and averaged individual cardiologists compared to the cardiologist committee consensus ($n = 328$)

	Algorithm AUC (95% CI)[a]		Algorithm F_1[b]		Average cardiologist F_1	
	Sequence[a]	Set[b]	Sequence	Set	Sequence	Set
Atrial fibrillation and flutter	0.973 (0.966–0.980)	0.965 (0.932–0.998)	0.801	0.831	0.677	0.686
AVB	0.988 (0.983–0.993)	0.981 (0.953–1.000)	0.828	0.808	0.772	0.761
Bigeminy	0.997 (0.991–1.000)	0.996 (0.976–1.000)	0.847	0.870	0.842	0.853
EAR	0.913 (0.889–0.937)	0.940 (0.870–1.000)	0.541	0.596	0.482	0.536
IVR	0.995 (0.989–1.000)	0.987 (0.959–1.000)	0.761	0.818	0.632	0.720
Junctional rhythm	0.987 (0.980–0.993)	0.979 (0.946–1.000)	0.664	0.789	0.692	0.679
Noise	0.981 (0.973–0.989)	0.947 (0.898–0.996)	0.844	0.761	0.768	0.685
Sinus rhythm	0.975 (0.971–0.979)	0.987 (0.976–0.998)	0.887	0.933	0.852	0.910
SVT	0.973 (0.960–0.985)	0.953 (0.903–1.000)	0.488	0.693	0.451	0.564
Trigeminy	0.998 (0.995–1.000)	0.997 (0.979–1.000)	0.907	0.864	0.842	0.812
Ventricular tachycardia	0.995 (0.980–1.000)	0.980 (0.934–1.000)	0.541	0.681	0.566	0.769
Wenckebach	0.978 (0.967–0.989)	0.977 (0.938–1.000)	0.702	0.780	0.591	0.738
Frequency-weighted average	0.978	0.977	0.807	0.837	0.753	0.780

[a]DNN algorithm area under the ROC compared to the cardiologist committee consensus. [b]DNN algorithm and averaged individual cardiologist F_1 scores compared to the cardiologist committee consensus. Sequence-level describes the algorithm predictions that are made once every 256 input samples (approximately every 1.3 s) and are compared against the gold-standard committee consensus at the same intervals. Set-level describes the unique set of algorithm predictions that are present in the 30-s record. Sequence AUC prediction, $n = 7{,}544$; set AUC prediction, $n = 328$.

Supplementary Table 1 shows the number of unique patients exhibiting each rhythm class.

We first compared the performance of the DNN against the gold standard cardiologist consensus committee diagnoses by calculating the AUC (Table 1a). Since the DNN algorithm was designed to make a rhythm class prediction approximately once per second (see Methods), we report performance both as assessed once every second—which we call "sequence-level" and consists of one rhythm class per interval—and once per record, which we call "set-level" and consists of the group of unique diagnoses present in the record. Sequence-level metrics help capture the duration of an arrhythmia, such as its onset and offset within a record, whereas set-level metrics focus only on the existence of a rhythm class within a record. The DNN achieved an AUC of greater than 0.91 for all rhythm classes; at the sequence-level all but one AUC was above 0.97. The class-weighted average AUC was 0.978 at the sequence-level and 0.977 at the set-level. The model demonstrated high AUCs for arrhythmias of greater clinical significance such as AF, atrio-ventricular block, and ventricular tachycardia. The sequence and set-level results were similar, though sequence-level AUC was higher in the majority of cases. In sensitivity analyses, we calculated multi-class AUC using the method described by Hand and Till28 and results were materially unchanged. Supplementary Table 2 shows the maximum sensitivity achieved by the DNN with specificity >90%, and vice versa. With one exception, all sensitivity and specificity pairs were >90%.

In addition to a cardiologist consensus committee annotation, each ECG record in the test dataset received annotations from six separate individual cardiologists who were not part of the committee (see Methods). Using the committee labels as the gold standard, we compared the DNN algorithm F_1 score to the average individual cardiologist F_1 score, which is the harmonic mean of the positive predictive value (PPV; precision) and sensitivity (recall) (Table 1). Cardiologist F_1 scores were averaged over six individual cardiologists. The trend of DNN F_1 scores tended to follow that of the averaged cardiologist F_1 scores: both had lower F_1 on similar classes, such as ventricular tachycardia and ectopic atrial rhythm (EAR). The set-level average F_1 scores weighted by the frequency of each class for the DNN (0.837) exceeded those for the averaged cardiologist (0.780). We performed multiple sensitivity analyses, all of which were consistent with our main results: both AUC and F_1 scores on the 10% development dataset ($n = 8{,}761$) were materially unchanged from the test dataset results, although they were slightly higher (Supplementary Tables 3 and 4). In addition, we retrained the DNN holding out an additional 10% of the training dataset as a second held-out test dataset ($n = 8{,}768$); the AUC and F_1 scores for all rhythms were materially unchanged (Supplementary Tables 5 and 6). We note that unlike the primary test dataset, which has gold-standard annotations from a committee of cardiologists, both sensitivity analysis datasets are annotated by certified ECG technicians.

We plotted receiver operating characteristic curves (ROCs) and precision-recall curves for the sequence-level analyses of three example classes: atrial fibrillation; trigeminy; and AVB (Fig. 1a,b). Individual cardiologist performance and averaged cardiologist performance are plotted on the same figure. Extended Data Fig. 2 presents ROCs for all classes, showing that the model met or exceeded the averaged cardiologist performance for all rhythm classes. Fixing the specificity at the average specificity level achieved by cardiologists, the sensitivity of the DNN exceeded the average cardiologist sensitivity for all rhythm classes (Table 2). We used confusion matrices to illustrate the discordance between the DNN's predictions (Fig. 2a) or averaged cardiologist predictions (Fig. 2b) and the committee consensus. The two confusion matrices exhibit a similar pattern, highlighting those rhythm classes that were generally more problematic to classify (that is, supraventricular tachycardia (SVT) versus atrial fibrillation, junctional versus sinus rhythm, and EAR versus sinus rhythm).

Finally, to demonstrate the generalizability of our DNN architecture to external data, we applied our DNN to the 2017 PhysioNet Challenge data (https://physionet.org/challenge/2017/), which contained four rhythm classes: sinus rhythm; atrial fibrillation; noise; and other. Keeping our DNN architecture fixed and without any other hyper-parameter tuning, we trained our DNN on the publicly available training dataset ($n = 8{,}528$), holding out a 10% development dataset for early stopping. DNN performance on the hidden test dataset ($n = 3{,}658$) demonstrated overall F_1 scores that were among those of the best performers from the competition (Supplementary Table 7)[24], with a class average F_1 of 0.83. This demonstrates the ability of our end-to-end DNN-based approach to generalize to a new set of rhythm labels on a different dataset.

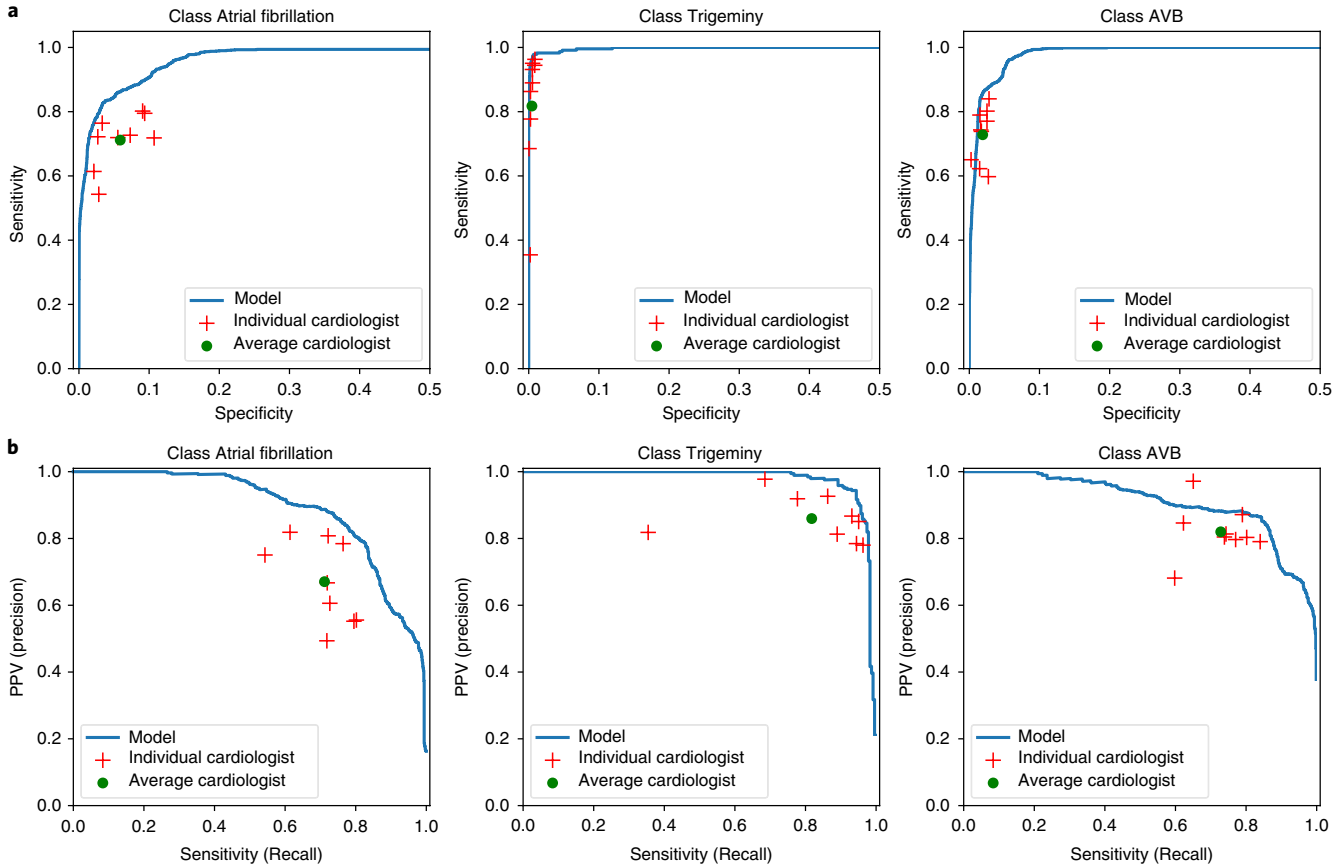


Fig. 1 | ROC and precision-recall curves. a, Examples of ROC curves calculated at the sequence level for atrial fibrillation (AF), trigeminy, and AVB. **b**, Examples of precision-recall curves calculated at the sequence level for atrial fibrillation, trigeminy, and AVB. Individual cardiologist performance is indicated by the red crosses and averaged cardiologist performance is indicated by the green dot. The line represents the ROC (**a**) or precision-recall curve (**b**) achieved by the DNN model. $n = 7{,}544$ where each of the 328 30-s ECGs received 23 sequence-level predictions.

Our study is the first comprehensive demonstration of a deep learning approach to perform classification across a broad range of the most common and important ECG rhythm diagnoses. Our DNN had an average class-weighted AUC of 0.97, with higher average F_1 scores and sensitivities than cardiologists. These findings demonstrate that an end-to-end DNN approach has the potential to be used to improve the accuracy of algorithmic ECG interpretation. Recent algorithmic and computational advances compel us to revisit the standard approaches to automated ECG interpretation. Furthermore, algorithmic approaches whose performance improves as more data become available, such as deep learning[2], can leverage the widespread digitization of ECG data and provide clear opportunities to bring us closer to the ideal of a learning health care system[29]. We emphasize our use in this study of a dataset large enough to evaluate an end-to-end deep learning approach to predict multiple diagnostic ECG classes, and our validation against the high standard of a cardiologist consensus committee. (Most cardiologists were subspecialized in rhythm abnormalities.) We believe this is the most clinically relevant gold standard, since cardiologists perform the final ECG diagnosis in nearly all clinical settings.

Our study demonstrates that the paradigm shift represented by end-to-end deep learning may enable a new approach to automated ECG analysis. The standard approach to automated ECG interpretation employs various techniques across a series of steps that include signal preprocessing, feature extraction, feature selection/reduction, and classification[30]. At each step, hand-engineered heuristics and derivations of the raw ECG data are developed with the ultimate aim to improve classification for a given rhythm, such as atrial fibrillation[31,32].

Table 2 | DNN algorithm and cardiologist sensitivity compared to the cardiologist committee consensus, with specificity fixed at the average specificity level achieved by cardiologists

	Specificity	Average cardiologist sensitivity	DNN algorithm sensitivity
Atrial fibrillation and flutter	0.941	0.710	0.861
AVB	0.981	0.731	0.858
Bigeminy	0.996	0.829	0.921
EAR	0.993	0.380	0.445
IVR	0.991	0.611	0.867
Junctional rhythm	0.984	0.634	0.729
Noise	0.983	0.749	0.803
Sinus rhythm	0.859	0.901	0.950
SVT	0.983	0.408	0.487
Ventricular tachycardia	0.996	0.652	0.702
Wenckebach	0.986	0.541	0.651

In contrast, DNNs enable an approach that is fundamentally different since a single algorithm can accomplish all of these steps 'end-to-end' without requiring class-specific feature extraction; in other words, the DNN can accept the raw ECG data as input and output diagnostic

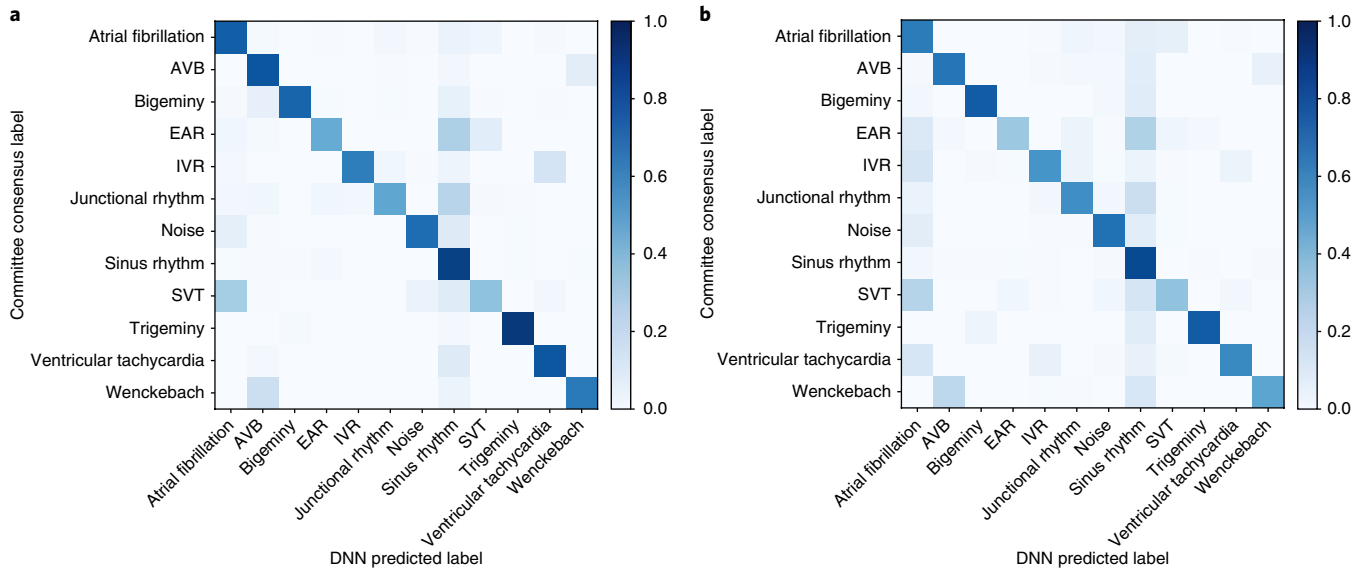


Fig. 2 | Confusion matrices. a, Confusion matrix for the predictions of the DNN versus the cardiology committee consensus. **b**, Confusion matrix for predictions of individual cardiologists versus the cardiology committee consensus. The percentage of all possible records in each category is displayed on a color gradient scale.

probabilities. With sufficient training data, using a DNN in this manner has the potential to learn all of the important previously manually derived features, along with as-yet-unrecognized features, in a data-driven way[2], and may learn shared features useful in predicting multiple classes. These properties of DNNs can serve to improve prediction performance, particularly since there is ample evidence to suggest that the currently recognized, manually derived ECG features represent only a fraction of the informative features for any diagnosis[33,34].

While artificial neural networks were first applied toward the interpretation of ECGs as early as two decades ago[3,35], until recently they only contained several layers and were constrained by algorithmic and computational limitations. More recent studies have employed deeper networks, although some only use DNNs to perform certain steps in the ECG processing pipeline, such as feature extraction[33] or classification[25]. End-to-end DNN approaches have been used more recently showing good performance for a limited set of ECG rhythms, such as atrial fibrillation[22,23,36], ventricular arrhythmias[21], or individual heartbeat classes[20,21,37,38]. While these prior efforts demonstrated promising performance for specific rhythms, they do not provide a comprehensive evaluation of whether an end-to-end approach can perform well across a wide range of rhythm classes, in a manner similar to that encountered clinically. Our approach is unique in using a 34-layer network in an end-to-end manner to simultaneously output probabilities for a wide range of distinct rhythm diagnoses, all of which is enabled by our dataset, which is orders of magnitude larger than most other datasets of its kind[26]. Distinct from some other recent DNN approaches[39], no substantial preprocessing of ECG data, such as Fourier or wavelet transforms[40], is needed to achieve strong classification performance.

Since arrhythmia detection is one of the most problematic tasks for existing ECG algorithms[1,5,6], if validated in clinical settings through clinical trials, our approach has the potential for substantial clinical impact. Paired with properly annotated digital ECG data, our approach has the potential to increase the overall accuracy of preliminary computerized ECG interpretations and can also be used to customize predictions to institution- or population-specific applications by additional training on institution-specific data. While expert provider confirmation will probably be appropriate in many clinical settings, the DNN could expand the capability of an expert over-reader in the clinical workflow, for example, by triaging urgent conditions or those for which the DNN has the least 'confidence'. Since ECG data collected from different clinical applications range in duration from 10 s (standard 12-lead ECGs) to multiple days (single-lead ambulatory ECGs), the application of any algorithm, including ours, must ultimately be tailored to the target clinical application. For example, even at the performance characteristics we report, applying our algorithm sequentially across an ECG record of long duration would result in nontrivial false-positive diagnoses. Faced with a similar problem, cardiologists probably incorporate additional mechanisms to improve their diagnostic performance, such as taking advantage of the increased context or knowledge about arrhythmia epidemiology. Similarly, additional algorithmic steps or post-processing heuristics may be important before clinical application.

An important finding from our study is that the DNN appears to recapitulate the misclassifications made by individual cardiologists, as demonstrated by the similarity in the confusion matrices for the model and cardiologists. Manual review of the discordances revealed that the DNN misclassifications overall appear very reasonable. In many cases, the lack of context, limited signal duration, or having a single lead limited the conclusions that could reasonably be drawn from the data, making it difficult to definitively ascertain whether the committee and/or the algorithm was correct. Similar factors, as well as human error, may explain the inter-annotator agreement of 72.8%.

Of the rhythm classes we examined, ventricular tachycardia is a clinically important rhythm for which the model had a lower F_1 score than cardiologists, but interestingly had higher sensitivity (94.1%) than the averaged cardiologist (78.4%). Manual review of the 16 records misclassified by the DNN as ventricular tachycardia showed that 'mistakes' made by the algorithm were very reasonable. For example, ventricular tachycardia and idioventricular rhythm (IVR) differ only in the heart rate being above or below 100 beats per minute (b.p.m.), respectively. In 7 of the committee-labeled IVR cases, the record contained periods of heart rate ≥ 100 b.p.m., making ventricular tachycardia a reasonable classification by the DNN; the remaining 3 committee-labeled IVR records had rates close to 100 b.p.m.. Of the 5 cases where the committee label was atrial fibrillation (4) or SVT (1), all but one displayed aberrant conduction, resulting in wide QRS complexes (the ECG waveform corresponding to ventricular activation) with a similar appearance to ventricular tachycardia. If we recategorize the 7 IVR records with a rate ≥ 100 b.p.m. as ventricular tachycardia, overall DNN performance on ventricular tachycardia exceeds that of cardiologists by F_1 score, with a set-level F_1 score of 0.82 (versus 0.77).

This study has several important limitations. Our input dataset is limited to single-lead ECG records obtained from an ambulatory monitor, which provides limited signal compared to a standard 12-lead ECG; it remains to be determined if our algorithm performance would be similar in 12-lead ECGs. However, it may be in applications such as this, which have lower signal-to-noise ratio and where the current standard of care leaves more room for improvement, that approaches such as deep learning may provide the greatest impact. As discussed earlier, a limitation facing this, or any algorithm, before clinical application would be tailoring it to the target application, which may require additional training or post-processing steps. Additionally, systematic differences in the way technicians versus cardiologists labeled records in our dataset could have decreased DNN performance, although we took precautions to limit this by establishing standard operating protocols for annotation. In addition, as revealed in our manual review of discordant predictions, in some cases there remains uncertainty in the correct label. Given the resource-intensive nature of cardiologist committee ECG annotation, our test dataset was limited to records from 328 patients; confidence intervals (CIs) with our test dataset size were acceptably narrow, as we report in Table 1, although our ability to perform subgroup analysis (such as by age/sex) is limited. Finally, we also note that to obtain a sufficient quantity of rare rhythms in our training and test datasets, we targeted patients exhibiting these rhythms during data extraction. This implies that prevalence-dependent metrics such as the F_1 score would not be expected to generalize to the broader population.

In summary, we demonstrate that an end-to-end deep learning approach can classify a broad range of distinct arrhythmias from single-lead ECGs with high diagnostic performance similar to that of cardiologists. If confirmed in clinical settings, this approach has the potential to improve the accuracy, efficiency, and scalability of ECG interpretation.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41591-018-0268-3.

Received: 12 March 2018; Accepted: 26 October 2018;
Published online: 7 January 2019

References

1. Schläpfer, J. & Wellens, H. J. Computer-interpreted electrocardiograms: benefits and limitations. *J. Am. Coll. Cardiol.* **70**, 1183–1192 (2017).
2. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
3. Holst, H., Ohlsson, M., Peterson, C. & Edenbrandt, L. A confident decision support system for interpreting electrocardiograms. *Clin. Physiol.* **19**, 410–418 (1999).
4. Schlant, R. C. et al. Guidelines for electrocardiography. A report of the American College of Cardiology/American Heart Association Task Force on assessment of diagnostic and therapeutic cardiovascular procedures (Committee on Electrocardiography). *J. Am. Coll. Cardiol.* **19**, 473–481 (1992).
5. Shah, A. P. & Rubin, S. A. Errors in the computerized electrocardiogram interpretation of cardiac rhythm. *J. Electrocardiol.* **40**, 385–390 (2007).
6. Guglin, M. E. & Thatai, D. Common errors in computer electrocardiogram interpretation. *Int. J. Cardiol.* **106**, 232–237 (2006).
7. Poon, K., Okin, P. M. & Kligfield, P. Diagnostic performance of a computer-based ECG rhythm algorithm. *J. Electrocardiol.* **38**, 235–238 (2005).
8. Amodei, D. et al. Deep Speech 2: end-to-end Speech recognition in English and Mandarin. In *Proc. 33rd International Conference on Machine Learning,* 173–182 (2016).
9. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In *Proc. International Conference on Computer Vision,* 1026–1034 (IEEE, 2015).
10. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
11. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).
12. Esteva, A. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
13. Poungponsri, S. & Yu, X. An adaptive filtering approach for electrocardiogram (ECG) signal noise reduction using neural networks. *Neurocomputing* **117**, 206–213 (2013).
14. Ochoa, A., Mena, L. J. & Felix, V. G. Noise-tolerant neural network approach for electrocardiogram signal classification. In *Proc. 3rd International Conference on Compute and Data Analysis,* 277–282 (Association for Computing Machinery, 2017).
15. Mateo, J. & Rieta, J. J. Application of artificial neural networks for versatile preprocessing of electrocardiogram recordings. *J. Med. Eng. Technol.* **36**, 90–101 (2012).
16. Pourbabaee, B., Roshtkhari, M. J. & Khorasani, K. Deep convolutional neural networks and learning ECG features for screening paroxysmal atrial fibrillation patients. *IEEE Trans. Syst. Man Cybern. Syst.* **99**, 1–10 (2017).
17. Javadi, M., Arani, S. A., Sajedin, A. & Ebrahimpour, R. Classification of ECG arrhythmia by a modular neural network based on mixture of experts and negatively correlated learning. *Biomed. Signal Process. Control* **8**, 289–296 (2013).
18. Acharya, U. R. et al. A deep convolutional neural network model to classify heartbeats. *Comput. Biol. Med.* **89**, 389–396 (2017).
19. Banupriya, C. V. & Karpagavalli, S. Electrocardiogram beat classification using probabilistic neural network. In *Proc. Machine Learning: Challenges and Opportunities Ahead* 31–37 (2014).
20. Al Rahhal, M. M. et al. Deep learning approach for active classification of electrocardiogram signals. *Inf. Sci. (NY)* **345**, 340–354 (2016).
21. Acharya, U. R. et al. Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network. *Inf. Sci. (NY)* **405**, 81–90 (2017).
22. Zihlmann, M., Perekrestenko, D. & Tschannen, M. Convolutional recurrent neural networks for electrocardiogram classification. *Comput. Cardiol.* https://doi.org/10.22489/CinC.2017.070-060 (2017).
23. Xiong, Z., Zhao, J. & Stiles, M. K. Robust ECG signal classification for detection of atrial fibrillation using a novel neural network. *Comput. Cardiol.* https://doi.org/10.22489/CinC.2017.066-138 (2017).
24. Clifford, G. et al. AF classification from a short single lead ECG recording: the PhysioNet/Computing in Cardiology Challenge 2017. *Comput. Cardiol.* https://doi.org/10.22489/CinC.2017.065-469 (2017).
25. Teijeiro, T., Garcia, C. A., Castro, D. & Felix, P. Arrhythmia classification from the abductive interpretation of short single-lead ECG records. *Comput. Cardiol.* https://doi.org/10.22489/CinC.2017.166-054 (2017).
26. Goldberger, A. L. et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101**, E215–E220 (2000).
27. Turakhia, M. P. et al. Diagnostic utility of a novel leadless arrhythmia monitoring device. *Am. J. Cardiol.* **112**, 520–524 (2013).

Acknowledgements

iRhythm Technologies, Inc. provided financial support for the data annotation in this work. M.H. and C.B. are employees of iRhythm Technologies, Inc. A.Y.H. was funded by an NVIDIA fellowship. G.H.T. received support from the National Institutes of Health (K23 HL135274). The only financial support provided by iRhythm Technologies, Inc. for this study was for the data annotation. Data analysis and interpretation was performed independently from the sponsor. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Author contributions

M.H., A.Y.N., A.Y.H., and G.H.T. contributed to the study design. M.H. and C.B. were responsible for data collection. P.R. and A.Y.H. ran the experiments and created the figures. G.H.T., P.R., and A.Y.H. contributed to the analysis. G.H.T., A.Y.H., and M.P.T. contributed to the data interpretation and to the writing. G.H.T., M.P.T., and A.Y.N. advised and A.Y.N. was the senior supervisor of the project. All authors read and approved the submitted manuscript.

Competing interests

M.H. and C.B. are employees of iRhythm Technologies, Inc. G.H.T. is an advisor to Cardiogram, Inc. M.P.T. is a consultant to iRhythm Technologies, Inc. None of the other authors have potential conflicts of interest.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-018-0268-3.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-018-0268-3.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to A.Y.H.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Study participants and sampling procedures. Our dataset contained retrospective, de-identified data from adult patients >18 years old who used the Zio monitor (iRhythm Technologies, Inc) from January 2013 to March 2017. All extracted data were de-identified according to the Health Insurance Portability and Accountability Act Safe Harbor. According to the iRhythm Technologies privacy policy, fully de-identified patient data may be shared externally for research purposes; patients may opt out of this sharing. Accordingly, written informed consent was not necessary for this study given that the 30-s ECG samples of both the training and test datasets were appropriately de-identified before use. The study was reviewed and exempted from full review by the Stanford University Institutional Review Board.

We extracted a median of one 30-s record per patient to construct the training dataset. ECG records were extracted based on the report summaries produced by iRhythm Technologies clinical workflow, which includes a full review by a certified ECG technician of initial annotations from an algorithm which is FDA 510(k) approved for clinical use. We randomly sampled patients exhibiting each rhythm; from these patients, we selected 30-s records where the rhythm class was present. Although the targeted rhythm class was typically present within the record, most records contained a mix of multiple rhythms. To further improve the balance of classes in the training dataset, rare rhythms such as AVB, were intentionally oversampled, with a median of two 30-s records per patient. For the test dataset, 30-s records of each rhythm were sampled in a similar manner to achieve a greater representation of rare rhythms; however, the test dataset included only a single record per patient. The training, development, and test datasets had completely disjointed sets of patients.

Annotation procedures. All ECG records in the training and test datasets underwent additional annotation procedures. We used separate procedures to annotate the training and test datasets, reserving the resource-intensive cardiologist annotation for use as the gold standard in the test dataset. To annotate the training dataset, a group of senior certified ECG technicians reviewed all records and noted the onset and offset of all rhythms on the record. Every record was randomly assigned to be reviewed by a single technician specifically for this task, not for any other purpose. All annotators received specific instructions and training regarding how to annotate transitions between rhythms to improve labeling consistency. We held out records from a random 10% of the training dataset patients for use as a development dataset to perform DNN hyper-parameter tuning.

Eight board-certified practicing cardiac electrophysiologists and one board-certified practicing cardiologist (all referred to as cardiologists) annotated records in the test dataset. All iRhythm Technologies clinical annotations were removed from the test dataset. Cardiologists were divided into three committees of three members each; each committee annotated a separate one-third of the test dataset (112 records). Cardiologist committees discussed records as a group and annotated by consensus, providing the gold standard for model evaluation. Each of the remaining six cardiologists that were not part of the committee for that record also provided individual annotations for that record. These annotations were used to compare the model's performance to that of the individual cardiologists. In summary, every record in the test dataset received one committee consensus annotation from a group of three cardiologists and six individual cardiologist annotations.

Many ECG records contained multiple rhythm class diagnoses since the onset and offset of all unique classes were labeled within each 30-s record. The atrial fibrillation class combined atrial fibrillation and atrial flutter. The AVB class combined both type 2 second-degree AVB (Mobitz II/Hay) and third-degree AVB. We combined these classes because they have similar clinical consequences. The noise label was selected whenever artifact in the signal precluded accurate interpretation of the underlying rhythm.

Algorithm development. We developed a convolutional DNN to detect arrhythmias (Extended Data Fig. 1), which takes as input the raw ECG data (sampled at 200 Hz, or 200 samples per second) and outputs one prediction every 256 samples (or every 1.28 s), which we call the output interval. The network takes as input only the raw ECG samples and no other patient- or ECG-related features. The network architecture has 34 layers; to make the optimization of such a network tractable, we employed shortcut connections in a manner similar to the residual network architecture[41]. The network consists of 16 residual blocks with two convolutional layers per block. The convolutional layers have a filter width of 16 and $32*2^k$ filters, where k is a hyper-parameter which starts at 0 and is incremented by 1 every fourth residual block. Every alternate residual block subsamples its inputs by a factor of 2. Before each convolutional layer, we applied batch normalization[42] and a rectified linear activation, adopting the pre-activation block design[43]. The first and last layers of the network are special-cased due to this pre-activation block structure. We also applied Dropout[44] between the convolutional layers and after the nonlinearity with a probability of 0.2. The final fully connected softmax layer produces a distribution over the 12 output classes.

The network was trained de novo with random initialization of the weights as described by He et al.[9]. We used the Adam optimizer[45] with the default parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a mini batch size of 128. We initialized the learning rate to 1×10^{-3} and reduced it by a factor of 10 when the developmentally set loss stopped improving for two consecutive epochs. We chose the model that achieved the lowest error on the development dataset.

In general, the hyper-parameters of the network architecture and optimization algorithm were chosen via a combination of grid search and manual tuning. For the architecture, we searched primarily over the number of convolutional layers, the size and number of the convolutional filters, as well as the use of residual connections. We found the residual connections useful once the depth of the model exceeded eight layers. We also experimented with recurrent layers including long short-term memory cells[46] and bidirectional recurrence, but found no improvement in accuracy and a substantial increase in runtime; thus, we abandoned this class of models. We manually tuned the learning rate to achieve fastest convergence.

Algorithm evaluation. Since the DNN outputs one class prediction every output interval, it makes a series of 23 rhythm predictions for every 30-s record. The cardiologists annotated the start and end point of each rhythm class in the record. We used this to construct a cardiologist label at every output interval by rounding the annotation to the nearest interval boundary. Therefore, model accuracy can be assessed at the level of every output interval, which we call 'sequence-level', or at the record level, which we call 'set-level'. To compare model predictions at the sequence level, the model predictions at each output interval were compared with the corresponding committee consensus labels for that same output interval. At the set level, the set of unique rhythm classes across a given ECG record that was predicted by the DNN was compared with the set of rhythm classes annotated across the record by the committee consensus. The set-level evaluation, unlike the sequence-level, does not penalize for time misalignment of a rhythm classification within a record.

Algorithm evaluation at the sequence level allows comparison against the gold standard at every output interval, providing the most comprehensive metric of algorithm performance, which we therefore employ for most metrics. The sequence-level evaluation is also similar to clinical applications for telemetry or Holter monitor analysis, whereby it is critical to identify the onset and the offset of rhythms. Evaluation at the set level is a useful abstraction, approximating how the DNN algorithm might be applied to a single ECG record to identify which diagnoses are present in a given record.

To train and evaluate our model on the Physionet Challenge data, which contains variable length recordings, we made minor modifications to the DNN. Without any change, the DNN can accept as input any record with a length that is a multiple of 256 samples. To handle examples that are not a multiple of 256, records were truncated to the nearest multiple. We used the given record label as the label for approximately every 1.3-s output prediction. To produce a single prediction for the variable length record we used a majority vote of the sequence-level predictions.

Statistical analysis. We calculated the ROC analysis and AUC to assess model discrimination for each rhythm class with a one versus other strategy[28,47]. AUCs for sequence-level and set-level analyses are presented separately. We give a two-sided CI for the AUC scores[48]. Sensitivity and specificity were calculated at binary decision thresholds for every rhythm class. We computed the precision-recall curve, which shows the relationship between PPV (precision) and sensitivity (recall)[49]. It provides complementary information to the ROC curve, especially with class-imbalanced datasets. To compare the relative performance of the DNN to the cardiologist committee labels, we calculated the F_1 score, which is the harmonic mean of the PPV and sensitivity. It ranges from 0 to 1 and rewards algorithms that maximize both PPV and sensitivity simultaneously, rather than favoring one over the other. The F_1 score is complementary to the AUC, which is particularly helpful in the setting of multi-class prediction and less sensitive than the AUC in settings of class imbalance[49]. For an aggregate measure of model performance, we computed the class frequency-weighted arithmetic mean for both the F_1 score and the AUC. To obtain estimates of how the DNN compares to an average cardiologist, the characteristics of cardiologist performance were averaged across the six cardiologists who individually annotated each record. We used confusion matrices to illustrate the specific examples of rhythm classes where the DNN prediction or the individual cardiologist's prediction were discordant with the committee consensus at the sequence level. Among the individual cardiologist annotations in the test dataset, we calculated inter-annotator agreement as the ratio of the number of times two annotators agreed that a rhythm was present at each output interval and the total number of pairwise comparisons.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability. Code for the algorithm development, evaluation, and statistical analysis is open source with no restrictions and is available from https://github.com/awni/ecg.

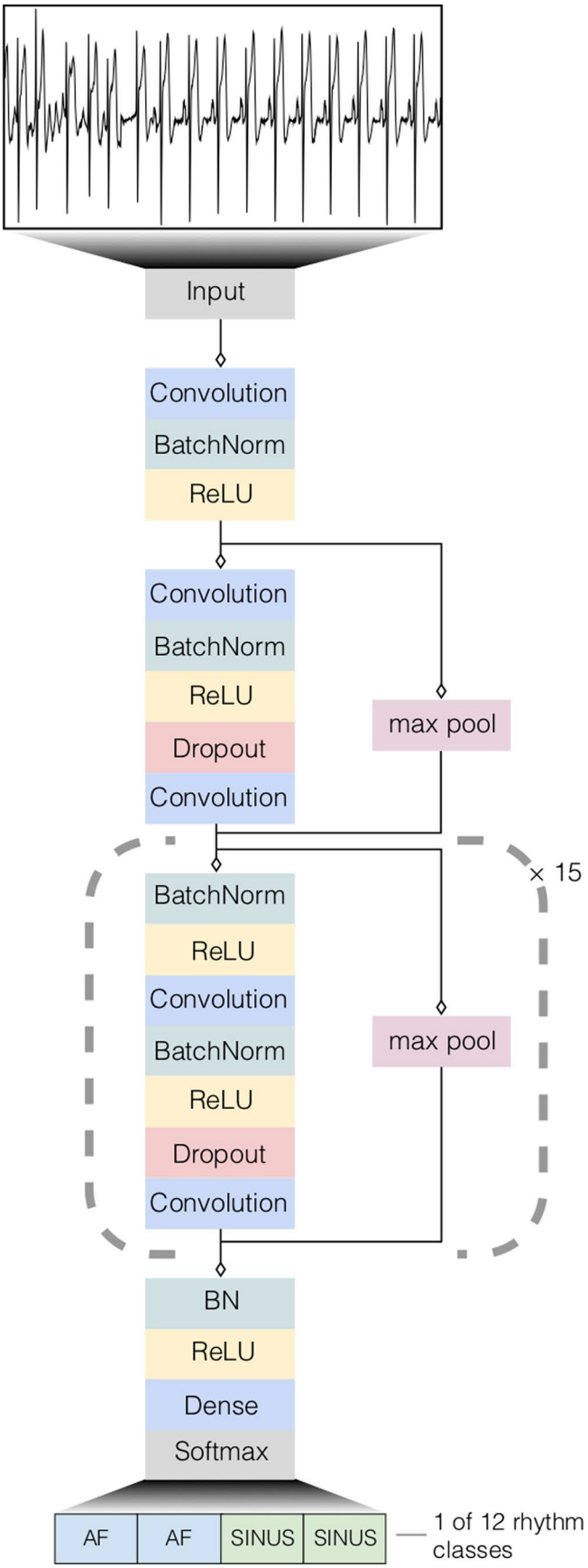
Data availability

The test dataset used to support the findings of this study is publicly available at https://irhythm.github.io/cardiol_test_set without restriction. Restrictions apply to the availability of the training dataset, which was used under license from iRhythm

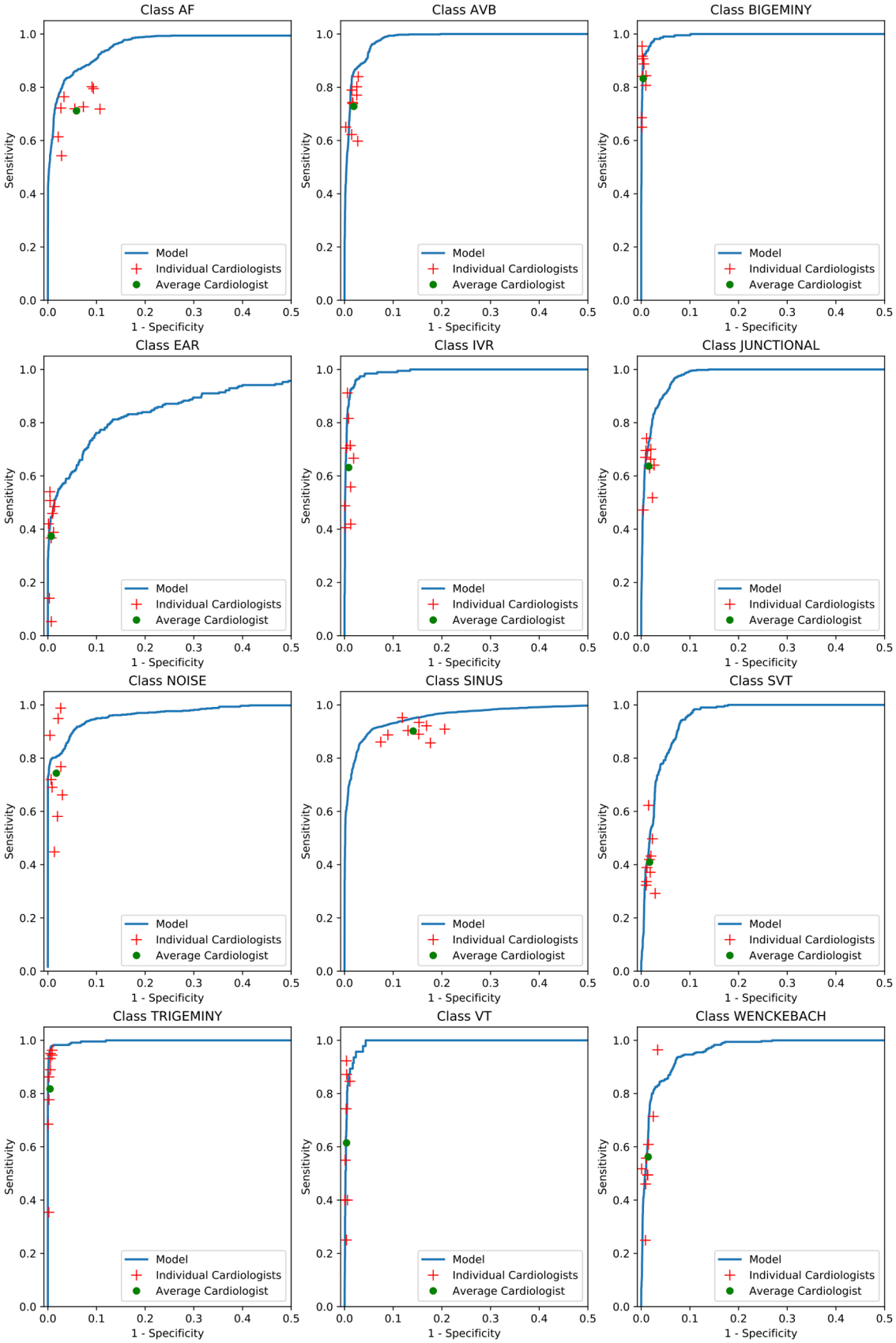
Technologies, Inc. for the current study. iRhythm Technologies, Inc. will consider requests to access the training data on an individual basis. Any data use will be restricted to noncommercial research purposes, and the data will only be made available on execution of appropriate data use agreements.

References

28. Hand, D. J. & Till, R. J. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Mach. Learn.* **45**, 171–186 (2001).
29. Smith, M. D. *et al.*in *Best Care at Lower Cost: the Path to Continuously Learning Health Care in America* (National Academies Press,: Washington, 2012).
30. Lyon, A., Mincholé, A., Martínez, J. P., Laguna, P. & Rodriguez, B. Computational techniques for ECG analysis and interpretation in light of their contribution to medical advances. *J. R. Soc. Interface* **15**, pii: 20170821 (2018).
31. Carrara, M. et al. Heart rate dynamics distinguish among atrial fibrillation, normal sinus rhythm and sinus rhythm with frequent ectopy. *Physiol. Meas.* **36**, 1873–1888 (2015).
32. Zhou, X., Ding, H., Ung, B., Pickwell-MacPherson, E. & Zhang, Y. Automatic online detection of atrial fibrillation based on symbolic dynamics and Shannon entropy. *Biomed. Eng. Online* **13**, 18 (2014).
33. Hong, S. et al. ENCASE: an ENsemble ClASsifiEr for ECG Classification using expert features and deep neural networks. *Comput. Cardiol.* https://doi.org/10.22489/CinC.2017.178-245 (2017).
34. Nahar, J., Imam, T., Tickle, K. S. & Chen, Y. P. Computational intelligence for heart disease diagnosis: a medical knowledge driven approach. *Expert Syst. Appl.* **40**, 96–104 (2013).
35. Cubanski, D., Cyganski, D., Antman, E. M. & Feldman, C. L. A neural network system for detection of atrial fibrillation in ambulatory electrocardiograms. *J. Cardiovasc. Electrophysiol.* **5**, 602–608 (1994).
36. Andreotti, F., Carr, O., Pimentel, M. A. F., Mahdi, A. & De Vos, M. Comparing feature-based classifiers and convolutional neural networks to detect arrhythmia from short segments of ECG. *Comput. Cardiol.* https://doi.org/10.22489/CinC.2017.360-239 (2017).
37. Xu, S. S., Mak, M. & Cheung, C. Towards end-to-end ECG classification with raw signal extraction and deep neural networks. *IEEE J. Biomed. Health Informatics* **14**, 1 (2018).
38. Ong, S. L., Ng, E. Y. K., Tan, R. S. & Acharya, U. R. Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats. *Comput. Biol. Med.* **102**, 278–287 (2018).
39. Shashikumar, S. P., Shah, A. J., Clifford, G. D. & Nemati, S. Detection of paroxysmal atrial fibrillation using attention-based bidirectional recurrent neural networks. In *Proc. 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,* 715–723 (Association for Computing Machinery, 2018).
40. Xia, Y., Wulan, N., Wang, K. & Zhang, H. Detecting atrial fibrillation by deep convolutional neural networks. *Comput. Biol. Med.* **93**, 84–92 (2018).
41. He, K., Zhang, X., Ren, S. & Sun, J. Identity mappings in deep residual networks. In *Proc. European Conference on Computer Vision,* 630–645 (Springer, 2016).
42. Ioffe, S. & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proc. International Conference on Machine Learning*, 448–456 (2015).
43. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition,* 770–778 (IEEE, 2016).
44. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
45. Kingma, D. P. & Ba, J. L. Adam: a method for stochastic optimization. In *Proc. International Conference on Learning Representations* 1–15 (2015).
46. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
47. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**, 861–874 (2006).
48. Hanley, J. A. & McNeil, B. J. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* **148**, 839–843 (1983).
49. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **10**, e0118432 (2015).



Extended Data Fig. 1 | Deep Neural Network architecture. Our deep neural network consisted of 33 convolutional layers followed by a linear output layer into a softmax. The network accepts raw ECG data as input (sampled at 200 Hz, or 200 samples per second), and outputs a prediction of one out of 12 possible rhythm classes every 256 input samples.



Extended Data Fig. 2 | Receiver operating characteristic curves for deep neural network predictions on 12 rhythm classes. Individual cardiologist performance is indicated by the red crosses and averaged cardiologist performance is indicated by the green dot. The line represents the ROC curve of model performance. AF-atrial fibrillation/atrial flutter; AVB- atrioventricular block; EAR-ectopic atrial rhythm; IVR-idioventricular rhythm; SVT-supraventricular tachycardia; VT-ventricular tachycardia. $n = 7{,}544$ where each of the 328 30-second ECGs received 23 sequence-level predictions.

nature research

Corresponding author(s): Awni Hannun

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	We used a small amount of custom software to build the data extraction and annotation tool. This software is not a major part of this work. The bulk of the custom and existing software we used is in the Data analysis section. The custom software we used for data collection was written in Python. It targeted records containing different rhythm classes, as described in the manuscript. Upon extraction, the data was saved in a PostgreSQL database. The extracted data was then pulled in a web-based tool to be reviewed. An html-based tool was designed for the purpose of accurately reviewing each record. Reviewers were able to see and scroll through 30-second ECG records and color code existing rhythms in the record from the onset to their offset. The full segmentation of the input record was then pushed to a different table in the same database. The reviewed records were finally de-identified and saved in json format to be used for model training purposes.
Data analysis	Code for the algorithm development, evaluation and statistical analysis is open source with no restrictions and available at https://github.com/awni/ecg.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The test dataset used to support the findings of this study is publicly available at https://irhythm.github.io/cardiol_test_set without restriction. Restrictions apply to the availability of the training dataset, which was used under license from iRhythm Technologies Inc. for the current study. iRhythm will consider requests to access the training data on an individual basis. Any data use will be restricted to non-commercial research purposes, and the data will only be made available upon execution of appropriate data use agreements.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Behavioural & social sciences

Study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We developed and validated the performance of a Deep Neural Network on ambulatory single-lead ECG. The data in the study is quantitative consisting of ECG records and their corresponding annotation.
Research sample	The dataset was a deidentified, retrospective dataset of adult patients >18 years of age who have used the iRhythm Zio monitor for clinical indications.
Sampling strategy	Records were chosen randomly from within the study period, though abnormal rhythms were intentionally over-sampled to provide more training examples for these rhythms. For the training set the sample size was chosen such that our model matched the performance of certified ECG technicians on a validation dataset. For the test set our sample size was chosen so that we would obtain roughly 20 (or more) examples for each rhythm class. We justified this sample size post-hoc with confidence intervals in our AUC computations.
Data collection	We extracted a median of 1 (and a maximum of 3) 30-second records per patient to construct the training dataset. To improve the balance of classes in the training dataset, records that exhibited less prevalent rhythms were intentionally oversampled, to a maximum of three 30-second records per patient. For the test dataset, 30-second records of each rhythm were randomly sampled from patients during the study period to achieve an equal number of records per class.
Timing	The data was collected retrospectively from a cohort who used the Zio monitor between January 2013 and March 2017
Data exclusions	We excluded patients under the age of 18 from the study. We preestablished this exclusion criteria in order to simplify any potential data use approval processes regarding the use of data from minors.
Non-participation	We did not require informed consent for this study given that the data belongs to iRhythm Technologies and that the data was fully de-identified.
Randomization	Our data was selected from the study period according to the selection strategy mentioned in the "Data Collection" section. Other than oversampling for certain arrhythmias all patients were randomly selected from pool of patients available in the study period.