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

Erick A. Perez Alday[1], Annie Gu[1], Amit Shah[2], Chad Robichaux[1], An-Kwok Ian Wong[3], Chengyu Liu[4], Feifei Liu[5], Ali Bahrami Rad[1], Andoni Elola[1,6], Salman Seyedi[1], Qiao Li[1], Ashish Sharma[1], Gari D. Clifford[1,7,*], Matthew A. Reyna[1,*]

[1]Department of Biomedical Informatics, Emory University, Atlanta, GA, USA
[2]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA, USA
[3]Department of Medicine, Division of Pulmonary, Allergy, Critical Care, and Sleep Medicine, Emory University, Atlanta, GA, USA
[4]School of Instrument Science and Engineering, Southeast University, Nanjing, Jiangsu, China
[5]School of Science, Shandong Jianzhu University, Jinan, Shandong, China
[6]Department of Communications Engineering, University of the Basque Country, Spain
[7]Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, USA
*These authors are joint senior authors.

E-mail: matthew.a.reyna@emory.edu

Abstract.

Objective: Vast 12-lead ECGs repositories provide opportunities to develop new machine learning approaches for creating accurate and automatic diagnostic systems for cardiac abnormalities. However, most 12-lead ECG classification studies are trained, tested, or developed in single, small, or relatively homogeneous datasets. In addition, most algorithms focus on identifying small numbers of cardiac arrhythmias that do not represent the complexity and difficulty of ECG interpretation. This work addresses these issues by providing a standard, multi-institutional database and a novel scoring metric through a public competition: the PhysioNet/Computing in Cardiology Challenge 2020.

Approach: A total of 66361 12-lead ECG recordings were sourced from six hospital systems from four countries across three continents. 43,101 recordings were posted publicly with a focus on 27 diagnoses. For the first time in a public competition, we required teams to publish open-source code for both training and testing their algorithms, ensuring full scientific reproducibility.

Main results: A total of 217 teams submitted 1395 algorithms during the Challenge, representing a diversity of approaches for identifying cardiac abnormalities from both academia and industry. As with previous Challenges, high-performing algorithms exhibited significant drops ($\lessapprox 10\%$) in performance on the hidden test data.

Significance: Data from diverse institutions allowed us to assess algorithmic generalizability. A novel evaluation metric considered different misclassification errors for

different cardiac abnormalities, capturing the outcomes and risks of different diagnoses. Requiring both trained models and code for training models improved the generalizability of submissions, setting a new bar in reproducibility for public data science competitions.

1. Introduction

Cardiovascular disease is the leading cause of death worldwide [1]. Early treatment can prevent serious cardiac events, and the most important tool for screening and diagnosing cardiac electrical abnormalities is the electrocardiogram (ECG) [2], [3]. The ECG is a non-invasive representation of the electrical activity of the heart that is measured using electrodes placed on the torso. The standard 12-lead ECG is widely used to diagnose a variety of cardiac arrhythmias such as atrial fibrillation and other cardiac anatomy abnormalities such as ventricular hypertrophy [2]. ECG abnormalities have also been identified as both short- and long-term mortality risk predictors [4], [5]. Therefore, the early and correct diagnosis of cardiac ECG abnormalities can increase the chances of successful treatments. However, manual interpretation of ECGs is time-consuming and requires skilled personnel with a high degree of training.

The automatic detection and classification of cardiac abnormalities can assist physicians in making diagnoses for a growing number of recorded ECGs. However, there has been limited success in achieving this goal [6], [7]. Over the last decade, the rapid development of machine learning techniques have also included a growing number of 12-lead ECG classifiers [8]–[10]. Many of these algorithms may identify cardiac abnormalities correctly. However, most of these methods are trained, tested, or developed in single, small, or relatively homogeneous datasets. In addition, most methods focus on identifying 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 [11]–[13]. The PhysioNet Challenge is an initiative that invites participants from academia, industry, and elsewhere to tackle clinically important questions that are either unsolved or not well-solved. Similar to previous years, the Challenge had both an unofficial phase and an official phase that ran over the course of several months. PhysioNet co-hosts the Challenge annually in cooperation with the Computing in Cardiology conference. The goal of the 2020 PhysioNet Challenge was to identify clinical diagnoses from 12-lead ECG recordings.

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. Like previous years, we facilitated

the development of the algorithms through the Challenge but did little to constrain the algorithms themselves. However, we required that each algorithm be reproducible from the provided training data. The winners of the Challenge are the team whose algorithm achieved the highest score for recordings in the hidden test set. We developed a new scoring function that awards partial credit to misdiagnoses that result in similar treatments or outcomes as the true diagnosis or diagnoses as judged by our cardiologists because traditional scoring metrics, such as common area under the curve (AUC) metrics, do not explicitly reflect the clinical reality that some misdiagnoses are more harmful than others and should be scored accordingly.

2. Methods

2.1. Data

For the PhysioNet/Computing in Cardiology Challenge 2020, we assembled multiple databases from across the world. Each database contained recordings with diagnoses and demographic data.

2.1.1. Challenge Data Sources We used data from five different sources. Two sources were split to form training, validation, and test sets; two sources were included only as training data; and one source was included only as test data. These sources of ECG data are described below and summarized in Table 1. We made the training data and clinical ECG diagnoses (labels) publicly available, but the validation and test data were kept hidden. The training, validation and test data were matched as closely as possible for age, sex and diagnosis. The completely hidden dataset has never been posted publicly, allowing us to assess common machine learning problems such as overfitting.

(i) **CPSC.** The first source is the China Physiological Signal Challenge 2018 (CPSC2018), held during the 7$^{\text{th}}$ International Conference on Biomedical Engineering and Biotechnology in Nanjing, China [14]. This source includes three databases: the original public training dataset (CPSC), an unused dataset (CPSC-Extra), and the test dataset (the hidden CPSC set) from the CPSC2018. The CPSC data and CPSC-Extra datasets were shared as training sets. The hidden CPSC set was split into validation and test set for this year's Challenge.

(ii) **INCART.** The second source is the public dataset from the St. Petersburg Institute of Cardiological Technics (INCART) 12-lead Arrhythmia Database, St. Petersburg, Russia, which is posted on in PhysioNet [15]. The dataset was shared as a training set.

(iii) **PTB and PTB-XL.** The third source is the Physikalisch-Technische Bundesanstalt (PTB) Database, Brunswick, Germany. This source includes two public databases: the

Database	Total Patients	Recordings in Training Set	Recordings in Validation Set	Recordings in Test Set	Total Recordings
CPSC	9458	10330	1463	1463	13256
INCART	32	74	0	0	74
PTB	19175	22353	0	0	22353
G12EC	15742	10344	5167	5167	20678
Undisclosed	Unknown	0	0	10000	10000
Total	Unknown	43101	6630	16630	66361

Table 1. Numbers of patients and recordings in the training, validation, and test databases for the Challenge. The training set includes data from the China Physiological Signal Challenge 2018 (CPSC), the St. Petersburg Institute of Cardiological Technics (INCART), the Physikalisch-Technische Bundesanstalt (PTB), and the Georgia 12-lead ECG Challenge (G12EC) databases. The validation set includes data from the CPSC and the G12EC databases. The test set includes data from the CPSC, the G12EC, and the undisclosed databases.

PTB Diagnostic ECG Database [16] and the PTB-XL Database [17], a large publicly available ECG dataset. These datasets were shared as training sets.

(iv) **G12EC.** 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 between the training, validation, and test sets. The validation and test set comprised the hidden G12EC set.

(v) **Undisclosed.** The fifth source is a dataset from an undisclosed American institution that is geographically distinct from the other dataset sources. This dataset has never been (and may never be) posted publicly, and is used as a test set for the Challenge.

2.1.2. Challenge Data Variables Each 12-lead ECG recording was acquired in a hospital or clinical setting. The specifics of the data acquisition depend on the source of the databases, which were assembled around the world and therefore vary. We encourage the readers to check the original publications for details but provide a summary below.

Each annotated ECG recording contained 12-lead ECG signal data with sample frequency varying from 257 Hz to 1 kHz. Demographic information, including age, sex, and a diagnosis or diagnoses, i.e., the labels for the Challenge data, were also included. The quality of the label depended on the clinical or research practices and included labels that were machine-generated, over-read by a single cardiologist, and adjudicated by multiple cardiologists.

Table 2 provides a summary of the age, sex, and recording information for the Challenge databases, indicating differences between the populations. Table 3 and Figure 1 provide

Dataset	Number of Recordings	Mean Duration (seconds)	Mean Age (years)	Sex (male/female)	Sample Frequency (Hz)
CPSC (all data)	13256	16.2	61.1	53%/47%	500
CPSC Training	6877	15.9	60.2	54%/46%	500
CPSC-Extra Training	3453	15.9	63.7	53%/46%	500
Hidden CPSC	2926	17.4	60.4	52%/48%	500
INCART	72	1800.0	56.0	54%/46%	257
PTB	516	110.8	56.3	73%/27%	1000
PTB-XL	21837	10.0	59.8	52%/48%	500
G12EC (all data)	20678	10.0	60.5	54%/46%	500
G12EC Training	10344	10.0	60.5	54%/46%	500
Hidden G12EC	10344	10.0	60.5	54%/46%	500
Undisclosed	10000	10.0	63.0	53%/47%	300

Table 2. Number of recordings, mean duration of recordings, mean age of patients in recordings, sex of patients in recordings, and sample frequency of recordings for each data set. Italicized dataset names indicate that the database is a subset of the source dataset above it. The training, validation and test data were matched as closely as possible for age, sex and diagnosis.

summaries of the diagnoses for the training and validation data. The training data contain 111 diagnoses or classes. We used 27 of these 111 diagnoses to evaluate participant algorithms because they were relatively common, of clinical interest, and more likely to be recognizable from ECG recordings. Table 3 contains the list of the scored diagnoses for the Challenge can be seen in Table 3 with long-form descriptions, the corresponding Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) codes, and abbreviations. Only these scored classes are shown in Table 3 and Figure 1, but all 111 classes were included in the training data so that participants could decide whether or not to use them with their algorithms. The test data contain 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 MATLAB- and WFDB-compatible format [11]. Each ECG recording had a binary MATLAB v4 file for the ECG signal data and an associated text file in WFDB header format describing the recording and patient attributes, including the diagnosis or diagnoses, i.e., the labels for the recording. 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 the diagnoses for the recordings, and (3) to change the amplitude resolution to save the data as integers as required for WFDB format. Saving the signals as integers help reduced storage size and compute times without degrading the

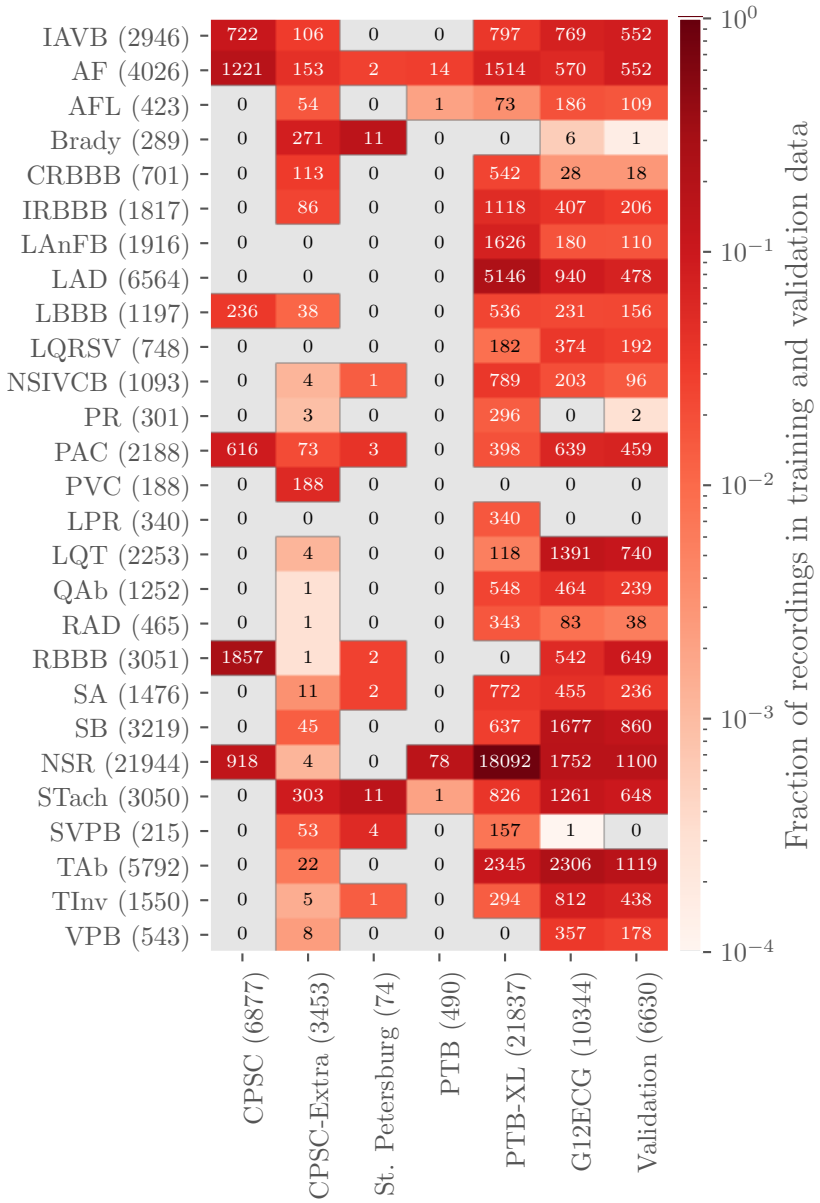

Figure 1. Numbers of recordings with each scored diagnosis in the training and validation sets. Colors indicate the fraction of recordings with each scored diagnosis in each data set, i.e., the total number of each scored diagnosis in a data set normalized by the number of recordings in each data set. Parentheses indicate the total numbers of records with a given label across training and the validation sets (rows) and the total numbers of recordings, including recordings without scored diagnoses, in each data set (columns).

Diagnosis	Code	Abbreviation
1st degree AV block	270492004	IAVB
Atrial fibrillation	164889003	AF
Atrial flutter	164890007	AFL
Bradycardia	426627000	Brady
Complete right bundle branch block	713427006	CRBBB
Incomplete right bundle branch block	713426002	IRBBB
Left anterior fascicular block	445118002	LAnFB
Left axis deviation	39732003	LAD
Left bundle branch block	164909002	LBBB
Low QRS voltages	251146004	LQRSV
Nonspecific intraventricular conduction disorder	698252002	NSIVCB
Pacing rhythm	10370003	PR
Premature atrial contraction	284470004	PAC
Premature ventricular contractions	427172004	PVC
Prolonged PR interval	164947007	LPR
Prolonged QT interval	111975006	LQT
Q wave abnormal	164917005	QAb
Right axis deviation	47665007	RAD
Right bundle branch block	59118001	RBBB
Sinus arrhythmia	427393009	SA
Sinus bradycardia	426177001	SB
Sinus rhythm	426783006	NSR
Sinus tachycardia	427084000	STach
Supraventricular premature beats	63593006	SVPB
T wave abnormal	164934002	TAb
T wave inversion	59931005	TInv
Ventricular premature beats	17338001	VPB

Table 3. Diagnoses, SNOMED CT codes and abbreviations in the posted training databases for diagnoses that were scored for the Challenge.

signal, as it only represents a change in the scaling factor for the signal amplitude.

2.2. 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 that teams submit code both for their trained models

and for training their models, which aided the generalizability and reproducibility of the research conducted during the Challenge. We ran the participants' trained models on the hidden validation and test data and evaluated their performance using a novel, expert-based evaluation metric that we designed for this year's Challenge.

2.2.1. Challenge Overview, Rules, and Expectations This year's Challenge is the 21st PhysioNet/Computing in Cardiology Challenge [11]. Similar to previous Challenges, this year's Challenge had an unofficial phase and an official phase. The unofficial phase (February 7, 2020 to April 30, 2020) provided an opportunity to socialize the Challenge and seek discussion and feedback from teams about the data, evaluation metrics, and requirements. The unofficial phase allowed 5 scored entries for each team. After a short break, the official phase (May 11, 2020 to August 23, 2020) introduced additional training, validation, and test data; a requirement for teams to submit their training code; and an improved evaluation metric. The official phase allowed 10 scored entries for each team. During both phases, teams were evaluated on a small validation set; evaluation on the test set occurred after the end of the official phase of the Challenge to prevent sequential training on the test data. Moreover, while teams were encouraged to ask questions, pose concerns, and discuss the Challenge in a public forum, they were prohibited from discussing their particular approaches to preserve the uniqueness of their approaches for solving the problem posed by the Challenge.

2.2.2. Classification of 12-lead ECGs We required teams to submit both their trained models along with code for training their models. We announced this requirement at the launch of this year's Challenge but did not start requiring the submission of training code until the official phase of the Challenge; by this time, we had a better idea of what teams would need to train their algorithms. Teams included any processed and relabeled training data in the training step; any changes to the training data are part of training a model.

We first ran each team's training code on the training data and then ran each team's trained code from the previous step on the hidden validation and test sets. We ran each algorithm sequentially on the recordings to use them as realistically as possible.

We allowed teams to submit either MATLAB or Python implementations of their code. Other languages, including Julia and R, were supported but received insufficient interest from participants during the unofficial phase. Participants containerized their code in Docker and submitted it in GitHub or Gitlab repositories. We downloaded their code and ran in containerized environments on Google Cloud. The computational environment is given more fully in [18], which describes the previous year's Challenge.

We used virtual machines on Google Cloud with 8 vCPUs, 64 GB RAM, and an optional NVIDIA T4 Tensor Core graphics processing unit (GPU) with a 72 hour time limit for training on the training set. We used virtual machines on Google Cloud with 2 vCPUs, 13 GB RAM, and an optional NVIDIA T4 Tensor Core GPU with a 24 hour time limit for

running the trained classifiers on the test set.

To aid teams, we shared baseline models that we implemented in MATLAB and Python. The Python baseline model was a random forest classifier that used age, sex, QRS amplitude, and RR intervals as features. QRS detection was implemented using the Pan-Tompkins algorithm [19]. The MATLAB baseline model was a hierarchical multinomial logistic regression classifier that used age, sex, and global electrical heterogeneity [20] parameters as features. The global electrical heterogeneity parameters were computed using a time coherent median beat and origin point calculation [21]. The QRS detection and RR interval was implemented using the heart rate variability (HRV) cardiovascular research toolbox [22], [23]. However, it was not the aim of these example models to provide a competitive classifier but instead to provide an example of how to read and extract features from the recordings.

2.2.3. Evaluation of Classifiers 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 are more harmful than others and should be scored accordingly. Moreover, it reflects the fact that it is less harmful to confuse some classes than others because the responses may be similar or the same.

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} is the normalized number of recordings in a database that were classified as belonging to class c_i but actually belong to class c_j (where c_i and c_j may be the same class or different classes). Since each recording can have multiple labels and each classifier can produce multiple outputs for a recording, we normalized the contribution of each recording to the scoring metric by dividing by the number of classes with a positive label and/or classifier output. Specifically, for each recording $k = 1, \ldots, n$, let x_k be the set of positive labels and y_k be the set of positive classifier outputs for recording k. We defined a multi-class confusion matrix $A = [a_{ij}]$ by

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

where

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

The quantity $|x_k \cup y_k|$ is the number of distinct classes with a positive label and/or classifier output for recording k. To incentivize teams to develop multi-class classifiers, 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 available 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 (where c_i and c_j may be the same class or different classes). The entries of W are defined by our cardiologists based on the similarity of treatments or differences in risks (see Figure 2). The highest values of the reward matrix are along its diagonal, associating full credit with correct classifier outputs, partial credit with incorrect classifier outputs, and no credit for labels and classifier outputs that are not captured in the weight matrix. Also, three similar classes (i.e., PAC and SVPB, PVC and VPB, CRBBB and RBBB) are scored as if they were the same class, so a positive label or classifier output in one of these classes is considered to be a positive label or classifier output for all of them. However, we did not change the labels in the training or test data to make these classes identical to preserve any institutional preferences or other information in the 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. This score is a generalized version of the traditional accuracy metric that awards full credit to correct outputs and no credit to incorrect outputs. To aid 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. A classifier that returns only positive outputs will typically receive a negative score, i.e., a lower score than a classifier that returns only negative outputs, which reflects the harm of false alarms.

Accordingly, this scoring metric was designed to award full credit to correct diagnoses and partial credit to misdiagnoses with similar risks or outcomes as the true diagnosis. The resources, populations, practices, and preferences of an institution all determine the ideal choice of the reward matrix W; the choice of W for the Challenge is just one example.

3. Results

We received a total of 1395 submissions of algorithms from 217 teams across academia and industry. The total number of successful entries was 707, with 397 successful entries during the unofficial phase of the Challenge and 310 successful entries during the official phase. During the official phase, we scored each entry on the validation set. The final score and

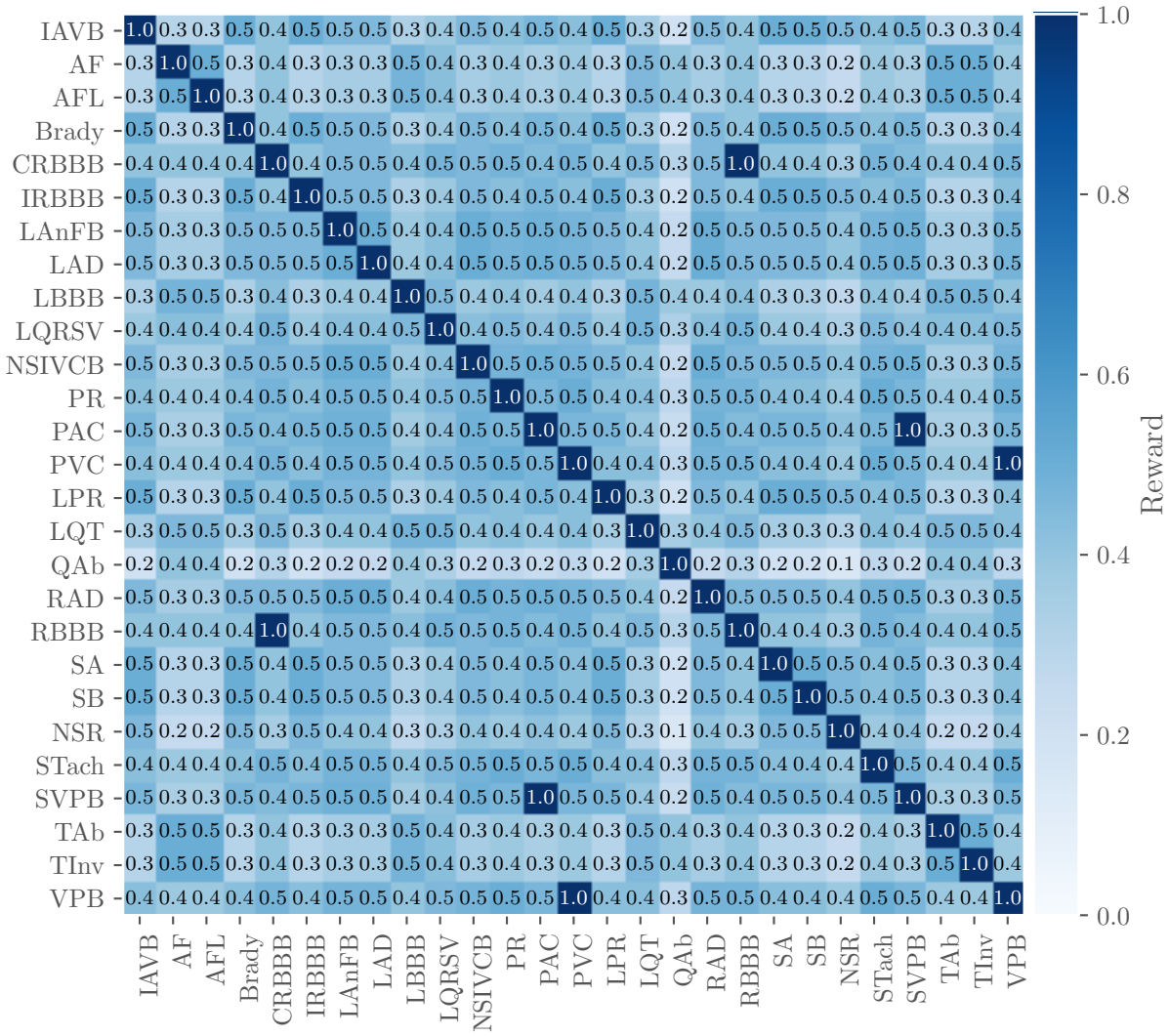

Figure 2. Reward matrix W for the diagnoses scored in the Challenge with rows and columns labeled by the abbreviations for the diagnoses in Table 3. Off-diagonal entries that are equal to 1 indicate similar diagnoses that are scored as if they were the same diagnosis. Each entry in the table was rounded to the first decimal place due to space constraints in this manuscript, but the shading of each entry reflects the actual value of the entry.

ranking were based on the test set. A total of 70 teams' codebases successfully ran on the test data. After final scoring, 41 teams were able to qualify for the final rankings [24]. Reasons for disqualification included: the training algorithm did not run, the trained model failed to run on the hidden undisclosed set (because of differences in sampling frequencies), the team failed to submit a preprint on time, the team failed to attend Computing in Cardiology (remotely or in person) and defend their work, and the team failed to submit their final article on time or address the reviewers' comments.

Figure 3 shows the performance of each team's final algorithm on the validation set, the hidden CPSC set, the hidden G12EC set, the hidden undisclosed set, and the test set. The line colors from red to blue indicates the higher to the lower scores on the test set. We observed the difference in score between each set. The higher scores were observed in the hidden CPSC dataset which contained a larger number of recordings in the training set as compared to the other three hidden dataset. We can also observe a drop on scores for the hidden undisclosed set for which none recording was included in the training or validation.

Figure 4 shows the rank performance of each team's final algorithm on the validation set, the hidden CPSC set, the hidden G12EC set, the hidden undisclosed set, and the test set. The points indicate the rank of each individual algorithm on each dataset. The line colors indicate the ranks on the test set.

On average, the Challenge scores dropped 47% from the hidden CPSC set to the hidden G12EC set and another 57% from the hidden G12EC set to the hidden undisclosed set. We observed an average drop of 50% from the validation score set to the test set.

The most common algorithmic approach was based on deep learning and convolutional neural networks (CNNs). However, over 70% of entries used standard clinical or hand-crafted features with classifiers such as support vector machines, gradient boosting, random forests, and shallow neural networks. The median training time was 6 hours, 49 minutes; nearly all approaches that required more than a few hours for training used deep learning frameworks.

4. Discussion

Figures 3 and 4 show how the performance of participant entries dropped on the hidden set. This under-performance on the hidden undisclosed dataset, and to a much lesser extent, on the hidden G12EC dataset could be due to the most teams over-trained on the CPSC data. The hidden CPSC data included fewer recordings than the other hidden sets. The poorer scores and ranks demonstrate the importance of including multiple sources for generalizability of the algorithms.

Deep learning approaches are one of the most popular machine learning techniques for classification problems, especially those of images. Some participants adapted previously developed algorithms for other classification problems and therefore this modification does not necessarily perform better than a custom-made machine learning algorithm.

It is important to note the class imbalance between the datasets, but the larger number and varying prevalences of diagnoses in different datasets represent the real-world problem of reading 12-lead ECGs in a clinical setting. In fact, most teams performed best on the CPSC dataset, which was the least representative dataset because it had fewer and more balanced diagnoses than the other datasets. Moreover, the scoring function that we proposed and used to evaluate the performance of each algorithm penalized classes non-uniformly, based on clinical importance. Balancing data would not only be artificial, but would provide

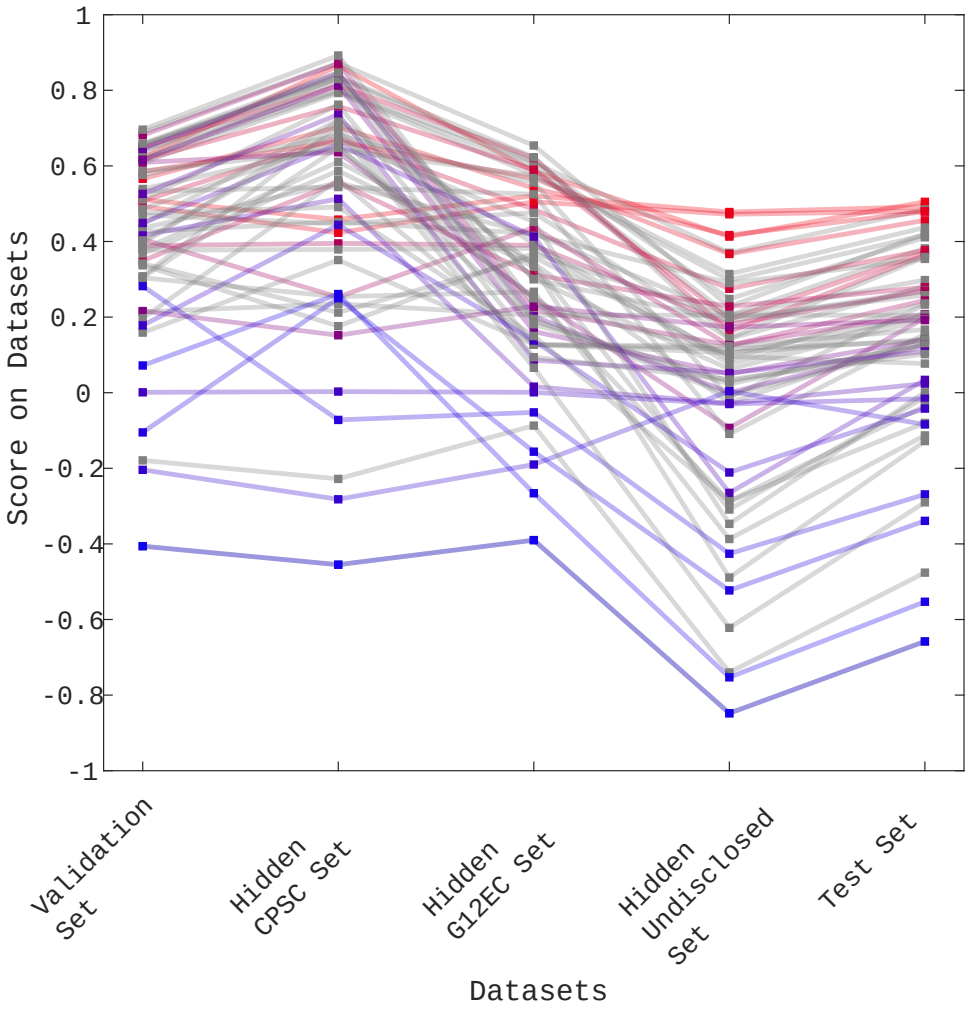

Figure 3. Scores of the final 70 algorithms that were able to completely evaluated on the validation set, the hidden CPSC set, the hidden G12EC set, the hidden undisclosed set, and the test set. The points indicate the score of each individual algorithm on each dataset, with the higher points showing algorithms with the highest scores on each dataset. The ranks on the test set are further indicated by color, with red indicating the best ranked algorithms and blue indicating the worst ranked algorithm on the test set.

an advantage to teams because the prevalence of the class would then be known. The Challenge was designed to discourage the use of *a priori* information on distributions, since the algorithms are likely to be used in a variety of unknown populations. Moreover, racial inequities and genetic variations are likely to lead to substantially different performances. While we cannot address that directly because the populations in the databases are not strictly matched, there is the potential to evaluate long-standing unknowns in algorithms that have been traditionally developed on predominately white, western hemisphere populations. (We note that the training and test data were matched as closely as possible for age, sex

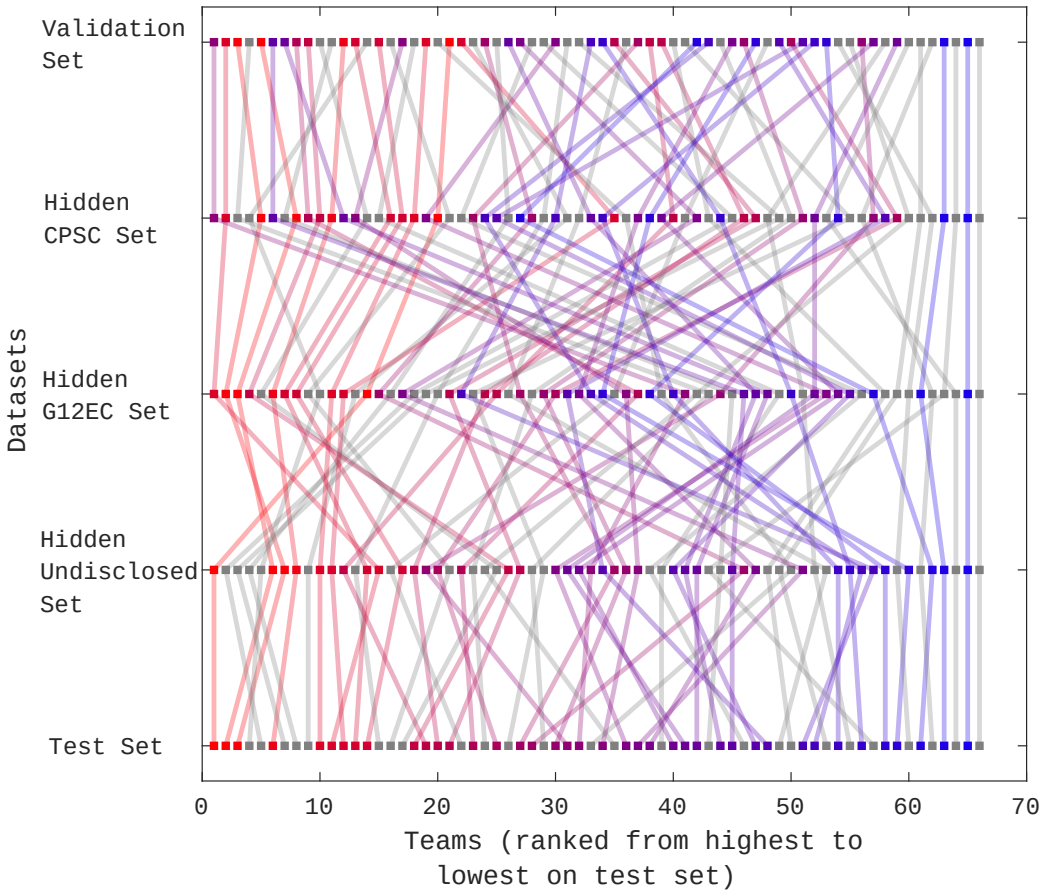

Figure 4. Ranks of the final 70 algorithms that were able to completely evaluated on the validation set, the hidden CPSC set, the hidden G12EC set, the hidden undisclosed set, and the test set. Lines from top to bottom indicate the rank of each individual algorithm on each dataset. Rank is indicated by color coding, with red indicating the best ranked algorithms, blue indicating the worst ranked algorithm on the test set, and gray indicating disqualified algorithms.

and diagnosis.) In future Challenges, we will re-use these databases and reveal per-class performances in the hidden test data to allow full evaluations of the algorithms in terms of class, age, race, and gender.

5. Conclusions

This article describes the world's largest open access database of 12-lead ECGs, together with a large hidden test database to provide objective comparisons. The data were drawn from three continents with diverse and distinctly different populations, encompassing 111 diagnoses with 27 diagnoses of special interest for the Challenge. Additionally, we introduced

a novel scoring matrix that rewards algorithms based on similarities between diagnostic outcomes, weighted by severity/risk.

The public training data and sequestered validation and test data provided the opportunity for unbiased and comparable repeatable research. Notably, 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 and generalizable body of work on the classification of electrocardiograms.

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 numbers 2R01GM104987-09 and R01EB030362 respectively, the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR002378, as well as the Gordon and Betty Moore Foundation, MathWorks, and AliveCor, Inc. under unrestricted gifts. Google also donated cloud compute credits for Challenge teams. GC has financial interests in Alivecor and Mindchild Medical. GC also holds a board position in Mindchild Medical. AIW holds equity and management roles in Ataia Medical and is supported by the NIGMS 2T32GM095442. AE receives financial support from the Basque Government through grant PRE_2019_2_0100. None of the aforementioned entities influenced the design of the Challenge or provided data for the Challenge. The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the above entities.

References

[1] E. Benjamin, P. Muntner, A. Alonso, M. Bittencourt, C. Callaway, A. Carson, A. Chamberlain, A. Chang, S. Cheng, S. Das, *et al.*, "Heart disease and stroke statistics-2019 update: A report from the American Heart Association.", *Circulation*, vol. 139, no. 10, e56, 2019.

[2] P. Kligfield, L. S. Gettes, J. J. Bailey, R. Childers, B. J. Deal, E. W. Hancock, G. Van Herpen, J. A. Kors, P. Macfarlane, D. M. Mirvis, *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*, vol. 49, no. 10, pp. 1109–1127, 2007.

[3] P. Kligfield, "The centennial of the Einthoven electrocardiogram", *Journal of Electrocardiology*, vol. 35, no. 4, pp. 123–129, 2002.

[4] I. Mozos and A. Caraba, "Electrocardiographic predictors of cardiovascular mortality", *Disease Markers*, vol. 2015, 2015.

[5] C. Gibbs, J. Thalamus, D. T. Kristoffersen, M. V. Svendsen, Ø. L. Holla, K. Heldal, K. H. Haugaa, and J. Hysing, "QT prolongation predicts short-term mortality independent of comorbidity", *EP Europace*, vol. 21, no. 8, pp. 1254–1260, 2019.

[6] J. L. Willems, C. Abreu-Lima, P. Arnaud, J. H. van Bemmel, C. Brohet, R. Degani, B. Denis, J. Gehring, I. Graham, G. van Herpen, *et al.*, "The diagnostic performance of computer programs for the interpretation of electrocardiograms", *New England Journal of Medicine*, vol. 325, no. 25, pp. 1767–1773, 1991.

[7] A. P. Shah and S. A. Rubin, "Errors in the computerized electrocardiogram interpretation of cardiac rhythm", *Journal of Electrocardiology*, vol. 40, no. 5, pp. 385–390, 2007.

[8] C. Ye, M. T. Coimbra, and B. V. Kumar, "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, pp. 1918–1921.

[9] A. H. Ribeiro, M. H. Ribeiro, G. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. Ferreira, C. R. Andersson, P. W. Macfarlane, M. Wagner Jr, *et al.*, "Automatic diagnosis of the 12-lead ECG using a deep neural network", *Nature Communications*, vol. 11, no. 1, pp. 1–9, 2020.

[10] T.-M. Chen, C.-H. Huang, E. S. Shih, Y.-F. Hu, and M.-J. Hwang, "Detection and classification of cardiac arrhythmias by a challenge-best deep learning neural network model", *Iscience*, vol. 23, no. 3, p. 100 886, 2020.

[11] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals", *Circulation*, vol. 101, no. 23, e215–e220, 2000.

[12] *PhysioNet Challenges*, `https://physionet.org/about/challenge/`, Accessed: 2020-02-07.

[13] *PhysioNet/Computing in Cardiology Challenge 2020*, `https://physionetchallenges.github.io/2020/`, Accessed: 2020-02-07.

[14] F. Liu, C. Liu, L. Zhao, X. Zhang, X. Wu, X. Xu, Y. Liu, C. Ma, S. Wei, Z. He, *et al.*, "An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection", *Journal of Medical Imaging and Health Informatics*, vol. 8, no. 7, pp. 1368–1373, 2018.

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

[16] R. Bousseljot, D. Kreiseler, and A. Schnabel, "Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet", *Biomedizinische TechnikBiomedical Engineering*, vol. 40, no. s1, pp. 317–318, 1995.

[17] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, "PTB-XL, a large publicly available electrocardiography dataset", *Scientific Data*, vol. 7, no. 1, pp. 1–15, 2020.

[18] M. A. Reyna, C. Josef, R. Jeter, S. P. Shashikumar, M. B. Westover, S. Nemati, G. D. Clifford, and A. Sharma, "Early prediction of sepsis from clinical data: The PhysioNet/Computing in Cardiology Challenge 2019", *Critical Care Medicine*, vol. 48, no. 2, pp. 210–217, 2019.

[19] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm", *IEEE Transactions on Biomedical Engineering*, vol. 32, no. 3, pp. 230–236, 1985.

[20] J. W. Waks, C. M. Sitlani, E. Z. Soliman, M. Kabir, E. Ghafoori, M. L. Biggs, C. A. Henrikson, N. Sotoodehnia, T. Biering-Sørensen, S. K. Agarwal, *et al.*, "Global electric heterogeneity risk score for prediction of sudden cardiac death in the general population: The atherosclerosis risk in communities (ARIC) and cardiovascular health (CHS) studies", *Circulation*, vol. 133, no. 23, pp. 2222–2234, 2016.

[21] E. A. Perez-Alday, Y. Li-Pershing, A. Bender, C. Hamilton, J. A. Thomas, K. Johnson, T. L. Lee, R. Gonzales, A. Li, K. Newton, *et al.*, "Importance of the heart vector origin point definition for an ECG analysis: The atherosclerosis risk in communities (ARIC) study", *Computers in Biology and Medicine*, vol. 104, pp. 127–138, 2019.

[22] A. N. Vest, G. D. Poian, Q. Li, C. Liu, S. Nemati, A. J. Shah, and G. D. Clifford, "An open source benchmarked toolbox for cardiovascular waveform and interval analysis", *Physiological Measurement*, vol. 39, no. 10, p. 105 004, 2018.

[23] A. N. Vest, G. D. Poian, Q. Li, C. Liu, S. Nemati, A. J. Shah, G. D. Clifford, and I. Sadiq, *PhysioNet-Cardiovascular-Signal-Toolbox*, version 1.0.2, Aug. 3, 2019. DOI: 10.5281/zenodo.1243111.

[24] *PhysioNet/Computing in Cardiology Challenge 2020 official results*, https://github.com/physionetchallenges/evaluation-2020/blob/master/Results/physionet_2020_official_scores.csv, Accessed: 2020-09-28.