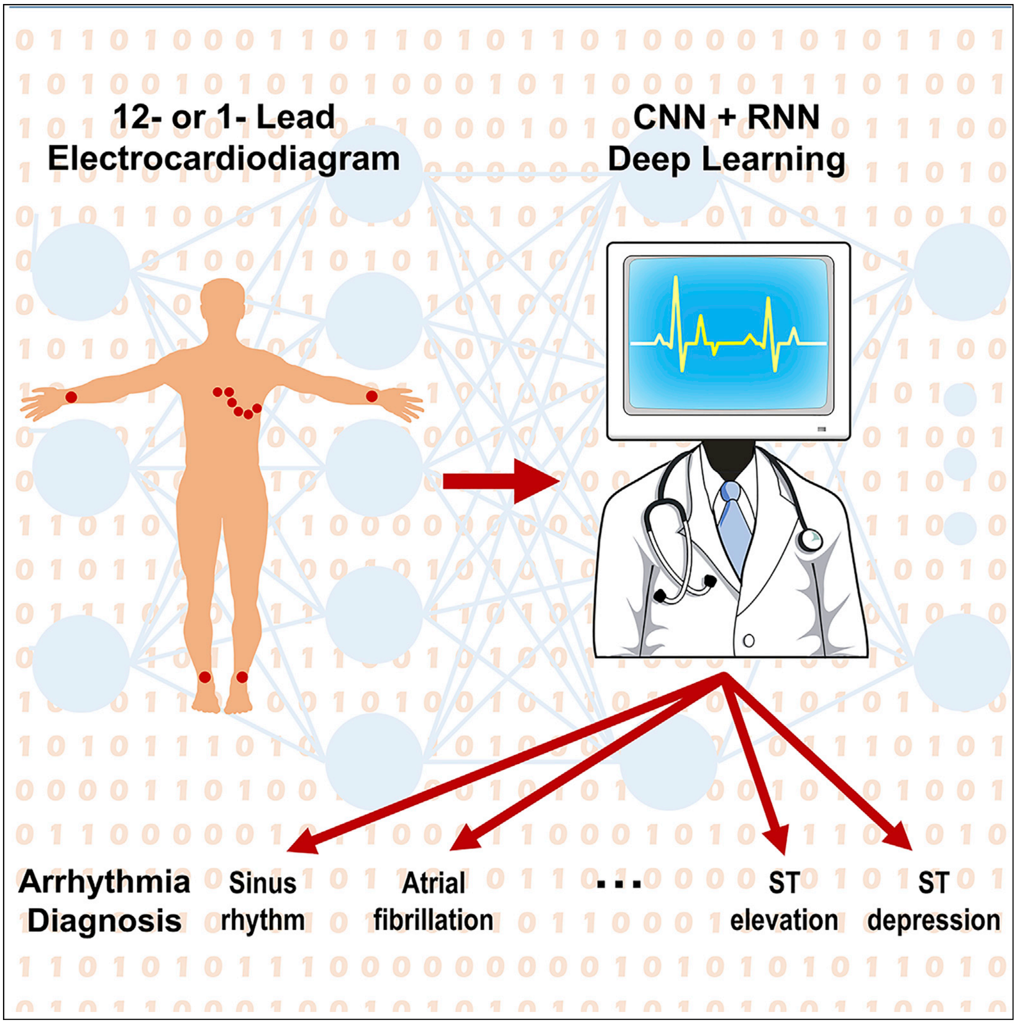
Article

Detection and Classification of Cardiac Arrhythmias by a Challenge-Best Deep Learning Neural Network Model



Tsai-Min Chen, Chih-Han Huang, Edward S.C. Shih, Yu-Feng Hu, Ming-Jing Hwang

mjhwang@ibms.sinica.edu.tw

HIGHLIGHTS

Accurate AI diagnosis of cardiac arrhythmia on ECG data from 11 hospitals

Capable of diagnosing concurrent cardiac arrhythmias

An ensemble model combining 12- and 1-lead models ranked first in CPSC2018

aVR and V1 found to be the best-performing single leads

Chen et al., iScience *23*, 100886
March 27, 2020 © 2020 The Author(s).
https://doi.org/10.1016/j.isci.2020.100886

iScience

CellPress

Article

Detection and Classification of Cardiac Arrhythmias by a Challenge-Best Deep Learning Neural Network Model

Tsai-Min Chen,[1,2,6] Chih-Han Huang,[1,3,6] Edward S.C. Shih,[1,6] Yu-Feng Hu,[1,4,5] and Ming-Jing Hwang[1,3,7,*]

[1]Institute of Biomedical Sciences, Academia Sinica, Taipei 11529, Taiwan

[2]Taiwan AI Academy, Science and Technology Ecosystem Development Foundation, New Taipei City 24158, Taiwan

[3]Genome and Systems Biology Degree Program, Academia Sinica and National Taiwan University, Taipei 10617, Taiwan

[4]Division of Cardiology, Department of Medicine, Taipei Veterans General Hospital, Taipei 11217, Taiwan

[5]Institute of Clinical Medicine and Cardiovascular Research Institute, National Yang-Ming University, Taipei 11221, Taiwan

[6]These authors contributed equally

[7]Lead Contact

*Correspondence: mjhwang@ibms.sinica.edu.tw

https://doi.org/10.1016/j.isci.2020.100886

SUMMARY

Electrocardiograms (ECGs) are widely used to clinically detect cardiac arrhythmias (CAs). They are also being used to develop computer-assisted methods for heart disease diagnosis. We have developed a convolution neural network model to detect and classify CAs, using a large 12-lead ECG dataset (6,877 recordings) provided by the China Physiological Signal Challenge (CPSC) 2018. Our model, which was ranked first in the challenge competition, achieved a median overall F1-score of 0.84 for the nine-type CA classification of CPSC2018's hidden test set of 2,954 ECG recordings. Further analysis showed that concurrent CAs were adequately predictive for 476 patients with multiple types of CA diagnoses in the dataset. Using only single-lead data yielded a performance that was only slightly worse than using the full 12-lead data, with leads aVR and V1 being the most prominent. We extensively consider these results in the context of their agreement with and relevance to clinical observations.

INTRODUCTION

Cardiac arrhythmias (CAs) are harbingers of cardiovascular diseases and the potential associated mortality (Kibos et al., 2013). CAs are usually diagnosed from electrocardiograms (ECGs), a noninvasive, inexpensive, and widely used clinical method for monitoring heart function. The diagnosis of CAs is based on wave-like features, such as the P wave, QRS wave, and T wave, of ECGs. A complete ECG usually contains recordings from six limb leads (I, II, III, aVR, aVL, aVF) and six chest leads (V1, V2, V3, V4, V5, V6), with each lead measuring electrical activity from a different angle of the heart, covering both the vertical plane (limb leads) and the horizontal plane (chest leads) (Malmivuo et al., 1995; Wilson et al., 1954).

The different leads exhibit distinct features of ECG signals that are associated with specific types of CA. For example, atrial fibrillation (AF) is characterized by the fibrillatory atrial waves and irregular conduction of QRS (Bayes de Luna et al., 1988; Platonov et al., 2012). Left bundle branch block (LBBB) is diagnosed by the distinct QRS morphology at leads I, aVL, V1, V2, V5, and V6, whereas right bundle branch block (RBBB) is diagnosed by the rsR′ pattern at V1 and V2 (Surawicz et al., 2009). First-degree atrioventricular block (I-AVB) is defined as constant PR intervals longer than 0.2 s (Wesley, 2016). The premature atrial contraction (PAC) and premature ventricular contraction (PVC) indicate the electrical impulse from an abnormal site; specifically, the P wave or QRS morphology of PAC and PVC differs from that in normal heart beats (Garcia and Miller, 2004; Kobayashi, 2018). ST segment is abnormal if either ST-segment elevation (STE) or ST-segment depression (STD) is greater than 0.1 mV (Hanna and Glancy, 2011).

To reliably recognize these complex CA-associated ECG characteristics, considerable training is required. Indeed, studies have shown that internists or cardiologists sometimes misdiagnose CA types (Hannun et al., 2019; Shiyovich et al., 2010). The significant growth of ECG examination, which increases physicians' workload, exacerbates the problem. This problem might be alleviated by developing computer algorithms that produce accurate and automatic diagnosis to assist the physicians. Although such a task would be difficult owing to the large variance in the geometrical and physiological features of ECG signals (Hoekema et al., 2001), significant progress has been made, especially in recent years (Lyon et al., 2018).

Two general approaches are available for developing an automatic CA diagnostic tool. The first splits ECG signals into units of the heartbeat, or cycles of the characteristic ECG waveforms. Thus, even with a small number of subjects, this beat-based approach can generate a large amount of data for machine learning to train predictive classification models. However, extracting ECG morphological features to delineate ECG signals is challenging because it is often an imprecise undertaking (Lyon et al., 2018). Although prediction accuracies as high as >99% have been reported in beat-based studies (Lyon et al., 2018), they could be confounded by both training and test beats coming from the same individuals. This issue is illustrated by a study in which test beats were taken from patients who were not included in the training set; the cross-validation accuracy of classification for six types of CA decreased from 99.7% to 81.5% (Qin et al., 2017).

The MIT-BIH Arrhythmia Database (MIT-BIH AD) (Goldberger et al., 2000; Moody and Mark, 2001) and the UCI Machine Learning Repository: Arrhythmia Data Set (UCIAD) (Guvenir et al., 1997), which, respectively, contain only 48 and 452 subjects, have been the source of publicly available ECG data for most previous CA prediction studies. However, databases with such small numbers of subjects can cause over-fitting problems for classification, especially for neural network algorithms (Begg, 2006). Data over-fitting can also arise from significantly unbalanced data, as with one or a few CA types being over-represented among cases. These problems can produce biased results from analyses of MIT-BIH AD and UCIAD (Mustaqeem et al., 2018; Nayak et al., 2016). For instance, in a study using UCIAD, a high accuracy (92%) of CA classification was achieved when 80% of the data were used for the training set and the remaining 20% for the test set, but the accuracy dropped to only 60% when the training-test split was 50-50 (Mustaqeem et al., 2018). Additional drawbacks are that ECG data in MIT-BIH AD only include two leads (e.g., leads II and V1, II and V5, II and V4, and V2 and V4), whereas the UCIAD only has extracted features available (average width of Q, amplitude of Q, etc.), not the raw data from 12-lead ECGs.

The second approach provides an end-to-end solution, avoiding the main difficulty of the beat-based approach. This approach requires a very large ECG database as well as the construction of a suitable deep learning artificial neural network to take advantage of it. Developments in both factors in recent years have made the second approach increasingly attractive. For example, to promote open-source research, the PhysioNet/Computing in Cardiology Challenge 2017 (CinC2017) released single-lead (lead I) ECG data from 8,528 subjects with four types of heart rhythms (AF, normal, other rhythms, noise) to the public (Clifford et al., 2017). Using convolutional neural network (CNN) plus three layers of long short-term memory (LSTM, one kind of recurrent neural network [RNN]), Xiong et al. (2018) produced the top performance in CinC2017 with an F1 score (the harmonic mean of the precision and recall) of 0.82 on its hidden test set (3,658 subjects).

Similar to CinC2017, the China Physiological Signal Challenge 2018 (CPSC2018), hosted by the seventh International Conference on Biomedical Engineering and Biotechnology (Liu et al., 2018), released a large ECG database for free download and set aside a hidden test set to assess models submitted by challenge participants from around the world. In contrast to CinC2017, CPSC2018 used 12-lead ECG data and subjects were grouped according to normal heart rhythm and eight types of CA: AF, I-AVB, LBBB, RBBB, PAC, PVC, STD, and STE. This represents the largest 12-lead ECG database with the most labeled CA types in the public domain to date. Here, we report a deep learning artificial neural network modeling of the CPSC2018 ECG data, and the results that won the first place in the competition.

RESULTS

Construction of a CNN-Based Model of CAs

Figure 1 depicts the architecture of our model for automating recognition of the CAs labeled in the CPSC2018 dataset. The model and the data used are described in more detail in Transparent Methods of the Supplemental Information. Briefly, the model consists of five CNN blocks, with several other types of neural network layers appended to achieve optimal performance while reducing over-fitting. To derive the model, CPSC2018's open-source ECG dataset was randomly divided into 10 equal folds, with 8 of the 10 folds serving as the training set and each of the two remaining folds serving as the validation set and the test set, respectively. This 10-fold cross-validation procedure of machine learning was repeated to produce hundreds of trained models, and the models producing the best validation results were selected for further evaluation on the test folds. For each model, a subject, identifiable by a unique ID number, would appear only once, exclusively in one of the three subsets (training, validation, or test). Single-lead models with the same architecture were similarly derived using this procedure. An ensemble model combining the best validation models of both 12-lead models and single-lead models was submitted to compete in the CPSC2018 challenge.

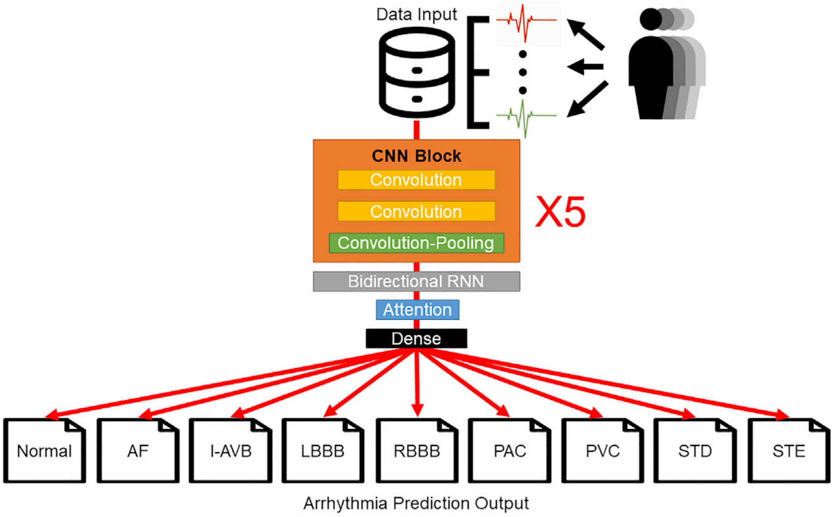


Figure 1. The Architecture of Deep Learning Artificial Neural Network for 12-Lead ECG CA Detection and Classification

Layers and blocks are specified in rectangle boxes; ''X5'' indicates that five CNN blocks are tandem connected before connecting to the bidirectional RNN layer, which is a gated recurrent unit layer. The output layer at the bottom contains the probabilities predicted by the model for each of the nine types of the CA classification. The type with the highest probability is the type predicted by the model for the input ECG recording.

Best Validation Models on 10-Fold Tests and Ensemble Model on Hidden Test

In Table 1, for each CA type the median accuracy, AUC (area under the receiver operating characteristic curve), and F1-score for the ten 10-fold tests from the best validation models are compared with those of the ensemble model, as well as with the F1-score of the ensemble model on the hidden test set of CPSC2018. The comparisons show that the ensemble model performed somewhat better than the best validation models, which is expected because the former combined and optimized the latter to produce the best 10-fold test results (see Methods). In addition, the ensemble model's performance was quite stable across all CA types, from the publicly available data to the hidden test data, reflecting the fairly similar compositions of the two sets of data, as mentioned in Transparent Methods.

Table 1 also reveals differential difficulties in predicting CA types. Namely, the prediction accuracy decreased from AF, bundle branch blocks, and premature contractions to ST abnormalities, with the normal type being one of the more difficult-to-predict types. The model's prediction for STE had the lowest F1-score (0.5–0.6), which may be due in part to physicians' variable opinions on how to diagnose STE (McCabe et al., 2013). The same trend, including the prediction of the normal type, was observed in all other top-performing models of CPSC2018 (Table S1). Indeed, almost all the top models produced very high F1-scores (>0.9) for AF and bundle branch blocks. Our model had significantly better predictions than the other models for several CA types, especially PAC, PVC, STD, and STE. This outcome explains how we outperformed others (Table S1). However, it should be noted that all top models performed well (overall F1-score > 0.8), and the difference between our model and the second-place model was minimal (Table S1).

Concurrent CA Types

One reason that models perform less accurately for certain CA types is that multiple CA types are predicted with almost equal probabilities for some patients. Figure 2 displays the probabilities output by the best validation models for ECG subjects when they were in the test fold of the 10-fold tests. As may be seen, normal, STD, and STE lack a probability score that can make them stand out from the other eight types, which is consistent with the model's performance results presented in Table 1. Further analysis on model probabilities showed that, for many patients with AF, a common concurrent CA was RBBB, whereas RBBB was often concurrent with PAC and PVC, in addition to AF (Figure 2). These probability results of concurrent CAs agreed well with the statistics for the 476 multi-labeled subjects; specifically, the three

	Best Validation Models			Ensemble Model			
CA Typ	Median Accuracy	Median AUC (95% CI)	Median F1-Score	Median Accuracy	Median AUC (95% CI)	Median F1-Score	Hidden Set F1-Score
Normal	0.940	0.890 (0.810–0.942)	0.795	0.949	0.867 (0.832–0.973)	0.808	**0.801**
AF	0.969	0.928 (0.902–0.985)	0.897	0.983	0.963 (0.914–0.993)	0.944	**0.933**
I-AVB	0.972	0.899 (0.864–0.988)	0.865	0.977	0.950 (0.875–0.990)	0.899	**0.875**
LBBB	0.990	0.914 (0.748–1.000)	0.821	0.995	0.942 (0.763–1.000)	0.899	**0.884**
RBBB	0.955	0.956 (0.887–0.988)	0.911	0.952	0.946 (0.871–0.976)	0.903	**0.910**
PAC	0.957	0.867 (0.749–0.955)	0.734	0.963	0.920 (0.779–0.981)	0.797	**0.826**
PVC	0.970	0.928 (0.841–0.988)	0.852	0.977	0.932 (0.864–0.996)	0.874	**0.869**
STD	0.951	0.878 (0.797–0.972)	0.788	0.959	0.906 (0.815–0.970)	0.834	**0.811**
STE	0.976	0.707 (0.558–0.995)	0.509	0.977	0.773 (0.603–0.993)	0.600	**0.624**

Table 1. Comparison of Model Performances on Tests

Results are from the best validation models and the ensemble model on the ten 10-fold tests, except for those in the last column (boldfaced), which are the ensemble model's median F1-scores for the hidden test set of CPSC2018 reported at its website http://2018.icbeb.org/Challenge.html, which did not provide accuracy or AUC results.

most multi-labeled incidences in these subjects were AF/RBBB, RBBB/PAC, and RBBB/PVC (Table 2). An ensemble model without these 476 multi-labeled subjects being added back to the training set (see Methods) performed well in predicting these multiple types of CA (Tables S3 and S4), indicating the model's ability to capture ECG features of concurrent CAs. These results are also generally compatible with clinical observations that rate-dependent (phase 3) block during ectopic atrial beats or AF can lead to RBBB (Nielsen et al., 2010; Gertsch, 2016). However, a larger dataset of multi-labeled subjects is required to fully evaluate our model's performance on concurrent CA diagnoses.

Model Performances with Single-Lead Data

The median F1-scores for models for a single lead in the 10-fold tests are presented in Figure 3. The performances for the best validation models using the 12-lead data in Table 1 were largely replicated by those using only single-lead data. In most cases, only minimal changes of F1-scores for the classification of individual CA types were noted between the analyses of 12-lead and single-lead ECGs. The results also indicate aVR was one of the best-performing single leads, with its performance ranking first in the overall average and the three individual CA types (normal, AF and STD), as well as within the top three for all CA types except STE and PAC. Another well-performing single lead is lead V1, which ranked first in three types (I-AVB, RBBB, and PAC) but did worse than most other leads for other types. In comparison, lead I, which was used by Apple Watch (Apple, 2019), was not as remarkable in our tests. Lead II, which is favored among the 12 leads by physicians for a quick impression of an ECG recording due to its clearest signal (Beebe and Myers, 2012), ranked fifth in the overall average but was statistically no different from the best performing leads (p value of paired t test < 0.05). These results are largely supported by a Bayes factor analysis (Goodman, 1999) that rigorously assessed statistical differences between these leads (see Tables S5–S7).

These performance rankings suggest that the current model identified the lead-specific morphology of the various CA types. For examples, the deep and broad S-waves in lead V1 and the broad clumsy R-waves in V6 have been used for the diagnosis of LBBB (Podrid et al., 2015), and V1 and V6 were identified as having the leading performance among the single leads. Meanwhile, the diagnostic criteria of RBBB included the rSR′ pattern in leads V1 and V2 (Chugh, 2014), which were also selected as top-performing single leads.

DISCUSSION

In recent years, deep learning of artificial intelligence (AI) has been successfully used to make medical diagnoses (Esteva et al., 2019). The present work for CA detection and classification is related to the

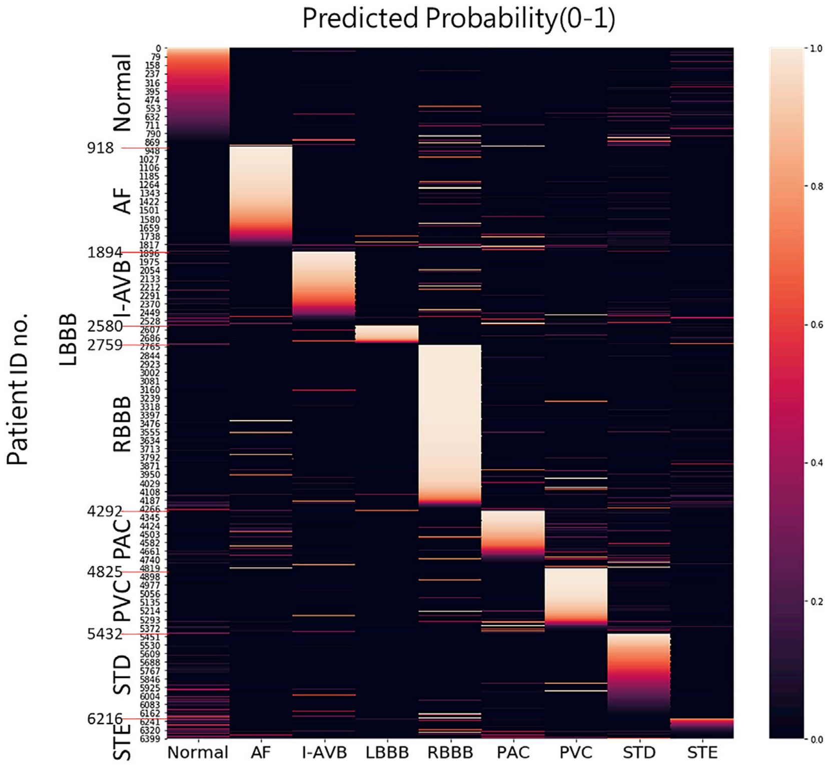


Figure 2. Probabilities Output by the Best Validation Models in the Test Folds of the 10-Fold Tests

On the right is the color-coded probability scale.

competition in CinC2017 (Clifford et al., 2017) and recently published studies (Clifford et al., 2017; Hannun et al., 2019). A direct comparison of performance between different studies is difficult because not all of them used publicly available ECG data and different CA types and type numbers were predicted. The complexity of these deep learning models also differed; for example, our model had a total of 18 neural network layers, compared with 33 (Hannun et al., 2019) and 5–7 (Clifford et al., 2017) in others.

Our model can be best compared with one reported very recently (Yao et al., 2020). The model of that study used an architecture that is quite similar to ours, albeit with some significant differences in detail as discussed in Transparent Methods, and its performance on the same CPSC2018 hidden test set was reported. As the comparisons in Table S8 showed, Yao et al.'s model achieved a better F1-score than our best validation models for most of the CA types, but the reverse was true when it was compared with our ensemble model. This reinforces the notion that an ensemble model can often achieve an even better performance than those achievable by individual models alone.

Interestingly, all these recent studies achieved an overall F1-score of 0.82–0.84 in predicting CA types. Although not fully tested in real-world scenarios, AI-based ECG diagnosis of CAs has been shown to significantly improve diagnosis accuracy, compared with general physicians and cardiologists (Hannun et al., 2019; Shiyovich et al., 2010) (also see Table S2 for a very small sampling). Therefore, these AI models are capable of reducing erroneous diagnoses and medical overload. Although this outcome is very encouraging, it is sobering to remember that until most of the ''ground truth'' CA diagnoses used to derive AI models are made by expert cardiologists, further improvement of model accuracy may be limited.

Our analyses suggest that the models built on single-lead information could predict CA types with minimal differences in performance from those based on 12 leads. The clinical diagnostic criteria of CA types are often lead specific. The top-ranking single lead for RBBB or LBBB in our model was compatible with the leads in the diagnostic criteria for RBBB and LBBB (Surawicz et al., 2009), solidifying the validity of the

	AF	I-AVB	LBBB	RBBB	PAC	PVC	STD	STE
AF	0	0	29	**172**	4	8	33	2
I-AVB		0	8	10	3	5	6	4
LBBB			0	0	10	6	3	4
RBBB				0	**55**	**51**	20	19
PAC					2	3	6	5
PVC						0	18	2
STD							0	2
STE								0

Table 2. Label Count Statistics of the 476 Multi-Labeled Subjects in the Released CPSC2018 Dataset

Only the upper triangle portion of the symmetrical concurrent CA label counts is shown. The three largest counts are boldfaced.

present AI diagnostic model. The performance of aVR, a lead that is often clinically ignored, in our AI model is intriguing and deserves attention. The leads I, II, and V1 are conventionally used as the modified leads in continuous monitoring or mobile devices for ECG (Apple, 2019; Brunner et al., 2010). In our AI model, aVR could predict several CA types with better performance than the conventional leads. Rather than considering reciprocal information from the left lateral side, the purpose of lead aVR is to obtain specific information from the right upper side of the heart, including the outflow tract of the right ventricle and the basal part of the septum. The vector of lead aVR is parallel to the anatomical and corresponding electrical axis from atrial base to ventricular apex, and thus it may maximize the electrical signals of atrial and ventricular depolarization. These factors may give the unique role of aVR to diagnose CAs with a potential mechanism to outperform the other leads. In comparison, lead I, which is used in Apple Watch for AF detection (Krueger, 2018), did not perform as well in our analysis. Our results suggest that the best predictive single lead for different CA types could be different for clinical applications. Our results may provide an impetus for future studies to investigate the potential use of lead aVR in different CA types and ECG devices (wearable or portable).

CAs are complex and concurrent CA types are not uncommon, especially for those that are related in cardiac electrophysiology. Although ECG-based CA diagnostic models have so far focused only on single-type predictions, our analysis shows that AI is capable of multi-type CA diagnosis. Detection and classification of concurrent CAs should be a subject for future studies, and our model is a first step in that direction. ECGs have been shown to be capable of disease/health detection beyond CAs, including, for example, the prediction of asymptomatic left ventricular dysfunction (Attia et al., 2019) and non-invasive potassium tracking (Attia et al., 2016). As methods of AI machine learning continue to advance and become friendlier for non-AI specialists to employ, we can expect ECGs to be explored for their diagnostic power in many more diseases and clinical applications.

Limitations of the Study

Although our model produced state-of-the-art accuracy on the hidden test set of CPSC2018, which was collected from 11 hospitals in China, it will most likely still need further refinement to achieve the same level of performance for other ECG datasets. However, a recent study (Hannun et al., 2019) has shown that this will probably only require transfer learning, that is, keeping the same model architecture while retraining network weights for the new data. Additionally, other deep learning architectures may exist that can achieve even better CA prediction accuracy and, as the comparisons with Yao et al.'s study (Yao et al., 2020) showed, even within a similar network architecture, different models (e.g., using different number of trainable parameters) exist to achieve similar performances. Furthermore, although different CA types might be better modeled by using information from different single leads, it remains to be studied whether different methods, or different network architectures, should be used for different single leads. In the absence of a systematic evaluation approach, as well as the lack of a truly gold standard training set, as alluded to above, these limitations are difficult to address at present.

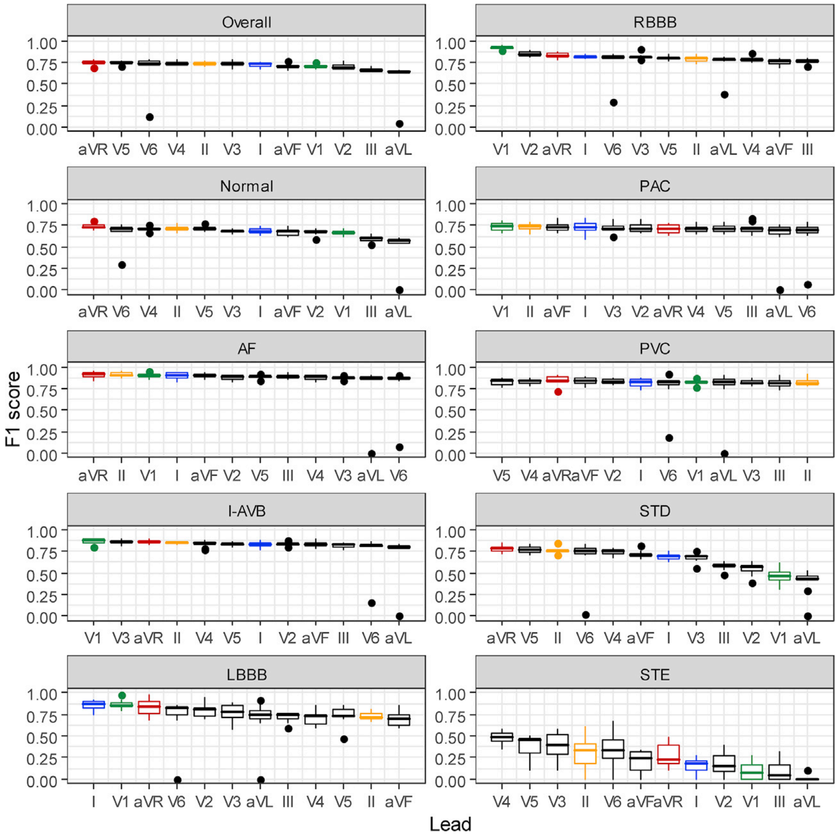


Figure 3. The Ranked F1-Score Results of Single-Lead Models

The F1-scores (on the y axis) are from the single-lead models performed on the 10-fold tests (see Methods). Lead aVR is shown in red, V1 in green, I in blue, and II in orange.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

DATA AND CODE AVAILABILITY

The CPSC2018 ECG data and the code of our model (entry no. CPSC0236) are available at http://2018.icbeb.org/Challenge.html.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.100886.

ACKNOWLEDGMENTS

We thankfully acknowledge the CPSC Challenge Chair Prof. Chengyu Liu and his conference coworkers for their help and for making the ECG data publicly available. The research of the Hwang laboratory is supported by Institute of Biomedical Sciences, Academia Sinica, and Ministry of Science and Technology, Taiwan, grant number MOST108-2311-B-001-017.

AUTHOR CONTRIBUTIONS

Conceptualization, T.-M.C., C.-H.H., E.S.C.S., and M.-J.H.; Methodology, T.-M.C., C.-H.H., E.S.C.S., and M.-J.H.; Investigation, T.-M.C., C.-H.H., and E.S.C.S.; Formal Analysis, T.-M.C., C.-H.H., and E.S.C.S.; Writing – Original Draft, T.-M.C., C.-H.H., E.S.C.S.; Writing – Review & Editing, Y.-F.H. and M.-J.H.; Resources, Y.-F.H. and M.-J.H.; Visualization, T.-M.C., C.-H.H., and E.S.C.S.; Supervision, M.-J.H.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 30, 2019
Revised: January 15, 2020
Accepted: January 30, 2020
Published: March 27, 2020

REFERENCES

Apple. (2019). Taking an ECG with the ECG App on Apple Watch Series 4. https://support.apple.com/en-us/HT208955.

Attia, Z.I., DeSimone, C.V., Dillon, J.J., Sapir, Y., Somers, V.K., Dugan, J.L., Bruce, C.J., Ackerman, M.J., Asirvatham, S.J., and Striemer, B.L. (2016). Novel bloodless potassium determination using a signal-processed single-lead ECG. J. Am. Heart Assoc. *5*, e002746.

Attia, Z.I., Kapa, S., Lopez-Jimenez, F., McKie, P.M., Ladewig, D.J., Satam, G., Pellikka, P.A., Enriquez-Sarano, M., Noseworthy, P.A., and Munger, T.M. (2019). Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nat. Med. *25*, 70.

Bayes de Luna, A., Cladellas, M., Oter, R., Torner, P., Guindo, J., Marti, V., Rivera, I., and Iturralde, P. (1988). Interatrial conduction block and retrograde activation of the left atrium and paroxysmal supraventricular tachyarrhythmia. Eur. Heart J. *9*, 1112–1118.

Beebe, R., and Myers, J. (2012). Professional Paramedic, Volume I: Foundations of Paramedic Care (Cengage Learning).

Begg, R. (2006). Neural Networks in Healthcare: Potential and Challenges: Potential and Challenges (IGI Global).

Brunner, L.S., Smeltzer, S.C.O.C., Bare, B.G., Hinkle, J.L., and Cheever, K.H. (2010). Brunner & Suddarth's Textbook of Medical-Surgical Nursing (Wolters Kluwer Health/Lippincott Williams & Wilkins).

Chugh, S. (2014). Textbook of Clinical Electrocardiography (Jaypee Brothers).

Clifford, G.D., Liu, C., Moody, B., Lehman, L.-w.H., Silva, I., Li, Q., Johnson, A., and Mark, R.G. (2017). AF classification from a short single lead ECG recording: the Physionet Computing in Cardiology Challenge 2017. Proc. Comput. Cardiol. *44*, 1.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., and Dean, J. (2019). A guide to deep learning in healthcare. Nat. Med. *25*, 24.

Garcia, T., and Miller, G. (2004). Arrhythmia Recognition: The Art of Interpretation (Jones & Bartlett Learning).

Gertsch, M. (2016). The ECG Manual: An Evidence-Based Approach (Springer).

Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.-K., and Stanley, H.E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation *101*, e215–e220.

Goodman, S.N. (1999). Toward evidence-based medical statistics. 2: the Bayes factor. Ann. Intern. Med. *130*, 1005–1013.

Guvenir, H.A., Acar, B., Demiroz, G., and Cekin, A. (1997). A supervised machine learning algorithm for arrhythmia analysis. Comput. Cardiol. *24*, 433–436.

Hanna, E.B., and Glancy, D.L. (2011). ST-segment depression and T-wave inversion: classification, differential diagnosis, and caveats. Cleveland Clinic J. Med. *78*, 404.

Hannun, A.Y., Rajpurkar, P., Haghpanahi, M., Tison, G.H., Bourn, C., Turakhia, M.P., and Ng, A.Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat. Med. *25*, 65.

Hoekema, R., Uijen, G.J., and Van Oosterom, A. (2001). Geometrical aspects of the interindividual variability of multilead ECG recordings. IEEE Trans. Biomed. Eng. *48*, 551–559.

Kibos, A.S., Knight, B.P., Essebag, V., Fishberger, S.B., Slevin, M., and Ţintoiu, I.C. (2013). Cardiac Arrhythmias: From Basic Mechanism to State-of-the-Art Management (Springer).

Kobayashi, Y. (2018). Idiopathic ventricular premature contraction and ventricular tachycardia: distribution of the origin, diagnostic algorithm, and catheter ablation. J. Nippon Med. Sch. *85*, 87–94.

Krueger, A.C. (2018). FDA Document of Electrocardiograph Software for Over-the-counter Use. https://www.accessdata.fda.gov/cdrh_docs/pdf18/DEN180044.pdf.

Liu, F., Liu, C., Zhao, L., Zhang, X., Wu, X., Xu, X., Liu, Y., Ma, C., Wei, S., He, Z., et al. (2018). An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. J. Med. Imaging Health Inform. *8*, 1368–1373.

Lyon, A., Mincholé, A., Martínez, J.P., Laguna, P., and Rodriguez, B. (2018). Computational techniques for ECG analysis and interpretation in light of their contribution to medical advances. J. R. Soc. Interf. *15*, 20170821.

Malmivuo, P., Malmivuo, J., and Plonsey, R. (1995). Bioelectromagnetism: Principles and Applications of Bioelectric and Biomagnetic Fields (Oxford University Press).

McCabe, J.M., Armstrong, E.J., Ku, I., Kulkarni, A., Hoffmayer, K.S., Bhave, P.D., Waldo, S.W., Hsue, P., Stein, J.C., and Marcus, G.M. (2013). Physician accuracy in interpreting potential ST-segment elevation myocardial infarction electrocardiograms. J. Am. Heart Assoc. *2*, e000268.

Moody, G.B., and Mark, R.G. (2001). The impact of the MIT-BIH arrhythmia database. IEEE Eng. Med. Biol. Mag. *20*, 45–50.

Mustaqeem, A., Anwar, S.M., and Majid, M. (2018). Multiclass classification of cardiac arrhythmia using improved feature selection and SVM invariants. Comput. Math. Methods Med. *1*, 1–10.

Nayak, C.G., Seshikala, G., Desai, U., and Nayak, S.G. (2016). Identification of arrhythmia classes using machine-learning techniques. Int. J. Biol. Biomed. *1*, 48–53.

Nielsen, J.B., Olesen, M.S., Tangø, M., Haunsø, S., Holst, A.G., and Svendsen, J.H. (2010). Incomplete right bundle branch block: a novel electrocardiographic marker for lone atrial fibrillation. Europace *13*, 182–187.

Platonov, P.G., Cygankiewicz, I., Stridh, M., Holmqvist, F., Vazquez, R., Bayes-Genis, A., McNitt, S., Zareba, W., and de Luna, A.B. (2012). Low atrial fibrillatory rate is associated with poor outcome in patients with mild to moderate heart failure. Circ. Arrhythmia Electrophysiol. *5*, 77–83.

Podrid, P., Rajeev Malhotra, M.D.M.S., Kakkar, R., and Noseworthy, P.A. (2015). Podrid's Real-World ECGs: Volume 4B, Arrhythmias [Practice Cases]: A Master's Approach to the Art and Practice of Clinical ECG Interpretation (Cardiotext Publishing).

Qin, Q., Li, J., Zhang, L., Yue, Y., and Liu, C. (2017). Combining low-dimensional wavelet features and support vector machine for arrhythmia beat classification. Sci. Rep. *7*, 6067.

Shiyovich, A., Wolak, A., Yacobovich, L., Grosbard, A., and Katz, A. (2010). Accuracy of diagnosing atrial flutter and atrial fibrillation from a surface electrocardiogram by hospital physicians: analysis of data from internal medicine departments. Am. J. Med. Sci. *340*, 271–275.

Surawicz, B., Childers, R., Deal, B.J., and Gettes, L.S. (2009). AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: part III: intraventricular conduction disturbances a scientific statement from the American heart association electrocardiography and arrhythmias committee, council on clinical Cardiology; the American college of Cardiology foundation; and the heart rhythm society endorsed by the international society for computerized electrocardiology. J. Am. Coll. Cardiol. *53*, 976–981.

Wesley, K. (2016). Huszar's ECG and 12-Lead Interpretation - E-Book (Elsevier Health Sciences).

Wilson, F.N., Kossmann, C.E., Burch, G.E., Goldberger, E., Graybiel, A., Hecht, H.H., Johnston, F.D., Lepeschkin, E., and Myers, G.B. (1954). Recommendations for standardization of electrocardiographic and vectorcardiographic leads. Circulation *10*, 564–573.

Xiong, Z., Nash, M.P., Cheng, E., Fedorov, V.V., Stiles, M.K., and Zhao, J. (2018). ECG signal classification for the detection of cardiac arrhythmias using a convolutional recurrent neural network. Physiol. Meas. *39*, 094006.

Yao, Q., Wang, R., Fan, X., Liu, J., and Li, Y. (2020). Multi-class Arrhythmia detection from 12-lead varied-length ECG using attention-based time-incremental convolutional neural network. Inf. Fusion *53*, 174–182.

iScience, Volume *23*

Supplemental Information

Detection and Classification of Cardiac Arrhythmias by a Challenge-Best Deep Learning Neural Network Model

Tsai-Min Chen, Chih-Han Huang, Edward S.C. Shih, Yu-Feng Hu, and Ming-Jing Hwang

Table S1. CPSC2018's top 10 models and results (reported by the conference on http://2018.icbeb.org/Challenge.html) [a]. Related to Table 1.

Rank	Overall F1	F_{af}	F_{block}	F_{pc}	F_{st}
1	**0.837**	**0.933**	**0.899**	**0.847**	**0.779**
2	0.830	0.931	0.912	0.817	0.761
3	0.806	0.914	0.879	0.801	0.742
4	0.802	0.918	0.89	0.789	0.718
5	0.791	0.924	0.882	0.779	0.709
6	0.783	0.905	0.902	0.722	0.708
7	0.782	0.911	0.891	0.775	0.670
8	0.778	0.921	0.858	0.797	0.676
9	0.776	0.906	0.876	0.773	0.711
10	0.766	0.894	0.857	0.733	0.683

[a] Our team's results are ranked first (boldfaced), and the highest scores of each sub-competition are indicated in red. Overall F1 is the average of the F1 values from each classification type. Faf: F1 of AF; Fblock: F1 of I-AVB, LBBB and RBBB; Fpc: F1 of PAC and PVC; Fst: F1 of STD and STE.

Table S2. Comparisons of CA diagnosis between conference-assigned label, our model prediction, and consensus from three expert cardiologists. Related to Table 1.

PatientID	Conf.	Model	Cardiologist 1	Cardiologist 2	Cardiologist 3	Consensus
6268	STE	Normal	Normal	SR (normal)	SR (normal)	Normal
6215	STD	AF	Narrow-QRS tachycardia, SVT, RBBB, RAD	PSVT	PSVT	PSVT
4237	RBBB	I-AVB	I-AVB	I-AVB, SR, TWI (V1-V4), rSR' (V1)	I-AVB,SR	I-AVB
6380	STE	LBBB	LBBB	LBBB,SR,LAE	LBBB,SR	LBBB
1452	AF	RBBB	AF, MVR, RBBB	AF, RBBB, Q wave, STE with reciprocal change, w/o old MI	AF, RBBB	RBBB
5398	PVC	PAC	PAC	PAC	PAC,SR	PAC
5963	STD	PVC	PVC	PVC,SR,STD(V3-V6)	PVC,SR	PVC
4278	RBBB	STD	Normal	SR, minimal STTC(II, III, aVF)	STD-like, SR	Normal
504	Normal	STE	STE-like	SR, early repolarization	SR	Normal

Conference assignments (regarded as "ground truth" for model training) and our model predictions in agreement with the consensus of the three expert cardiologists are highlighted in red. SVT: Supraventricular tachycardia; RAD: Right Axis Deviation; MVR: mitral valve replacement; SR: sinus rhythm; PSVT: Paroxysmal supraventricular tachycardia; LAE: left atrial enlargement; MI: myocardial infarction; STTC: ST-T change

Table S3. The performances of models trained without multi-labeled data[a]. Related to Table 1.

	Best Validation Models			Ensemble Model		
CA Type	Median Accuracy	Median AUC (95% CI)	Median F1-score	Median Accuracy	Median AUC (95% CI)	Median F1-score
Normal	0.940	0.908 (0.791-0. 916)	0.807	0.937	0.901 (0. 808-932)	0.794
AF	0.974	0.949 (0.885-0.992)	0.915	0.980	0.955 (0.929-0.995)	0.935
I-AVB	0.973	0.918 (0.852-0.991)	0.876	0.976	0.912 (0.900-0.996)	0.879
LBBB	0.993	0.927 (0.748-1.000)	0.870	0.993	0.913 (0.770-1.000)	0.862
RBBB	0.961	0.954 (0.895-0.981)	0.922	0.944	0.925 (0.880-0.972)	0.885
PAC	0.957	0.852 (0.747-0.961)	0.747	0.965	0.889 (0.798-0.984)	0.796
PVC	0.973	0.926 (0.832-0.989)	0.858	0.974	0.950 (0.820-0.992)	0.870
STD	0.950	0.869 (0.800-0.966)	0.786	0.956	0.914 (0.790-0.950)	0.821
STE	0.974	0.667 (0.491-0.995)	0.394	0.975	0.663 (0.491-0.995)	0.444

[a] For these models, the 476 multi-labeled recordings were not included in the training set. These results are from the best validation models and the ensemble model for the ten 10-fold tests. These performances are comparable with those presented in Table 1.

Table S4. Successfully predicted CA types for the 476 multi-labeled subjects in the released dataset of CPSC2018[a]. Related to Table 2.

	AF	I-AVB	LBBB	RBBB	PAC	PVC	STD	STE
AF	0/0	0/0	17/29	**154/172**	0/4	6/8	4/33	0/2
I.AVB		0/0	2/8	8/10	0/3	2/5	0/6	0/4
LBBB			0/0	0/0	2/10	4/6	0/3	0/4
RBBB				0/0	**34/55**	**49/51**	16/20	5/19
PAC					0/0	3/3	4/6	2/5
PVC						0/0	6/18	0/2
STD							0/0	2/2
STE								0/0

[a] Only the upper triangle portion of the symmetrical concurrent CA label counts is shown. The numbers shown are the number of correctly predicted subjects/the total number of multi-labeled subjects for a given CA type. The two CA types with the highest and the second highest probabilities are the predicted concurrent CA types. Boldfaced are the three most common concurrent CA labels in these subjects (see Table 2).

Table S5. The Bayes factors[a] (in log scale) of each lead's performance (F1 score) relative to that of the best performing lead in each CA type (see Fig. 3). Related to Figure 3.

	I	II	III	aVR	aVL	aVF	V1	V2	V3	V4	V5	V6
Normal	3.4	0.3	14.1	-0.9	3.1	4.1	7.3	5.3	5.0	0.9	0.4	-0.1
AF	-0.8	-0.8	-0.6	-0.9	-0.3	-0.7	-0.9	0.2	1.1	0.8	-0.3	-0.2
I-AVB	1.2	0.1	3.7	-0.4	0.1	0.8	-0.9	1.3	-0.5	0.5	1.1	-0.1
LBBB	-0.9	6.1	6.4	-0.8	0.4	6.1	-0.9	0.2	1.0	3.8	1.9	-0.1
RBBB	18.3	15.5	22.2	11.4	4.1	18.8	-0.9	9.9	12.9	16.0	18.6	1.8
PAC	-0.9	-0.9	-0.6	-0.3	-0.1	-0.9	-0.9	-0.8	-0.7	-0.2	-0.4	-0.2
PVC	-0.8	-0.9	-0.8	-0.8	-0.6	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.6
STD	6.0	-0.9	15.4	-0.9	10.2	2.4	14.9	11.9	5.6	0.6	-0.8	-0.5
STE	11.9	1.7	11.3	3.7	23.1	6.7	12.2	7.1	0.3	-0.9	0.4	0.8

[a] Computed using the BayesFactor routine in the R package.

Table S6. Leads in the top-performing group (threshold: Bayes factor<3.0) [a]. Related to Figure 3.

	I	II	III	aVR	aVL	aVF	V1	V2	V3	V4	V5	V6
Normal	0	1	0	1	0	0	0	0	0	0	1	1
AF	1	1	1	1	1	1	1	1	0	0	1	1
I-AVB	0	1	0	1	1	0	1	0	1	0	0	1
LBBB	1	0	0	1	1	0	1	1	0	0	0	1
RBBB	0	0	0	0	0	0	1	0	0	0	0	0
PAC	1	1	1	1	1	1	1	1	1	1	1	1
PVC	1	1	1	1	1	1	1	1	1	1	1	1
STD	0	1	0	1	0	0	0	0	0	0	1	1
STE	0	0	0	0	0	0	0	0	1	1	1	0
Total	4	**6**	3	**7**	5	3	**6**	4	4	3	**6**	**7**

[a] The leads in the top-performing group, indicated by 1 (0 for those excluded from this group), for a given CA type are considered to perform equally well statistically based on the threshold of Bayes factor < 3.0, which indicates the null hypothesis of no difference from the best-performing lead holds. Based on this threshold, the sum total shows that leads aVR and V6 received the most top-performing group counts, followed by leads II, V1, and V5.

Table S7. Leads in the top-performing group (threshold: Bayes factor < 0.33) [a]. Related to Figure 3.

	I	II	III	aVR	aVL	aVF	V1	V2	V3	V4	V5	V6
Normal	0	0	0	1	0	0	0	0	0	0	0	0
AF	1	1	1	1	0	1	1	0	0	0	0	0
I-AVB	0	0	0	0	0	0	1	0	1	0	0	0
LBBB	1	0	0	1	0	0	1	0	0	0	0	0
RBBB	0	0	0	0	0	0	1	0	0	0	0	0
PAC	1	1	1	0	0	1	1	1	1	0	0	0
PVC	1	1	1	1	1	1	1	1	1	1	1	1
STD	0	1	0	1	0	0	0	0	0	0	1	1
STE	0	0	0	0	0	0	0	0	0	1	0	0
Total	4	4	3	**5**	1	3	**6**	2	3	2	2	2

[a] The leads in the top-performing group, indicated by 1 (0 for those excluded from this group), for a given CA type are considered to perform equally well statistically based on the threshold of Bayes factor < 0.03, which indicates the null hypothesis of no difference from the best-performing lead holds. Using this threshold, the sum total shows that lead V1 received most top-performing group counts, followed by lead aVR.

Table S8. The F1 scores of our best validation models and the ensemble model, and Yao et al.'s model.* Related to Table 1.

CA type	Best validation models	Yao et al.'s model (ATI-CNN)	Ensemble model
Normal	0.795	0.789	0.801
AF	0.897	0.920	0.933
I-AVB	0.865	0.850	0.875
LBBB	0.821	0.872	0.884
RBBB	0.911	0.933	0.910
PAC	0.734	0.736	0.826
PVC	0.852	0.861	0.869
STD	0.788	0.789	0.811
STE	0.509	0.556	0.624

*F1 scores better than Yao et al.'s model are shown in red. The F1 scores of our models have also appeared in Table 1. The F1 scores for Yao et al.'s model and the ensemble model were on CPSC2018's hidden test set, while the best validation models were tested on its publicly released data set.

Table S9. Numbers and sizes of hyperparameters and trainable parameters of our model. Related to Figure 1.

CNN block & other network layers	CNN layer	Kernel	Kernel size	Padding	Stride	Parameters*
1	1	12	3	1	1	12×(12×3+1)=444
	2	12	3	1	1	12×(12×3+1)=444
	3 (pooling)	12	24	1	2	12×(12×24+1)=3,468
2	4	12	3	1	1	12×(12×3+1)=444
	5	12	3	1	1	12×(12×3+1)=444
	6 (pooling)	12	24	1	2	12×(12×24+1)=3,468
3	7	12	3	1	1	12×(12×3+1)=444
	8	12	3	1	1	12×(12×3+1)=444
	9 (pooling)	12	24	1	2	12×(12×24+1)=3,468
4	10	12	3	1	1	12×(12×3+1)=444
	11	12	3	1	1	12×(12×3+1)=444
	12 (pooling)	12	24	1	2	12×(12×24+1)=3,468
5	13	12	3	1	1	12×(12×3+1)=444
	14	12	3	1	1	12×(12×3+1)=444
	15 (pooling)	12	48	1	2	12×(12×48+1)=6,924
Bi-GRU	-	-	-	-	-	2×12×3×(12+12+2)=1,872
Attention	-	-	-	-	-	12×2×(12×2+2)=624
Batch-normalization	-	-	-	-	-	12×2×4=96
Dense	-	-	-	-	-	9×(12×2+1)=225
Total						28,035

*Computed as follows:

CNN layer: #lead×(#kernels×kernel size+bias)

Bi-GRU: #directions×#cells×#gates×(#leads+#leads+bias)

Attention: #leads×2×(#leads×2+bias)

Batch-normalization: #leads×2×4

Dense: #output×(#leads×2+bias)

#directions = 2 (bi-direction)

#gates = 3 (gates of reset, update, and current memory)

#leads = 12 (12 ECG leads)

#output = 9 (9 rhythm types)

Transparent Methods

The CPSC2018 ECG database has been described in detail by Liu and coworkers (Liu et al., 2018). Briefly, a total of 9,831 12-lead ECG recordings from 9,458 individuals were collected from 11 hospitals in China. The ECGs were sampled by a frequency of 500 Hertz for a few seconds to a minute, with a few exceptions including one lasting as long as 144 seconds. Each recording was also labeled as a normal type or one of eight abnormal CA types as mentioned above. The database was divided by a random 70-30 training-test split, and only the training set was made available to the public. Gender and age distribution between the training set and the test set were fairly balanced, as were the distributions of the subjects from the 11 hospitals and the CA types (Liu et al., 2018). Of the 6,877 training-set recordings, 470 received two CA-type labels and six received three.

Our model was built on a combined architecture of five CNN blocks, followed by a bidirectional gated recurrent unit (GRU), an attention layer (Schuster and Paliwal, 1997; Yang et al., 2016), and finally a dense (i.e., fully connected) layer (Figure 1). Our choice of using five CNN blocks and the specific number and types of neural network layers was determined by a limited trial and error process, as we made no attempt to examine many other potentially good models, either of a similar or of a very different architecture. In our model, each CNN block contained two convolution layers that were followed by a pooling layer to reduce the amount of parameters and computation in the network and to control over-fitting (Hearty, 2016). Furthermore, between these CNN blocks or between other independent layers, including the one between the last CNN block and the bidirectional GRU layer, we randomly dropped 20% of their connections. We chose to use CNN and RNN because of their demonstrated ability to handle noisy signals and time series data (Pal and Prakash, 2017; Rutkowski, 2008) and of recent studies that included ECG classification (Tan et al., 2018). GRU is a new form of RNN that was recently proposed, and it may require less training time and fewer iterations than LSTM (Cho et al., 2014; Chung et al., 2014). We used batch normalization to adjust and scale the input from the attention layer, which is in a special form proposed by Yang et al. (Yang et al., 2016) and which determines a vector of importance weights, to the dense layer (Ioffe and Szegedy, 2015). LeakyReLU activation function, a leaky version of Rectified Linear Unit, was used for each layer, except for the dense layer, for which Sigmoid activation function was used (Maas, 2013).

Table S9 lists the types and sizes of hyperparameters as well as the number of trainable parameters for each network layer used in our model. In comparison to a recently published model (Yao et al., 2020), which is quite similar to ours in terms of network architecture, main differences include our use of one bi-directional GRU layer, instead of two layers of uni-directional LSTM, for the RNN, and our placement of batch normalization near the end of the network as opposed to at the end of every CNN layer. Intriguingly, the total numbers of trainable parameters between the two models are vastly different, with ours being the much smaller of the two (~28K vs. ~5M) mainly because of very different kernel sizes used (cf. Table S9 and Yao et al.'s Table 3). This suggests the existence of likely numerous different deep learning models that can achieve similar performances, at least for the case of CA detection using ECG data.

In our implementation, the CPSC2018 ECG data were processed in a matrix form consisting of three elements: the first was the subject ID; the second identified which of the ECG's 12 leads was being considered; and the third contained its 72,000 ECG values, which corresponded to the recordings taken for the maximum recording time (144 seconds) and at a frequency of 500 Hertz. We padded zeroes up front for any recording that was less than the maximum time. The 476 multi-labeled subjects were extracted when the other 6,401 subjects were randomly divided into 10 equal parts to set up an 8-1-1 train, validation, and test scheme of machine learning. The extracted multi-labeled subjects were then added back to be included for the training. Our classification training was carried out using categorical-cross-entropy loss function and ADAM optimizer in the GPU version of TensorFlow from the Keras package (Abadi et al., 2016; Charles, 2013; Kingma and Ba, 2014). Models were evaluated on their performance on the validation set for 100 training epochs (an epoch refers to one cycle through the full training dataset in artificial neural network learning). The best model, which was the one with the smallest loss on the validation set, was further evaluated by computing its F1-score on the test set. The procedure was repeated 10 times to complete the 10-fold training and validation plus test to produce 10 best validation models. The median F1-score for each CA label, including the normal type, for the 10 test sets was calculated using the F1-score package from Scikit-learn (Pedregosa et al., 2011).

We further investigated the performance of using only single-lead data. To do that, for a given lead we simply assigned zero to all the ECG values of the other 11 leads and derived the model using the same network architecture and the same 10-fold cross-validation plus test procedure described above. This process resulted in 120 best single-lead validation models and a median F1-score for each of the 12 single leads on each of the nine CA labels.

To compete for CPSC2018, the 130 best validation models (10 from full-lead training and 120 from single-lead training) were combined into one ensemble model for which the average of the output probabilities from the 130 models for each CA type was adjusted by a weight vector to produce the final probability for that CA type. The vector's nine weights, each for each of the nine CA types, were optimized by a genetic algorithm (GA) (Goldberg, 1989) to produce the best overall median F1-score on the 10 test sets. In this GA optimization, a mating system consisting of 40 genes, each of a DNA length of 9, and a population of 100 was set up, and the mating (optimization) process was followed for 100 generations using a mating probability (DNA crossover) of 0.5 and a mutation probability also of 0.5. Given an input of an ECG recording, the CA type receiving the largest probability from the ensemble model would then be the type of CA predicted for that ECG recording. The ensemble model was our model submitted to CPSC2018, and its performances on the hidden test set (2,954 recordings) as computed and reported by CPSC2018 organizers are presented in Table 1.

Supplemental references

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., *et al.* (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. In ArXiv e-prints.

Charles, P.W.D. (2013). Project Title. GitHub repository *https://github.com/charlespwd/project-title*.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:14091259.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:14123555.

Goldberg, D.E. (1989). Genetic Algorithms in Search Optimization and Machine Learning. Addison-Wesley Longman Publishing.

Hearty, J. (2016). Advanced Machine Learning with Python (Packt Publishing).

Ioffe, S., and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:150203167.

Kingma, D.P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. In ArXiv e-prints.

Liu, F., Liu, C., Zhao, L., Zhang, X., Wu, X., Xu, X., Liu, Y., Ma, C., Wei, S., He, Z., *et al.* (2018). An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection. Journal of Medical Imaging and Health Informatics *8*, 1368-1373.

Maas, A.L. (2013). Rectifier Nonlinearities Improve Neural Network Acoustic Models Proc. icml *30 (1)*, 3.

Pal, A., and Prakash, P. (2017). Practical Time Series Analysis: Master Time Series Data Processing, Visualization, and Modeling using Python (Packt Publishing).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al.* (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research *12*, 2825-2830.

Rutkowski, L. (2008). Computational Intelligence: Methods and Techniques (Springer Berlin Heidelberg).

Schuster, M., and Paliwal, K.K. (1997). Bidirectional recurrent neural networks. Ieee Transactions on Signal Processing *45*, 2673-2681.

Tan, J.H., Hagiwara, Y., Pang, W., Lim, I., Oh, S.L., Adam, M., Tan, R.S., Chen, M., and Acharya, U.R. (2018). Application of stacked convolutional and long short-term memory network for accurate identification of CAD ECG signals. Computers in biology and medicine, *94*, 19-26

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A.J., and Hovy, E.H. (2016). Hierarchical Attention Networks for Document Classification. Paper presented at: HLT-NAACL.

Yao, Q., Wang, R., Fan, X., Liu, J., and Li, Y. (2020). Multi-class Arrhythmia detection from 12-lead varied-length ECG using Attention-based Time-Incremental Convolutional Neural Network. Information Fusion, *53*, 174-182.