Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021

Matthew A Reyna[1], Nadi Sadr[1], Erick A Perez Alday[1], Annie Gu[1], Amit J Shah[2,3], Chad Robichaux[1], Ali Bahrami Rad[1], Andoni Elola[1,4], Salman Seyedi[1], Sardar Ansari[5], Hamid Ghanbari[5], Qiao Li[1], Ashish Sharma[1], Gari D Clifford[1,6]

[1]Department of Biomedical Informatics, Emory University, USA
[2]Department of Epidemiology, Emory University, USA
[3]Department of Medicine, Division of Cardiology, Emory University, USA
[4]Department of Communications Engineering, University of the Basque Country, Spain
[5]Department of Emergency Medicine, University of Michigan, USA
[6]Department of Biomedical Engineering, Georgia Institute of Technology, USA

Abstract

The PhysioNet/Computing in Cardiology Challenge 2021 focused on the identification of cardiac abnormalities from electrocardiograms (ECGs) and assessed the diagnostic potential of reduced-lead ECGs relative to the standard but less-accessible twelve-lead ECG.

We sourced 131,155 recordings with clinical diagnoses from seven institutions in four countries, sharing 88,253 annotated recordings publicly and withholding the remaining recordings for validation and testing.

We asked the Challenge participants to design working, open-source algorithms for identifying cardiac abnormalities from twelve-lead, six-lead, four-lead, three-lead, and two-lead ECG recordings. By sourcing data from diverse populations, requiring the submission of reusable training code, and designing an evaluation metric specifically for this task, we encouraged the development of generalizable, reproducible, and clinically relevant algorithms for identifying cardiac abnormalities from ECGs.

A total of 68 teams submitted a total of 1056 algorithms during the Challenge. Of these, 39 teams were ultimately successful, representing a diversity of approaches from both academia and industry.

1. Introduction

The PhysioNet/Computing in Cardiology Challenge is an annual competition that supports the development of open-source solutions to complex physiological signal processing and medical classification problems [1]. In 2021, the Challenge's 22nd year, we extended the previous year's Challenge to ask participants to develop automated techniques for detecting and classifying cardiac abnormalities from both twelve-lead electrocardiogram (ECG) recordings and reduced-lead ECG recordings [2–4].

Cardiovascular disease is a leading cause of death worldwide, but different cardiovascular diseases have different causes, risks, and treatments [5]. The standard twelve-lead ECG is a widely used, non-invasive tool for monitoring cardiac function and diagnosing cardiac disorders [6]. However, breakthroughs in ECG technologies have led to the development of smaller, lower-cost, and easier-to-use devices that improve access in low-resource and home settings with remote patient monitoring programs. Subsets of the standard twelve leads can be comparable to the full set of leads in limited contexts, and there is limited evidence that reduced-lead ECGs can capture the wide range of diagnostic information captured by twelve-lead ECGs [7–9].

The PhysioNet/Computing in Cardiology Challenge 2021 provided an opportunity to investigate the ability of algorithms to use various reduced-lead ECGs to detect a variety of cardiac abnormalities from a diverse sources with a wide range of cardiac abnormalities. The goal of the 2021 Challenge was to perform clinical diagnoses from twelve-lead, six-lead, four-lead, three-lead, and two-lead ECG recordings [2, 4][1]. We asked participants to design and implement working, open-source algorithms that, based only on the provided ECG recordings and routine demographic data, could automatically identify any cardiac abnormalities present in a recording. The prizes were awarded for the top-performing algorithms.

[1]The 2017 Challenge focused on the identification of atrial fibrillation from single-lead ECGs, and the 2020 Challenge focused on the identification of 27 cardiac abnormalities from twelve-lead ECGs [2, 10].

We sourced annotated ECG recordings from seven institutions in four countries across three continents to encourage and assess model generalizability to different demographics and institutional practices. We also developed a scoring function that awards partial credit to misdiagnoses that result in similar treatments or outcomes as the true diagnoses. Finally, we required that each model be reproducible from the provided training data.

2. Challenge Data

We assembled nine databases with 131,149 twelve-lead ECG recordings from across the world. We sourced 88,253 recordings from seven databases for training, 6,630 recordings from two databases for validation, and 36,266 recordings from four databases for testing, including two sources that were not represented in the training or validation sets. We posted the training data and labels publicly but withheld the validation and test data and labels to avoid common machine learning problems such as overfitting.

We introduced the first six of the below databases in the 2020 Challenge [2, 3] and the last three of the below databases for the 2021 Challenge, increasing the total number of annotated ECG recordings from 66,361 to 131,149.

- **CPSC.** This database contains 13,256 recordings from the China Physiological Signal Challenge (CPSC) 2018 [11]. We used the training data and unused data from CPSC 2018 as training data and the test data from CPSC 2018 as validation and test data.
- **INCART.** This database contains 74 recordings from the St. Petersburg INCART 12-lead Arrhythmia Database [12]. We used this database as training data.
- **PTB.** This database contains 516 recordings from the Physikalisch-Technische Bundesanstalt (PTB) database [13]. We used this database as training data.
- **PTB-XL.** This database contains 21,837 recordings from the PTB-XL database [14]. We used this database as training data.
- **G12EC.** This database contains 20,672 recordings from the Georgia 12-lead ECG Challenge (G12EC) Database. We split this dataset into training, validation, and test data.
- **Undisclosed.** This database contains 10,000 recordings from an undisclosed American institution that is geographically distinct from the other sources. This database has never been (and may never be) posted publicly. We used this database as test data.
- **Chapman-Shaoxing.** This database contains 10,247 recordings from Chapman University and Shaoxing People's Hospital [15]. We use this database as training data.
- **Ningbo.** This database contains 34,905 recordings from Ningbo First Hospital [16]. We used this database as training data.
- **UMich.** This database contains 19,642 recordings from the University of Michigan[2]. We used this database as test data.

Number of leads	Lead combination
12	I, II, III, aVR, aVL, aVF, V1, V2, V3, V4, V5, V6
6	I, II, III, aVR, aVL, aVF
4	I, II, III, V2
3	I, II, V2
2	I, II

Table 1. Lead combinations used for the Challenge validation and test sets.

The annotated ECG recordings contained ECG signal data and demographic information, including age, sex, and diagnoses of cardiac abnormalities, i.e., the labels for the Challenge data. Participants were given signal data from all twelve leads in the training set, but they were only given the signal data for the lead combinations in Table 1 for each of the validation and test sets.

The training set contained 133 diagnoses or classes. The validation and test sets contained subsets of these 133 diagnoses in potentially different proportions, but each diagnosis in the validation and test sets was represented in the training data.

All data were provided in WFDB format with SNOMED CT codes as the labels for the recordings [1, 17]. We did not change the data or labels from the original databases except to provide consistent, Health Insurance Portability and Accountability Act (HIPPA)-compliant identifiers for age and sex, to provide approximate SNOMED CT codes for the labels, and to encode the data in WFDB format.

See [2, 4] for a more complete description of the Challenge data.

3. Challenge Objective

We asked the participants to design and implement working, open-source algorithms to automatically identify any cardiac abnormalities present in twelve-lead, six-lead, four-lead, three-lead, and two-lead ECGs recordings. Like the 2020 Challenge, we required participants to provide their trained models and the code for training their models, improving the generalizability and reproducibility of the research conducted during the Challenge. We ran each

[2]De-identified data collected under U-M HUM00092309: Approximately 20,000 ten-second-long twelve-lead ECGs obtained from the University of Michigan Section of Electrophysiology. The sample was randomly selected from the patients who had a routine ECG test from 1990 to 2013 to approximately match the demographics of the training databases. The dataset was de-identified and contains only basic demographics information such as age (any age over the age of 90 is denoted as 90+) and sex, the ECG waveforms and the diagnosis statements associated with the record.

team's training algorithm on the training data and ran the resulting models on the hidden validation and test data, evaluating their performance using an expert-based evaluation metric that we designed for the 2021 Challenge.

3.1. Submissions

We required teams to submit the code for training their models along with their trained models. Teams included any processed data and labels as a part of training.

We first ran each team's training code on the full training data. We then ran the resulting trained model on twelve-lead, six-lead, four-lead, three-lead, and two-lead versions of the validation and test sets. In each case, we ran the model sequentially, requiring the model to return classifier outputs for each recording before accessing the next one. We allowed up to 72 hours for training and 24 hours for validation and testing. See [4] for details about the run time environment and resources.

3.2. Scoring

We extended the 2020 Challenge scoring metric to incorporate additional data and diagnoses for the 2021 Challenge. This scoring metric awarded full credit to correct diagnoses and partial credit to misdiagnoses that result in similar outcomes or treatments as the given diagnoses.

We used 30 of the 133 diagnoses in the Challenge data to evaluate the algorithms. Our cardiologists chose these 30 diagnoses because they were relatively prevalent, of clinical interest, and electrophysiological and therefore able to be accurately diagnosed using ECG recordings alone. They determined the amount of credit given for misdiagnoses; see Table 2. We did not score the other 103 diagnoses.

For each classifier, we compared the classifier outputs for the recordings with the diagnoses given by the recording labels and awarded the credit or reward shown in Table 2 to the classifier for each recording. We calculated the sum of these values for each recording in a database and normalized the sum so that a classifier that always identifies the correct diagnoses received a score of 1 and a classifier that always identifies the sinus rhythm diagnosis received a score of 0. See [2,4] for a complete mathematical description of the Challenge scoring.

4. Challenge Results

A total of 68 teams submitted 1,056 attempts, 618 of which were successful. Of these, 39 teams qualified to be ranked. Table 3 summarize the highest ranked teams for prize-winning categories: the highest Challenge metric scores on the two-lead version of the hidden test data, the highest scores on the three-lead version of the test data, and

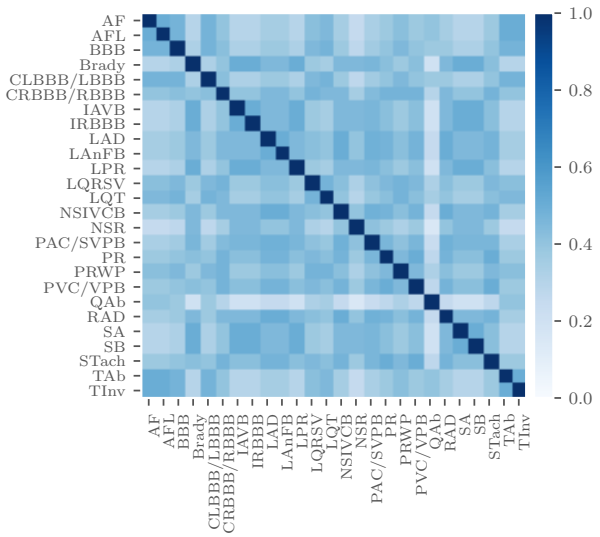

Table 2. Reward matrix for the diagnoses scored in the Challenge, where columns are diagnoses and rows are the classifier outputs. Columns/rows with multiple diagnoses have different labels but are scored identically.

Rank	Team Name	All-lead score	3-lead score	2-lead score
1	ISIBrno-AIMT	0.58	0.58	0.58
2	DSAIL_SNU	0.57	0.57	0.57
3	NIMA	0.56	0.55	0.55

Table 3. Three highest ranked teams and their Challenge metric scores for the three prize-winning categories. The all-lead category was calculated as the mean score for the twelve-lead, three-lead, and two-lead versions of the test data.

the highest mean of the scores on the two-lead, three-lead, and twelve-lead versions. These categories were chosen to emphasize the independent lead combinations. Additional scores are available on [4].

There was little change in overall performance on the test data across different lead combinations (median change ≤ 0.036 and two-sided signed rank-sum test $p \geq 0.50$ for all pairwise comparisons of the lead combinations). There was also little change in performance from the CPSC and G12EC validation databases to the CPSC and G12EC test databases (median change 0.012 and two-sided signed rank-sum test $p = 0.28$ for the two-lead category), which were sources that were represented in the training data. However, much like the 2020 Challenge, there was a significant drop in performance from the CPSC and G12EC validation databases to the completely hidden Undisclosed and UMich test databases (median change 0.21 and 0.080 and two-sided rank-sum test $p = 1.53 \cdot 10^{-7}$ and $p = 6.70 \cdot 10^{-3}$ for the Undisclosed

and UMich test databases, respectively, for the two-lead category), which were unrepresented in the training data.

5. Conclusions

This article describes the augmentation of the world's largest open-access database of twelve-lead ECGs with data drawn from nine sources in four countries across three continents, together with an international competition ('Challenge') based on these data. The data were annotated with 133 diagnoses; 30 diagnoses were the focus of a scoring metric that rewarded algorithms based on similarities between diagnostic outcomes that we weighted by severity or risk. This year's Challenge also differed from the 2020 Challenge by asking teams to classify with as few as two leads.

The results suggest that 'two will do' for some classes, but small differences in overall performance across different lead combinations belie larger differences in performance for individual diagnoses. Conversely, large differences in performance on the hidden test sets demonstrate the challenge of generalizing models to new databases.

The public training data and the sequestered validation and test data provided opportunities for unbiased and comparable repeatable research, as well as a corpus of reusable algorithms for the identification of cardiac abnormalities from standard twelve-lead and reduced-lead ECGs.

Acknowledgements

This work was supported by the National Institute of General Medical Sciences (NIGMS) and the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under NIH grant number R01EB030362, the National Center for Advancing Translational Sciences of the National Institutes of Health under award number UL1TR002378, the Gordon and Betty Moore Foundation, MathWorks, 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, et al. 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, et al. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. Physiological Measurement 2020;41.

[3] PhysioNet/Computing in Cardiology Challenge 2020. https://physionetchallenges.org/2020/. Accessed: 2021-06-08.

[4] PhysioNet/Computing in Cardiology Challenge 2021. https://physionetchallenges.org/2021/. Accessed: 2021-09-20.

[5] Virani SS, Alonso A, Aparicio HJ, Benjamin EJ, Bittencourt MS, Callaway CW, et al. Heart Disease and Stroke Statistics – 2021 Update: a Report from the American Heart Association. Circulation 2021;143(8):e254–e743.

[6] Kligfield P. The Centennial of the Einthoven Electrocardiogram. J Electrocardiol 2002;35(4):123–129.

[7] Aldrich HR, Hindman NB, Hinohara T, Jones MG, Boswick J, Lee KL, et al. Identification of the Optimal Electrocardiographic Leads for Detecting Acute Epicardial Injury in Acute Myocardial Infarction. Am J Cardiol 1987; 59(1):20–23.

[8] Drew BJ, Pelter MM, Brodnick DE, Yadav AV, Dempel D, Adams MG. Comparison of a New Reduced Lead Set ECG with the Standard ECG for Diagnosing Cardiac Arrhythmias and Myocardial Ischemia. J Electrocardiol 2002;35(4, Part B):13–21.

[9] Green M, Ohlsson M, Lundager Forberg J, Björk J, Edenbrandt L, Ekelund U. Best Leads in the Standard Electrocardiogram for the Emergency Detection of Acute Coronary Syndrome. J Electrocardiol 2007;40(3):251–256.

[10] Clifford GD, Liu C, Moody B, Lehman LwH, Silva I, Li Q, et al. AF Classification from a Short Single Lead ECG Recording: the PhysioNet/Computing in Cardiology Challenge 2017. In 2017 Computing in Cardiology (CinC), volume 44. IEEE, 2017; 1–4.

[11] Liu F, Liu C, Zhao L, Zhang X, Wu X, Xu X, 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.

[12] 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.

[13] Bousseljot R, Kreiseler D, Schnabel A. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. Biomedizinische Technik 1995;40(S1):317–318.

[14] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, et al. PTB-XL, a Large Publicly Available Electrocardiography Dataset. Sci Data 2020;7(1):1–15.

[15] Zheng J, Zhang J, Danioko S, Yao H, Guo H, Rakovski C. A 12-lead Electrocardiogram Database for Arrhythmia Research Covering More Than 10,000 Patients. Sci Data 2020; 7(48):1–8.

[16] Zheng J, Cui H, Struppa D, Zhang J, Yacoub SM, El-Askary H, et al. Optimal Multi-Stage Arrhythmia Classification Approach. Sci Data 2020;10(2898):1–17.

[17] Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED Clinical Terms: Overview of the Development Process and Project Status. In Proceedings of the AMIA Symposium. AMIA, 2001; 662–666.

Address for correspondence:

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