Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020

Matthew A Reyna[1], Erick A Perez Alday[1], Annie Gu[1], Chengyu Liu[2], Salman Seyedi[1], Ali Bahrami Rad[1], Andoni Elola[1,3], Qiao Li[1], Ashish Sharma[1], Gari D Clifford[1,4]

[1]Department of Biomedical Informatics, Emory University, USA
[2]School of Instrument Science and Engineering, Southeast University, China
[3]Department of Communications Engineering, University of the Basque Country, Spain
[4]Department of Biomedical Engineering, Georgia Institute of Technology, USA

Abstract

The PhysioNet/Computing in Cardiology Challenge 2020 focused on the identification of cardiac abnormalities in 12-lead electrocardiogram (ECG) recordings. A total of 66,405 recordings were sourced from five hospital systems in four countries and annotated with clinical diagnoses. We shared 43,101 annotated recordings publicly and withheld the remaining recordings for testing.

We challenged participants to design working, open-source algorithms for identifying cardiac abnormalities in 12-lead ECG recordings. For this year's Challenge, we sourced data from several institutions with different demographics, required participants to submit code for training their models, and proposed a novel evaluation metric that awards partial credit for misclassified cardiac abnormalities with low risks or similar outcomes as the actual abnormalities. These innovations encouraged the development of generalizable, reproducible, and clinically relevant algorithms.

To date, over 200 teams submitted over 900 algorithms during the Challenge, representing a diversity of approaches from both academia and industry for identifying cardiac abnormalities. The official phase of the Challenge is currently ongoing.

1. Introduction

The PhysioNet/Computing in Cardiology Challenge is an international competition for open-source solutions to complex physiological signal processing and medical classification problems [1]. In 2020, the Challenge's 21st year, we asked participants to develop automated techniques for detecting and classifying cardiac abnormalities in 12-lead electrocardiogram (ECG) recordings [2–4].

Cardiovascular disease is the leading cause of death worldwide, but different cardiovascular diseases have different causes, different risks, and different treatment options [5]. The ECG is an essential tool for screening and diagnosing cardiac abnormalities [6, 7]. ECGs provide a representation of the electrical activity of the heart using measurements from electrodes that are placed on the torso. Painless, harmless, and non-invasive, the standard 12-lead ECG is widely used to diagnose a variety of cardiac arrhythmias (atrial fibrillation, etc.) and other cardiac anatomy abnormalities (ventricular hypertrophy, etc.) [7]. ECG abnormalities have also been identified as both short- and long-term mortality risk predictors [8, 9]. Therefore, early and accurate diagnoses of cardiac ECG abnormalities can improve outcomes.

The manual interpretation of ECGs is time-consuming and requires skilled personnel with a high degree of training, but a number of 12-lead ECG classifiers have emerged over the past decade [10–12]. However, most of these methods have only been tested or developed in single, small, or relatively homogeneous datasets using a small number of cardiac arrhythmias that do not represent the complexity and difficulty of ECG interpretation.

The PhysioNet/Computing in Cardiology Challenge 2020 provided an opportunity to address these problems by providing data from a wide set of sources with a large set of cardiac abnormalities [1, 3, 4]. We asked participants to design and implement a working, open-source algorithm that can, based only on the clinical data provided, automatically identify any cardiac abnormalities present in a 12-lead ECG recording. The winners of the Challenge were the team whose algorithm achieved the highest score for recordings in the hidden test set.

For this year's Challenge, we sourced data for several countries to encourage and assess generalizability to different demographics and institutional practices. We also required that each model be reproducible from the provided training data to improve the reproducibility of the participants' approaches. Finally, we developed a new scoring function that explicitly awards partial credit to misdiagnoses that result in similar treatments or outcomes as the true diagnosis or diagnoses as judged by our cardiologists.

This year's Challenge is ongoing. We will update this manuscript with the results after the end of the Challenge.

2. Challenge Data

For the PhysioNet/Computing in Cardiology Challenge 2020, we assembled multiple databases from across the world. Each database contained 12-lead ECG recordings with diagnoses and demographic information. We shared data from four sources publicly for training and retained data from three sources for testing. Two of the three sources of test data are also sources of training data, but very few, if any, individuals had ECG recordings in both the training and test sets. We posted the training data and labels but did not post the test data or labels to avoid common machine learning problems such as overfitting. The completely hidden dataset has never been posted publicly.

- **CPSC.** The first source is the China Physiological Signal Challenge in 2018 (CPSC2018), held during the 7$^{\text{th}}$ International Conference on Biomedical Engineering and Biotechnology in Nanjing, China [13]. This source includes two databases: a public training dataset (CPSC) and unused data (CPSC-Extra) from CPSC2018. The unused data is not the test data from the CPSC2018, which remains hidden and was used for testing in this Challenge 2020.
- **INCART.** The second source set is the public dataset from the St. Petersburg INCART 12-lead Arrhythmia Database, St. Petersburg Institute of Cardiological Technics, St. Petersburg, Russia, which is posted on in PhysioNet [14].
- **PTB.** The third source is the Physikalisch Technische Bundesanstalt (PTB), Brunswick, Germany, which includes two public databases: the PTB Diagnostic ECG Database [15] and the PTB-XL [16], a large publicly available electrocardiography dataset.
- **Georgia.** The fourth source is the Georgia 12-lead ECG Challenge (G12EC) Database, Emory University, Atlanta, Georgia, USA. This is a new database, representing a large population from the Southeastern United States, and is split in two for training and testing in this Challenge.
- **Undisclosed.** The fifth source is an undisclosed American institution that is geographically distinct from the other sources. This dataset has never been (and may never be) posted publicly and is used for testing in this Challenge.

Each annotated ECG recording contained 12-lead ECG signal data and demographic information, including age, sex, and diagnoses of cardiac abnormalities, i.e., the labels for the Challenge data. See [2] for a fuller description of the data.

The training data contain 111 diagnoses/classes. We used 27 of the 111 total diagnoses to evaluate participant algorithms; see [2]. These 27 diagnoses were relatively common, of clinical interest, and more likely to be recognizable from ECG recordings. However, all 111 classes were included in the training data so that participants can decide whether or not to use them with their algorithms. The test data contained a subset of the 111 diagnoses in potentially different proportions, but each diagnosis in the test data was represented in the training data.

All data were provided in WFDB format [1]. Each ECG recording had a binary MATLAB v4 file for the ECG signal data and a text file in WFDB header format describing the recording and patient attributes, including the diagnosis or diagnoses. We did not change the original data or labels from the databases, except (1) to provide consistent and Health Insurance Portability and Accountability Act (HIPPA)-compliant identifiers for age and sex, (2) to add approximate SNOMED CT codes as diagnoses for each recording, and (3) to change the amplitude resolution to save the data as integers as required for WFDB format.

3. Challenge Objective

We asked participants to design working, open-source algorithms for identifying cardiac abnormalities in 12-lead ECG recordings. To the best of our knowledge, for the first time in any public competition, we required code both for a team's trained models and the code for training their models, which improved the generalizability and reproducibility of the research conducted during the Challenge. We ran the participants' trained models on the hidden test data and evaluated their performance using a novel, expert-based evaluation metric that we designed for this year's Challenge.

3.1. Classification of 12-lead ECGs

We required teams to submit both their trained models along with code for training their models. Teams included any processed and relabeled training data in this step; any changes to the training data were considered to be part of training the model.

We first ran each team's training code on the full training data and then ran each team's trained code from the previous step sequentially on the recordings from the hidden test sets.

3.2. Challenge Scoring

For this year's Challenge, we developed a new scoring metric that awards partial credit to misdiagnoses that result in similar outcomes or treatments as the true diagnoses as judged by our cardiologists. This scoring metric reflects the clinical reality that some misdiagnoses have low risks or similar outcomes to the same diagnoses.

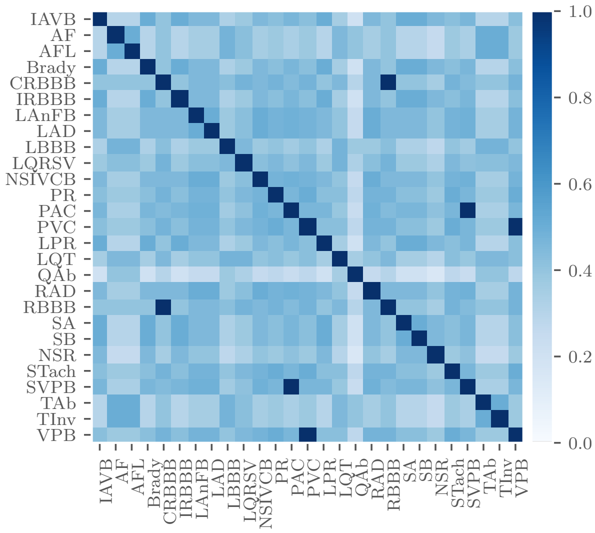

Table 1. Reward matrix W for the diagnoses scored in the Challenge, where columns are the actual diagnoses and columns and rows are the classifier outputs.

Let $C = \{c_i\}_{i=1}^{m}$ be a collection of m distinct diagnoses for a database of n recordings. First, we defined a multi-class confusion matrix $A = [a_{ij}]$, where

$$a_{ij} = \sum_{k=1}^{n} a_{ijk}, \tag{1}$$

with

$$a_{ijk} = \begin{cases} \frac{1}{\{x_k \cup y_k\}}, & \text{if } c_i \in x_k \text{ and } c_j \in y_k, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The quantity $\{x_k \cup y_k\}$ is the set of distinct classes with a positive label and/or classifier output for recording k. We allowed classifiers to receive slightly more credit from recordings with multiple labels than from those with a single label, but each additional positive label or classifier output may reduce the potential credit for that recording.

Next, we defined a reward matrix $W = [w_{ij}]$, where w_{ij} is the reward for a positive classifier output for class c_i with a positive label c_j. The entries in W are defined by our cardiologists based on the similarity of treatments or differences in risks (see Table 1). The matrix W awards full credit to correct classifier outputs, partial credit to incorrect classifier outputs, and no credit for labels and classifier outputs that are not captured in the weight matrix. Three similar classes (i.e., PAC and SVPB, PVC and VPB, CRBBB and RBBB) are scored as if they were the same class. However, we did not change the labels in the training or test data.

Finally, we defined a score

$$s_{\text{unnormalized}} = \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} a_{ij} \tag{3}$$

for each classifier as a weighted sum of the entries in the confusion matrix. For better interpretability, we normalized this score so that a classifier that always outputs the true class or classes receives a score of 1 and an inactive classifier that always outputs the normal class receives a score of 0, i.e.,

$$s_{\text{normalized}} = \frac{s_{\text{unnormalized}} - s_{\text{inactive}}}{s_{\text{true}} - s_{\text{inactive}}}, \tag{4}$$

where s_{inactive} is the score for the inactive classifier and s_{true} is the score for ground-truth classifier.

4. Results

At the time of writing, we have received over 700 submissions of algorithms from nearly 200 teams across academia and industry. The most common algorithmic approach was based on deep learning and convolutional neural networks. However, over 70% of entries used standard, hand-crafted features with classifiers such as support vector machines, gradient boosting, random forests, and shallow neural networks.

5. Conclusions

This article describes the world's largest open-access database of 12-lead ECGs with data drawn from five institutions in four countries across three continents. The data were annotated with 111 diagnoses; 27 of these diagnoses representing were the focus of a novel scoring matrix that rewarded algorithms based on similarities between diagnostic outcomes that we weighted by severity or risk.

The public training data and the sequestered test data provided the opportunity for unbiased and comparable repeatable research. To the best of our knowledge, this is the first public competition that has required the teams to provide both their original source code and the framework for (re)training their code. In doing so, this creates the first truly repeatable body of work on electrocardiograms and many related areas of research.

Acknowledgements

This research is supported by the National Institute of General Medical Sciences (NIGMS) and the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under NIH grant number 2R01GM104987-09, the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number

UL1TR002378, the Gordon and Betty Moore Foundation, and AliveCor, Inc. The content is solely the responsibility of the authors and does not necessarily represent the official views of these entities.

References

[1] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. circulation 2000;101(23):e215–e220.

[2] Perez Alday EA, Gu A, Shah A, Robichaux C, Wong AKI, Liu C, Liu F, Rad BA, Elola A, Seyedi S, Li Q, Sharma A, Clifford GD, Reyna MA. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. Under Review 2020;.

[3] PhysioNet Challenges. https://physionet.org/about/challenge/. Accessed: 2020-02-07.

[4] PhysioNet/Computing in Cardiology Challenge 2020. https://physionetchallenges.github.io/2020/. Accessed: 2020-02-07.

[5] Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Das SR, et al. Heart Disease and Stroke Statistics – 2019 Update: a report From the American Heart Association. Circulation 2019;.

[6] Kligfield P. The centennial of the Einthoven electrocardiogram. Journal of Electrocardiology 2002;35(4):123–129.

[7] Kligfield P, Gettes LS, Bailey JJ, Childers R, Deal BJ, Hancock EW, Van Herpen G, Kors JA, Macfarlane P, Mirvis DM, et al. Recommendations for the standardization and interpretation of the electrocardiogram: part i: the electrocardiogram and its technology 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. Journal of the American College of Cardiology 2007;49(10):1109–1127.

[8] Mozos I, Caraba A. Electrocardiographic predictors of cardiovascular mortality. Disease markers 2015;2015.

[9] Gibbs C, Thalamus J, Kristoffersen DT, Svendsen MV, Holla ØL, Heldal K, Haugaa KH, Hysing J. QT prolongation predicts short-term mortality independent of comorbidity. EP Europace 2019;21(8):1254–1260.

[10] Ye C, Coimbra MT, Kumar BV. Arrhythmia detection and classification using morphological and dynamic features of ECG signals. In 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology. IEEE, 2010; 1918–1921.

[11] Ribeiro AH, Ribeiro MH, Paixão GM, Oliveira DM, Gomes PR, Canazart JA, Ferreira MP, Andersson CR, Macfarlane PW, Wagner Jr M, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nature Communications 2020;11(1):1–9.

[12] Chen TM, Huang CH, Shih ES, Hu YF, Hwang MJ. Detection and Classification of Cardiac Arrhythmias by a Challenge-Best Deep Learning Neural Network Model. Iscience 2020;23(3):100886.

[13] Liu F, Liu C, Zhao L, Zhang X, Wu X, Xu X, Liu Y, Ma C, Wei S, He Z, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. Journal of Medical Imaging and Health Informatics 2018;8(7):1368–1373.

[14] Tihonenko V, Khaustov A, Ivanov S, Rivin A, Yakushenko E. St Petersburg INCART 12-lead arrhythmia database. PhysioBank PhysioToolkit and PhysioNet 2008;Doi: 10.13026/C2V88N.

[15] Bousseljot R, Kreiseler D, Schnabel A. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. Biomedizinische TechnikBiomedical Engineering 1995;40(s1):317–318.

[16] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, Schaeffter T. PTB-XL, a large publicly available electrocardiography dataset. Scientific Data 2020; 7(1):1–15.

Address for correspondence:

Matthew A Reyna
DBMI, 101 Woodruff Circle, 4th Floor East, Atlanta, GA 30322
matthew.a.reyna@emory.edu