Issues in the automated classification of multilead ECGs using heterogeneous labels and populations

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[6], Qiao Li[1], Ashish Sharma[1], Gari D Clifford[1,7]

[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 Electronics Technology, University of the Basque Country, Spain
[5]Department of Emergency Medicine, University of Michigan, USA
[6]Division of Cardiovascular Medicine, University of Michigan, USA
[7]Department of Biomedical Engineering, Georgia Institute of Technology, USA

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

Abstract. *Objective:* The standard twelve-lead electrocardiogram (ECG) is a widely used tool for monitoring cardiac function and diagnosing cardiac disorders. The development of smaller, lower-cost, and easier-to-use ECG devices may improve access to cardiac care in low-resource environments, but the diagnostic potential of these devices is unclear. This work explores these issues through a public competition: the PhysioNet/Computing in Cardiology Challenge 2021.

Approach: We sourced 131,149 twelve-lead ECG recordings from ten international sources. We posted 88,253 annotated recordings as public training data and withheld the remaining recordings as hidden validation and test data. We challenged teams to submit containerized, open-source algorithms for diagnosing cardiac abnormalities using various ECG lead combinations, including the code for training their algorithms. We designed and scored algorithms using an evaluation metric that captures the risks of different misdiagnoses for 30 conditions. After the Challenge, we implemented a semi-consensus voting model on all working algorithms.

Main results: A total of 68 teams submitted 1,056 algorithms during the Challenge, providing a variety of automated approaches from both academia and industry. The performance differences across the different lead combinations were smaller than the performance differences across the different test databases, showing that generalizability posed a larger challenge to the algorithms than the choice of ECG leads. A voting model improved performance by 3.5%.

Significance: The use of different ECG lead combinations allowed us to assess the diagnostic potential of reduced-lead ECG recordings, and the use of different data sources allowed us to assess the generalizability of algorithms to diverse institutions and populations. The submission of working, open-source code for both training and testing and the use of

a novel evaluation metric improved the reproducibility, generalizability, and applicability of the research conducted during the Challenge.

1. Introduction

Cardiovascular diseases are the leading cause of death worldwide [1]. Early treatment of cardiovascular disease can prevent serious cardiac events and improve outcomes, and the electrocardiogram (ECG) is a critical screening tool for a range of cardiac abnormalities, such as atrial fibrillation and ventricular hypertrophy [2], [3]. Moreover, recent advances in ECG technologies have allowed the development of smaller, lower-cost, and easier-to-use devices with the potential to improve access to cardiac screening and diagnoses in low-resource environments. However, due to the large variety of potential diagnoses, manual interpretation of ECGs is a time-consuming task that requires highly skilled and well trained personnel [4], [5].

Automatic detection and classification of cardiac abnormalities from ECGs can assist clinicians and ECG technicians, especially in the context of the ever-increasing number of recorded ECGs. Recent progress in machine learning techniques combined with standard clinical or handcrafted features has led to the development of algorithms that may identify cardiac abnormalities [6]–[10]. However, most published methods have only been developed in or tested on a single populations and/or relatively small and relatively homogeneous populations. Moreover, few publications provide software that allows for redeveloping and evaluating these models, so the work is often not scientifically repeatable or extendable. Additionally, most algorithms focus on identifying a limited number of cardiac issues that do not represent the complexity and difficulty of pathologies present in the ECG, and the reported performance for these cardiac issues does not reflect the cost of misclassification in a multi-class classification problem, where most outcomes have very different burdens on the individual. Finally, most publications that focus on automated approaches do not consider different lead combinations. The wide diagnostic potential of more accessible devices that use subsets of the standard twelve leads is largely unknown [11]–[13].

The PhysioNet/Computing in Cardiology Challenge 2021 provided an opportunity to address these issues by inviting teams to develop algorithms for diagnosing 30 cardiac abnormalities from various twelve-lead, six-lead, four-lead, three-lead, and two-lead ECG recordings [14]–[16]. The PhysioNet/Computing in Cardiology Challenges (recently renamed the 'George B. Moody PhysioNet Challenge') invite participants from academia and industry to address clinically important questions that have not been adequately addressed. PhysioNet co-hosts these annual Challenges in cooperation with Computing in Cardiology. Similarly to previous years, the Challenge ran over the course of nine months to allow for the development and refinement of the Challenge objective, data, and evaluation metric.

Previous Challenges addressed arrhythmia detection from ECGs: the 2017 Challenge considered the identification of atrial fibrillation in single-lead ECG recordings, and the 2020 Challenge considered the identification of 27 cardiac abnormalities from twelve-lead ECG recordings [14], [17]. However, this was the first PhysioNet/Computing in Cardiology Challenge that explored the diagnostic potential of reduced-lead ECGs for a variety of diagnoses [15], [16].

For the 2021 Challenge, we sourced 131,149 twelve-lead ECG recordings from ten databases from around the world; we shared two-thirds of the recordings as public training data and retained one-third of the recordings as hidden validation and test data, including recordings from two completely hidden databases. We designed an evaluation metric to capture the risks of different outcomes and misdiagnoses for 30 diagnoses, and we used it to evaluate the submitted algorithms on twelve-lead, six-lead, four-lead, three-lead, and two-lead versions of the ECG recordings. We required the teams to submit containerized, open-source code for training and testing their algorithms to ensure full scientific reproducibility.

2. Methods

2.1. Data

For the PhysioNet/Computing in Cardiology Challenge 2021, we sourced data from several countries across three continents. Each database contained electrocardiogram (ECG) recordings with diagnoses and basic demographic information. We use twelve-lead ECG recordings for the public training data and twelve-lead, six-lead, four-lead, three-lead, and two-lead versions of ECG recordings for the hidden validation and test data.

2.1.1. Challenge Data Sources We created ten databases for the Challenge [15]. Tables 1 and 2 describe the sources and splits (training, validation, and test) of these data. We publicly released the training data and the clinical ECG diagnoses for the training data, but we kept the validation and test data hidden to allow us to assess algorithmic generalizability and other common machine learning issues. For sources represented in multiple splits, the training, validation, and test data were matched as closely as possible to preserve the distributions of age, sex, and diagnoses.

(i) **Chapman-Shaoxing.** The Chapman-Shaoxing database is derived from the database in [18]. We posted this database as training data.

(ii) **CPSC.** The CPSC database is derived from the China Physiological Signal Challenge 2018 (CPSC 2018), held during the 7th International Conference on Biomedical Engineering and Biotechnology in Nanjing, China [19]. We posted the training data from CPSC 2018 as training data, and we split the test data from CPSC 2018 into validation and test data.

(iii) **CPSC-Extra.** The CPSC-Extra database contains unused data from CPSC 2018 [19]. We posted this database as training data.

(iv) **G12EC.** The G12EC Database is a new database representing a large population from the Southeastern United States. We split this database into training, validation, and test data.

(v) **INCART.** The INCART database is derived from the St. Petersburg Institute of Cardiological Technics (INCART) 12-lead Arrythmias Database [20]. We posted this database as training data.

(vi) **Ningbo.** The Ningbo database is derived from the database in [21]. We posted this database as training data.

(vii) **PTB.** The PTB database is derived from the Physikalisch-Technische Bundesanstalt (PTB) Database [22]. We posted this database as training data.

(viii) **PTB-XL.** The PTB-XL database is derived from the Physikalisch-Technische Bundesanstalt XL (PTB-XL) Database [23]. We posted this database as training data.

(ix) **UMich.** The UMich database is a database from the University of Michigan‡. We used this database as test data.

(x) **Undisclosed.** The Undisclosed database is a new database from an undisclosed American institution that is geographically distinct from the sources for the other databases. This database has never been publicly released, and it may never be publicly released. We used this database as test data.

2.1.2. Challenge Data Variables Each ECG recording was acquired in a hospital or clinical setting and included signal data, basic demographics data, and clinical diagnoses. The specifics of the data acquisition processes depended on the source of the databases and could vary from institution to institution. We have provided a summary of the clinical variables, and we encourage the readers to check the original publications for the details of each database and to cite them directly when used in their research.

We shared the full twelve-lead ECG signal data with the public training data, and we used twelve-lead, six-lead, four-lead, three-lead, and two-lead versions of the signal data for the hidden validation and test data. Table 3 summarizes the lead combinations included in the different versions of the hidden data. The choices of lead combination were made to test the notion of over-completeness. That is, in theory, the twelve-lead ECG represents a

‡ 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.

Database	Source	Locations(s)	Reference
Chapman-Shaoxing	Shaoxing People's Hospital	Shaoxing, Zhejiang, China	[18]
CPSC	CPSC 2018	Various Locations, China	[19]
CPSC-Extra	CPSC 2018	Various Locations, China	[19]
G12EC	Emory University Hospital	Atlanta, Georgia, USA	[14]
INCART	St. Petersburg Institute of Cardiological Technics	St. Petersburg, Russia	[20]
Ningbo	Ningbo First Hospital	Ningbo, Zhejiang, China	[21]
PTB	University Clinic Benjamin Franklin	Berlin, Germany	[22]
PTB-XL	Physkalisch Technische Bundesanstalt	Various Countries, Europe	[23]
UMich	University of Michigan	Ann Arbor, Michigan, USA	[15]
Undisclosed	N/A	USA	[14]

Table 1: Sources, locations, and references for each database in the Challenge.

Database	Total Patients	Training Set Recordings	Validation Set Recordings	Test Set Recordings	Total Recordings
Chapman-Shaoxing	10247	10247	0	0	10247
CPSC	Unknown	6877	1463	1463	9803
CPSC-Extra	Unknown	3453	0	0	3453
G12EC	15738	10344	5167	5161	20672
INCART	32	74	0	0	74
Ningbo	34905	34905	0	0	34905
PTB	262	516	0	0	516
PTB-XL	18885	21837	0	0	21837
UMich	N/A	0	0	19642	19642
Undisclosed	N/A	0	0	10000	10000
Total	N/A	88253	6630	36266	131149

Table 2: Numbers of patients and recordings in the training, validation, and test splits of the databases in the Challenge. The numbers of patients for the completely hidden test sets are not given.

spatial over-sampling of a three-dimensional dipole, and so, one might conclude that only three orthogonal leads are required to capture all activity. However, the heart is a not a point source dipole, and motion, physical distortion, and near-field electromagnetic effects come into play. It is well-known that precordial leads 'image' the ventricles far better than the limb leads, for example. For this Challenge, we asked 'will two do?' In other words, we wanted to know if two leads (I and II) would allow researchers to do as well as twelve leads, at least for the diagnoses being evaluated in this Challenge. However, we added three

extra categories, all of which are approximately equivalent to the same two leads, in order to test whether any subtle extra information was buried in these signals. We note that the addition of lead III should add no extra information, since it is part of 'Einthoven's Triangle' (lead I + lead III = lead II). In addition, the augmented leads (aVR, aVL and AVF), which are unipolar (in contrast to the bipolar limb leads) while providing different 'viewpoints', theoretically do not add any new information since they are also formed by recording the potential difference at the right arm, left arm and left leg. There are two reasons for including these apparently redundant leads. First, they provide clinicians with a deeper intuition, which may be a reflection of the varying amplitude resolution on each lead when the vectors are represented in a low-resolution format (e.g., paper, or on-screen). Second, the vectors are referenced to different grounds. There is an assumption that there is a stable voltage reference (with negligible variation during the cardiac cycle), known as the 'Wilson Central Terminal' (WCT), which is obtained by averaging the three active limb electrode voltages measured with respect to the return ground electrode. In the case of the augmented leads, the reference is provided by averaging limb electrodes ('Goldberger's Central Terminal'); see Figure 1. However, concerns have been raised by researchers about the ambiguous value and behavior of this reference voltage, which may lead to misdiagnoses or biases in certain circumstances [24]–[28]. Notably, Bacharova et al. [24] found significant diagnostic differences based on the reference in the case of ischemia. See Jin *et al.* [29] for further discussion on this point.

The precordial leads provide information in the transverse plane, in addition to the frontal plane information provided by the limb leads; see Figure 2. We chose to swap the augmented leads for lead V2 to provide information on the ventricular septum and anterior wall. For the two-lead case, we removed the precordial lead to stick with the most common configurations seen in ambulatory ECG monitoring. In summary, the full list of lead combinations are given in Table 3, where 12 leads provides the maximum information available in the recordings (and is the same as the previous year's Challenge [14]). The four-lead and four-lead combinations should be equivalent and provide the next largest amount of information (limb and precordial). Finally, the six-lead and two-lead combinations should be equivalent to each other and provide equivalent (and the least) information. Figure 3 illustrates the equivalence of the different lead combinations considered in the Challenge.

The sampling frequency of the signals varied from 250 Hz to 1 kHz, and the duration of the signals ranged from from 5 seconds to 30 minutes. The age and sex of the subjects were provided with most recordings. Table 4 provides a summary of the age, sex, and recording information for the Challenge databases, including splits of the CPSC and G12EC databases between the training, validation, and test sets.

The diagnoses or labels were provided with the training data; neither the teams nor their algorithms had access to the diagnoses for the validation and test data. The quality of the labels depended on the clinical or research practices, and the datasets included labels

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 3: Lead combinations used for the hidden validation and test sets in the Challenge.

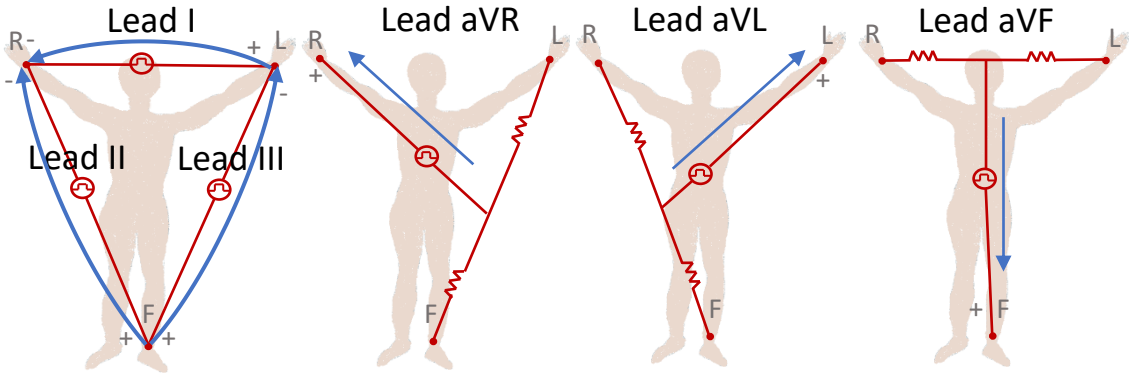


Figure 1: Illustrations of the three Einthoven limb leads (I, II, and III; left-most figure) and three circuits of the three Goldberger augmented leads (aVR, aVL, and aVF; three right-most figures). This figure is recreated from [30].

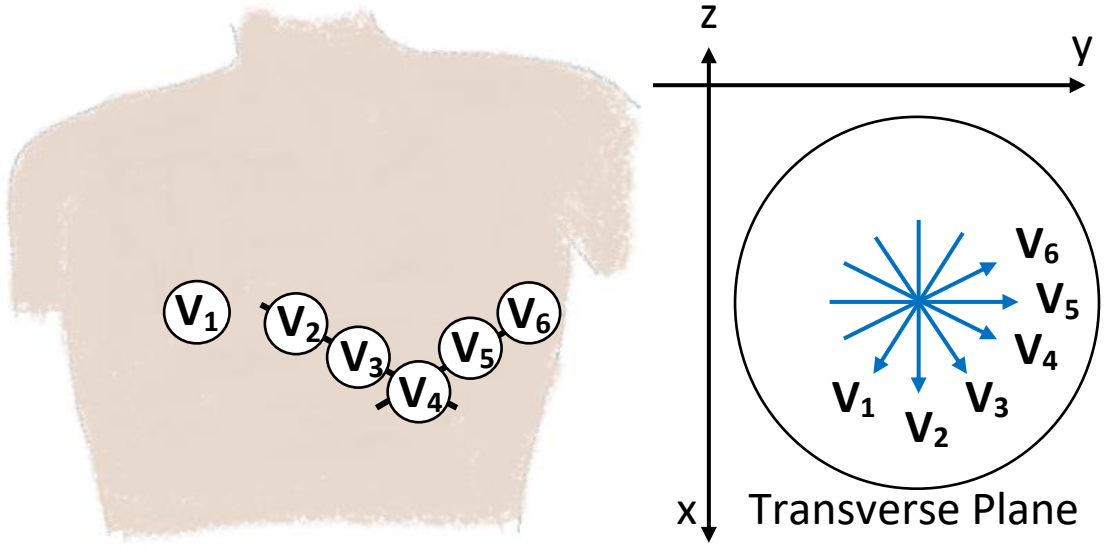


Figure 2: Illustration of the electrode locations for the six precordial leads, V1 to V6, on a human torso (left figure) and projections of these lead vectors centrally located on a transverse/horizontal plane (right figure). This figure is recreated from [30].

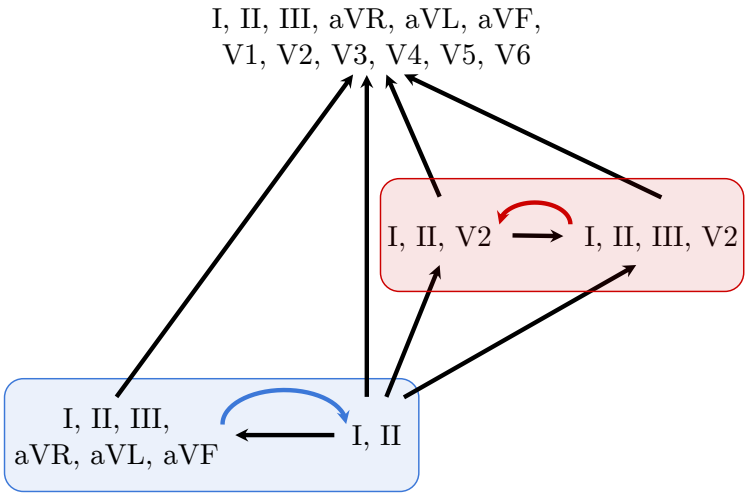


Figure 3: Hasse diagram for the lead combinations in the 2021 Challenge. The vertices are lead combinations, and the edges indicate that one lead combination is a proper subset of the other lead combination. The red and blue boxes and arrows indicate lead combinations that are functionally equivalent: I, II and I, II, III, aVR, aVL, aVF (blue) as well as I, II, V2 and I, II, III, V2 (red).

that were machine-generated, over-read by a single cardiologist, and adjudicated by multiple cardiologists. Human experts may have used different criteria for ECG interpretation for some abnormalities; see e.g., [31]. We did not correct for differences in labeling practices except to encode the diagnoses using approximate Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) codes for all of the datasets.

The data include 133 diagnoses or classes. All 133 diagnoses were represented in the training data, and a subset of these diagnoses were represented in the validation and test data. We evaluated the participant algorithms using 30 of the 133 diagnoses that were chosen by our cardiologists because they were relatively common, of clinical interest, and more likely to be recognizable from ECG recordings. Table 5 contains the list of the scored diagnoses for the Challenge with long-form descriptions, the corresponding SNOMED-CT codes, and abbreviations. While we only scored the algorithms using the diagnoses in Table 5 and Figure 4, we included all 133 classes in the data so that that participants could choose whether or not to use them with their algorithms.

All data were provided in MATLAB- and WFDB-compatible format [32]. Each ECG recording had a binary MATLAB v4 file for the ECG signal data and an associated plain text file in WFDB header format describing the recording and patient attributes, including the diagnosis or diagnoses 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 (HIPAA)-compliant identifiers for age and sex, (2) to encode the

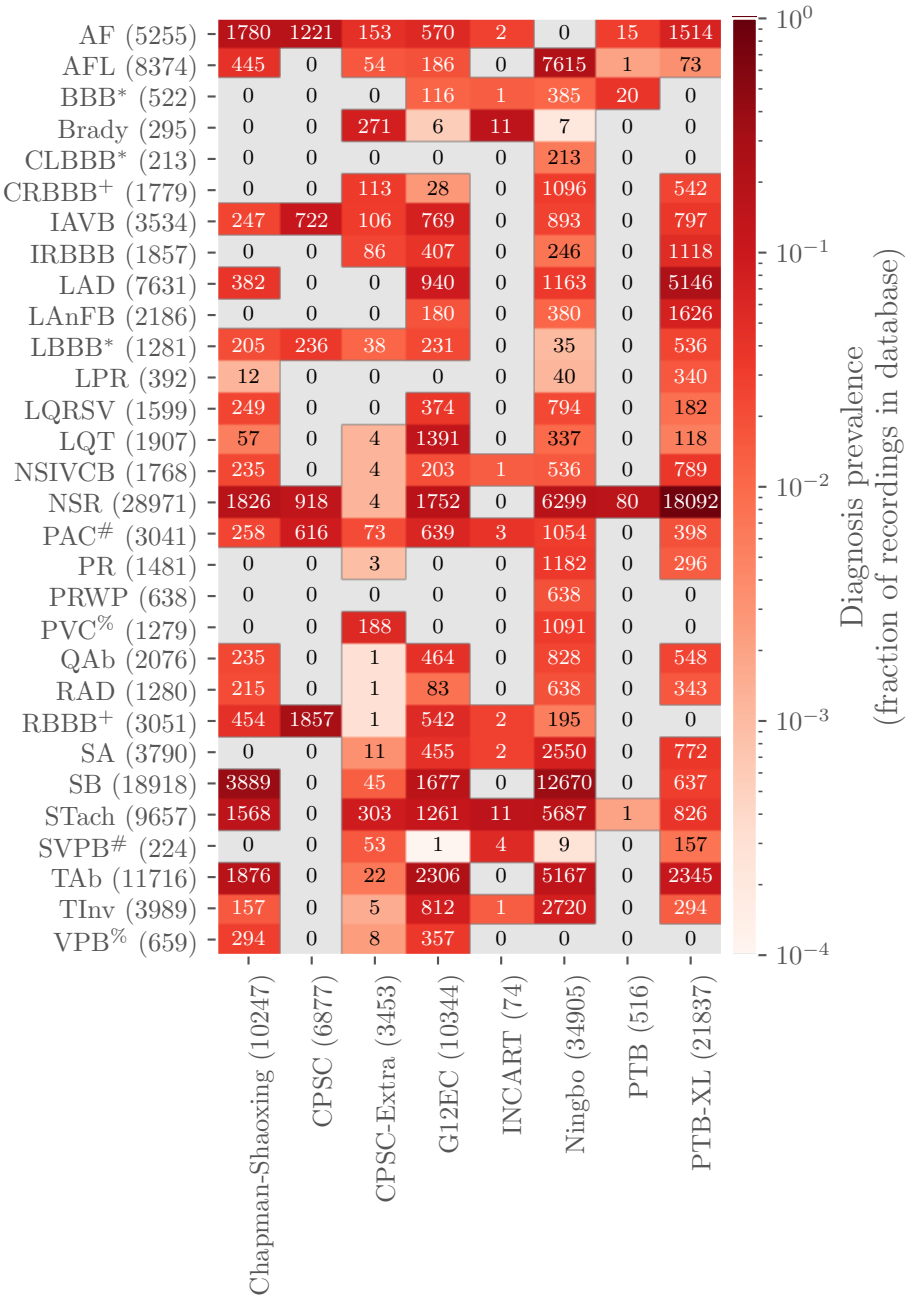


Figure 4: Numbers of recordings with each scored diagnosis in the training databases in the Challenge. Colors indicate the fraction of recordings with each scored diagnosis in each database, i.e., the total number of each scored diagnosis in a database normalized by the number of recordings in each database. Parentheses indicate the total numbers of records with a given label across the training data (rows) and the total numbers of recordings, including recordings without scored diagnoses, in each database (columns). The symbols $*$, $+$, $\#$, and $\%$ indicate that distinct diagnoses were scored as if they were the same diagnosis.

Database	Number of Recordings	Sampling Frequency (Hz)	Mean Duration (seconds)	Mean Age (years)	Sex (female/male)
Chapman-Shaoxing	10247	500	10.0	60.1	44%/56%
CPSC	9803	500	16.4	60.0	47%/53%
- CPSC training	6877	500	15.9	60.2	46%/54%
- CPSC validation	1463	500	17.2	58.9	49%/51%
- CPSC test	1463	500	17.5	60.0	47%/53%
CPSC-Extra	3453	500	15.9	63.7	47%/53%
G12EC	20672	500	10.0	60.5	46%/54%
- G12EC training	10344	500	10.0	60.5	46%/54%
- G12EC validation	5167	500	10.0	60.3	47%/53%
- G12EC test	5161	500	10.0	60.7	46%/54%
INCART	74	257	1800.0	56.0	46%/54%
Ningbo	34905	500	10.0	57.7	43%/56%
PTB	516	1000	110.8	56.3	27%/73%
PTB-XL	21837	500	10.0	59.8	48%/52%
UMich	19642	250 or 500	10.0	60.2	47%/53%
Undisclosed	10000	300	10.0	63.0	47%/53%

Table 4: 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 database in the Challenge. The CPSC and G12EC databases were represented in the training, validation, and test data and include summary statistics for the entire database and splits of the database.

diagnoses as approximate SNOMED CT codes, and (3) to store the signal data using 16-bit signed integers for WFDB format.

2.2. Challenge Objective

We asked the Challenge participants to design working, open-source algorithms for identifying cardiac abnormalities from standard twelve-lead and several reduced-lead ECG recordings. We required that the Challenge teams submit code both for training their models and for applying their trained models, which aided the 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 an expert-based evaluation metric that we designed for this year's Challenge.

2.2.1. Challenge Overview, Rules, and Expectations This year's Challenge was the 22$^{\text{nd}}$ PhysioNet/Computing in Cardiology Challenge [32]. Similarly to previous Challenges,

Diagnosis	Code	Abbreviation
Atrial fibrillation	164889003	AF
Atrial flutter	164890007	AFL
Bundle branch block	6374002	BBB
Bradycardia	426627000	Brady
Complete left bundle branch block	733534002	CLBBB
Complete right bundle branch block	713427006	CRBBB
1st degree AV block	270492004	IAVB
Incomplete right bundle branch block	713426002	IRBBB
Left axis deviation	39732003	LAD
Left anterior fascicular block	445118002	LAnFB
Left bundle branch block	164909002	LBBB
Prolonged PR interval	164947007	LPR
Low QRS voltages	251146004	LQRSV
Prolonged QT interval	111975006	LQT
Nonspecific intraventricular conduction disorder	698252002	NSIVCB
Sinus rhythm	426783006	NSR
Premature atrial contraction	284470004	PAC
Pacing rhythm	10370003	PR
Poor R wave progression	365413008	PRWP
Premature ventricular contractions	427172004	PVC
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 tachycardia	427084000	STach
Supraventricular premature beats	63593006	SVPB
T wave abnormal	164934002	TAb
T wave inversion	59931005	TInv
Ventricular premature beats	17338001	VPB

Table 5: Diagnoses, SNOMED CT codes, and abbreviations that were scored for the Challenge. CLBBB and LBBB, CRBBB and RBBB, PAC and SVPB, and PVC and VPB are distinct diagnoses, but we scored them as if they were the same diagnosis.

this year's Challenge had an unofficial phase and an official phase. The unofficial phase (December 24, 2020 to April 8, 2021) provided an opportunity to socialize the Challenge and seek discussions and feedback from teams about the data, scoring, and requirements. The unofficial phase allowed 5 scored entries from each team on the hidden validation data. After a short break, the official phase (May 1, 2021 to August 15, 2021) introduced additional training data. The official phase allowed 10 scored entries from each team on the hidden validation data. After the end of the official phase, we evaluated one algorithm from each team on the hidden test data 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 or sharing their work for the Challenge to preserve the diversity and uniqueness of the approaches to the problem posed by the Challenge.

2.2.2. Classification of ECGs We required teams to submit their code for both training and testing their models, including any code for processing or relabeling the data. We ran each team's training code on the training data to create a model, and we ran this model on the hidden validation and test sets. We ran the trained model on the recordings sequentially, instead of providing them all of the recordings at the same time, to apply them as realistically as possible. We then scored the outputs from the models. Figure 5 illustrates this computational pipeline.

We allowed teams to submit either MATLAB or Python implementations of their code. Participants containerized their code in Docker and submitted it in GitHub or Gitlab repositories. We downloaded their code and ran it in containerized environments on Google Cloud. We described the computational environment that we used to run entries more fully in [33]. We used virtual machines on Google Cloud with 10 virtual central processing units (vCPUs), 65 GB of random-access memory (RAM), and an optional NVIDIA T4 Tensor Core graphics processing unit (GPU). We imposed a 72 hour time limit for training with a GPU on the full training set and a 48 hour time limit for training without a GPU on the full training set. We used virtual machines on Google Cloud with 6 vCPUs, 39 GB of RAM, and an optional NVIDIA T4 Tensor Core GPU with a 24 hour time limit with or without a GPU for running the trained classifiers on each of the validation and test sets.

To aid teams, we shared example entries that we implemented in MATLAB and Python. The MATLAB example model was a multinomial logistic regression classifier that used age, sex, and the root-mean square of the signal for each ECG lead as features. The Python example model was a random forest classifier that also used age, sex, and the root-mean square of the signal for each ECG lead as features. We did not design these example models to be competitive but instead to provide working examples of how to read and extract features from the recordings that teams could easily run in several minutes on a personal computer.

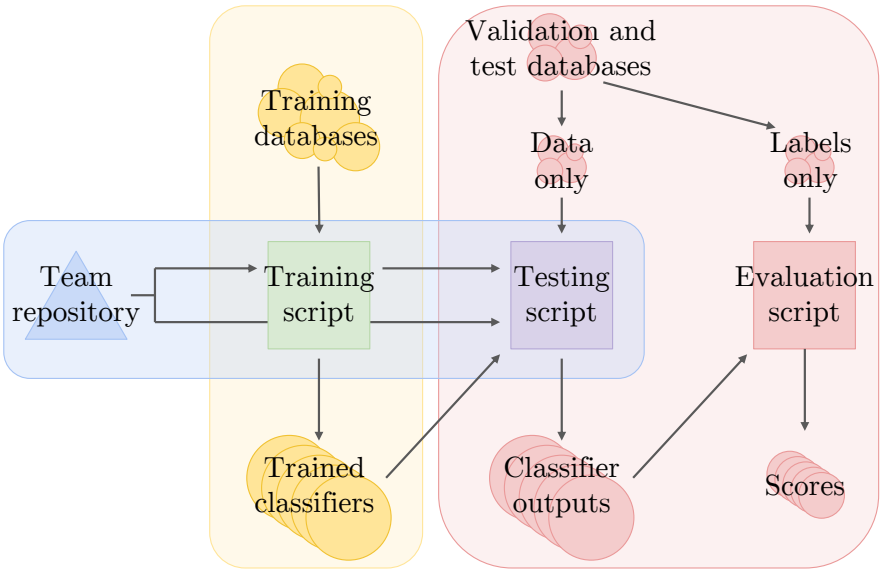


Figure 5: Computational pipeline for the 2021 Challenge. The vertices show the code, data, and results, and the edges show the relationships between the code, data, and results. Teams share their training and test scripts, which we run and score on the training, validation, and test sets; the test scripts running the trained models never see the labels for the validation and test sets.

2.2.3. Evaluation of Classifiers We introduced a scoring metric that awarded partial credit to misdiagnoses that resulted in similar outcomes or treatments as the true diagnoses that were given by our cardiologists. This scoring metric reflected the clinical reality some misdiagnoses are more harmful than others and should be scored accordingly.

First, let $C = \{c_i\}_{i=1}^{m}$ be a collection of m distinct diagnoses for a database of n recordings, and let $A = [a_{ij}]$ be a multi-class confusion matrix, where a_{ij} is a normalized number of recordings in a database that were classified as belonging to class c_i but actually belonged to class c_j, where c_i and c_j may be the same class or different classes.

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 $A = [a_{ij}]$ such that

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

where

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

where the quantity $|x_k \cup y_k|$ is the number of distinct classes with a positive label and/or classifier output for recording k. We normalized the counts in the confusion matrix because

each recording could have multiple true or predicted diagnoses, and we wanted incentivize teams to develop multi-class classifiers, but we did not want recordings with many diagnoses to dominate the scores that the algorithms received.

Next, let $W = [w_{ij}]$ be a reward matrix, 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 were defined by our cardiologists based on the similarity of treatments or differences in risks (see Figure 6). This matrix provided 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. Also, four pairs of similar classes (i.e., CLBBB and LBBB, CRBBB and RBBB, PAC and SVPB, and PVC and VPB) were scored as if they were the same class by assigning full credit to off-diagonal entries, so a positive label or classifier output in one diagnosis in the pair was considered to be a positive label or classifier output for both diagnoses in the pair. We did not change the labels in the training, validation, or test data to make these classes identical so that we could preserve any institutional preferences and other information in these data.

Finally, let

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

be 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 (ones in diagonal entries of the matrix) and no credit to incorrect outputs (zeros in off-diagonal entries of the matrix). To aid interpretability, we normalized this score so that a classifier that always output the true class or classes received a score of one and an inactive classifier that always output the sinus rhythm class received a score of zero, i.e.,

$$s_{\mathrm{N}} = \frac{s_{\mathrm{U}} - s_{\mathrm{I}}}{s_{\mathrm{T}} - s_{\mathrm{I}}}, \tag{4}$$

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

We used the same values of W for each algorithm and database, but each algorithm received different values of A and s_{N} for each database. For any particular lead combination, the algorithm with the highest value of s_{N} on the hidden test data for a specific lead combination won.

This scoring metric was designed to award full credit to correct diagnoses and partial credit to misdiagnoses with risks or outcomes that were similar to the true diagnosis. The resources, populations, practices, and preferences of an institution all affect how such a reward matrix W should be defined; the definition of this scoring metric from our cardiologists for the Challenge provides one such example.

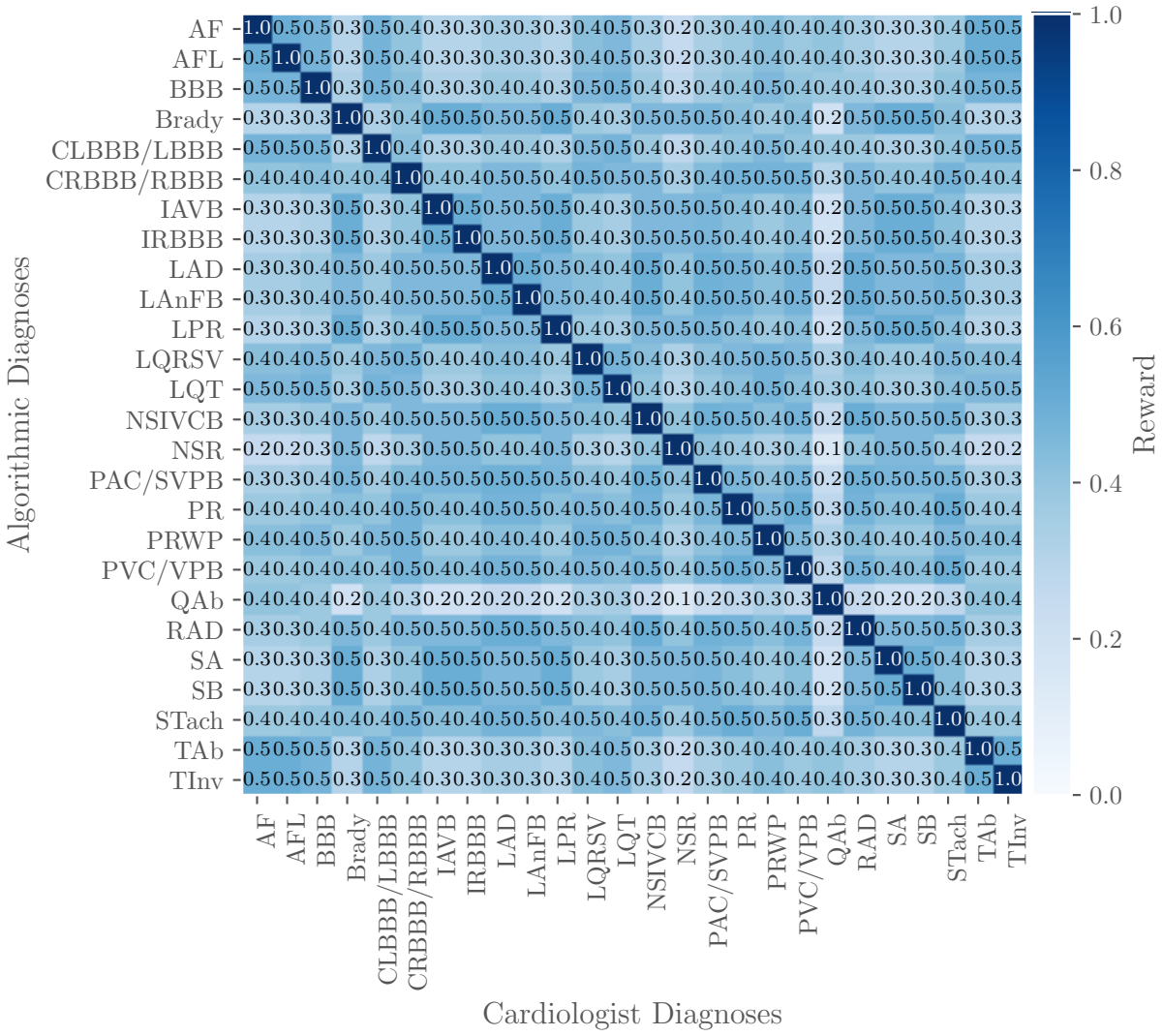


Figure 6: Reward matrix W for the 30 diagnoses scored in the Challenge. The rows and columns are the abbreviations for the ground-truth and predicted diagnoses in Table 5. Off-diagonal entries that are equal to 1 indicate similar diagnoses that are scored as if they were the same diagnosis. Each entry 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.

3. Results

3.1. Entries

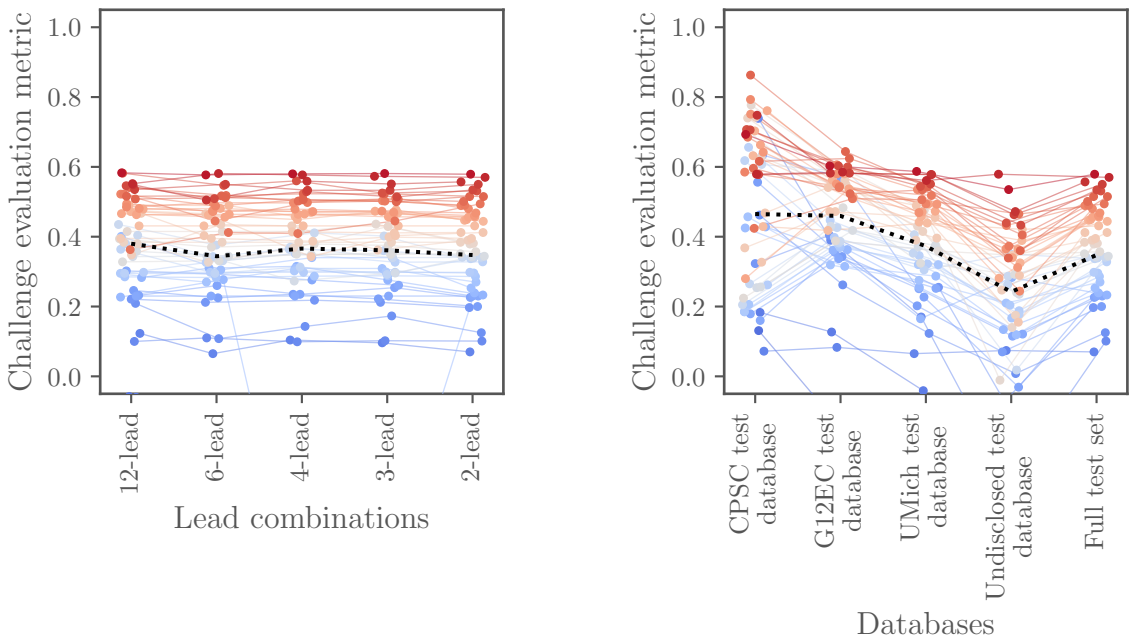
A total of 68 teams from academia and industry submitted 1,056 entries throughout the unofficial and official phases of the 2021 Challenge. During the unofficial and official phases, we trained the teams' models on the public training data and scored the trained models on the hidden validation set. After the end of the official phase, we scored a final entry from each team on the twelve-lead, six-lead, four-lead, three-lead, and two-leads versions of the hidden test set using the Challenge evaluation metric (4). The qualifying teams with the highest score on each version of the test set won the lead combination category for the Challenge.

There were 430 successful entries, including 165 successful entries during the unofficial phase and 265 successful entries during the official phase. There were 636 unsuccessful entries, including 234 unsuccessful entries and 309 unsuccessful entries that we were unable to train during the unofficial and official phases, respectively, highlighting the importance of sharing training code for the reproducibility of the models.

A total of 39 teams met all of the conditions for the Challenge, including the submission of algorithms that we could successfully run on the training, validation, and test databases [16]. There were several reasons for disqualification of teams including the following: the training code failed to train on the training data or simply loaded a pretrained model, the trained model failed to run on the validation or test data, the team failed to submit a preprint to Computing and Cardiology by the conference preprint deadline, the team failed to attend Computing in Cardiology either remotely or in person to present and defend their work. The algorithms from teams that met all of the conditions for the challenge are called 'official entries', and the other algorithms are called 'unofficial entries'.

Deep learning (DL) approaches were common (35 algorithms, 90% of the official entries), including convolutional neural networks (CNNs) in general and ResNet-based approaches in particular [34]. Although only 4 (10%) entries used other classifiers such as random forest classifiers [35], logistic regression (LogitBoost) [36], and XGBoost [37], [38], ten of the DL algorithms (about 30%) extracted hand-crafted features for their DL models. By combining hand-crafted extracted features with the DL models, these teams tried to generate more robust multi-label classification.

Across the different approaches, common trends appeared for the different lead combinations and data sources. Figure 7a shows the scores (the Challenge evaluation metric (4)) of the 39 official entries. It shows that the algorithms had similar performance (mean change of 2.6% for the Challenge evaluation metric) across different lead combinations on the different test databases. Figure 7b demonstrates the scores of the 39 official entries on the individual and full test sets. It shows that the algorithms have varied performance on different test databases, including noticeably lower performance for the completely hidden



(a) Algorithm scores for different lead combinations on the full hidden test set.

(b) Algorithm scores for the two-lead combination on different test databases.

Figure 7: Scores of the 39 official entries that we were able to evaluate on the hidden validation and test databases for the Challenge and met the other conditions for the Challenge. The points indicate the score of each individual algorithm on each dataset, with the higher points showing algorithms with the highest scores on each database. The ranks on the test set are further indicated by color, with red indicating the algorithms with the best rankings on the database and blue indicating the algorithms with the worst rankings on the database. The dashed line shows the median score for each lead combination or database.

test databases from sources for which no training data was provided, demonstrating the difficulty of generalizing to new databases.

Table 6 further quantifies the changes in performance for different lead combinations and databases using the median relative change in the Challenge evaluation metric, again showing small changes across different lead combinations and larger ones across different test databases.

Supplemental Table 1 provides a list of the 39 official entries that met all of the conditions for the Challenge, including their scores and ranks using the Challenge evaluation metric on the two-lead test set. This table also summarizes the libraries, model architectures, data processing, and optimization methods used by the algorithms, and it includes citations of

Lead combinations	CPSC test	G12EC test	UMich test	Undisclosed test	Full test
12-lead	-10%	-3%	-14%	-39%	-17%
6-lead	-13%	-3%	-14%	-36%	-19%
4-lead	-13%	-3%	-12%	-37%	-17%
3-lead	-14%	-3%	-13%	-38%	-19%
2-lead	-13%	-3%	-13%	-41%	-18%

Table 6: The median relative change in the Challenge evaluation metric from the full validation set to the individual test sets and full test set for the five different lead combinations.

the Computing and Cardiology papers for more information about the methods.

3.2. Voting algorithm

We developed and applied a naïve voting approach to combine individual algorithms into a single algorithm. This approach leveraged the different strengths of the individual algorithms while outperforming any single individual algorithm. In particular, we built a simple model that considered the classifier outputs of k different models that returned a positive vote for a diagnosis if at least αk different algorithms returned a positive vote for that diagnosis. This approach provided a majority votes-like voting model that used the data to determine a more optimal amount of consensus between methods than a simple majority.

We chose the voting model parameters k and α as follows for the two-lead versions of the data and the Challenge evaluation metric; the same approach applies to other versions of the data and other scoring schemes. We first ranked the 39 official entries from highest to lowest performance according to the Challenge evaluation metric (4) on the training set, and we defined a voting model using the top k algorithms that returned a positive vote for a diagnosis if at least αk different algorithms returned a positive vote for that diagnosis. We found that $k = 10$ and $\alpha = 0.4$ resulted in the voting model with the highest score on the validation set; Figure 8 illustrates this parameter search. We then ranked the algorithms from highest to lowest performance on the validation set, and we defined a voting model using the $k = 10$ highest scoring algorithms that returned a positive vote for a diagnosis if at least $\alpha k = 0.4 \cdot 10 = 4$ algorithms returned a positive vote for that diagnosis. We finally applied this model to the test data.

This voting model received a Challenge evaluation metric of 0.60 on the two-lead version of the test data, outperforming the highest-performing individual algorithm, which received a Challenge evaluation metric of 0.58. Notably, this voting model used classifier outputs from an algorithm that was ranked 24th on the test data; some individual algorithms had lower overall performance but higher performance on some diagnoses or patient groups, which the voting model was able to utilize.

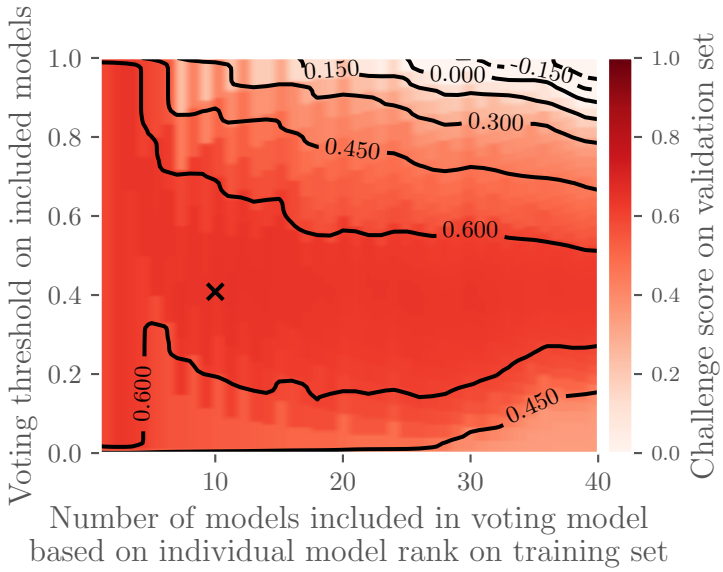


Figure 8: Parameter selection for voting model. The top k models (x-axis) were included in the voting model based on performance on the training set that returns a positive vote if at least αk models (y-axis) returned positive votes. The heat map shows the performance of the resulting voting model using the Challenge evaluation metric on validation data, and the x marks the parameters that define the voting model with the highest performance on the validation data; these parameters define the voting model to be applied to the test set. This figure shows performance for two-lead versions of the data with similar results on other datasets.

4. Follow-up Entries from the 2021 Challenge

As with most Challenges, we provided the community with another chance to evaluate their code on the test data for the 2021 Challenge. For this opportunity, we required the authors to submit updated code (or entirely new code) and a preprint describing the novelty of their updated or new approach that they planned to submit to the special issue containing this article. Tables 7, 8, and 9 provide the updated metrics for the twelve-lead, three-lead and two-lead of the full test set. The metrics include the area under the receiver operating characteristic curve (AUROC), the area under the precision recall curve (AUPRC), accuracy (defined here as the fraction of correctly classified recordings), F-measure, and the Challenge evaluation metric. We received 33 entries and successfully ran 13 entries.

In general, these post-Challenge entries improved on the performance of the original entries, and they again showed small changes in performance across the different lead combinations and larger changes across the different test databases from different sources.

Rank	Team [Reference]	AUROC	AUPRC	Accuracy	F-measure	Challenge metric [± original score]
-	Voting	N/A_a	N/A_a	0.32	0.49	0.62 [N/A_b]
1	CeZIS [39]	0.87	0.53	0.34	0.51	0.62 [+0.10]
2	ISIBrnoAIMT [40]	0.87	0.45	0.29	0.41	0.59 [+0.01]
3	HeartBeats [41]	0.94	0.52	0.21	0.44	0.57 [-0.01]
3	DrCubic [42]	0.93	0.51	0.26	0.45	0.57 [+0.08]
5	AIRCAS_MEL1 [43]	0.91	0.47	0.24	0.42	0.52 [+0.14]
6	iadi-ecg [44]	0.87	0.45	0.27	0.41	0.48 [+0.00]
6	SMS+1 [45]	0.87	0.36	0.21	0.32	0.48 [-0.04]
8	skylark [46]	0.83	0.30	0.02	0.26	0.39 [+0.03]
8	itaca-UPV [47]	0.82	0.32	0.05	0.29	0.39 [+0.05]
10	Revenger [48]	0.83	0.45	0.37	0.43	0.38 [N/A_c]
11	Medics [49]	0.74	0.26	0.07	0.26	0.36 [N/A_c]
12	Biomedic2ai [50]	0.79	0.35	0.28	0.32	0.24 [-0.12]
13	WEAIT [51]	0.52	0.07	0.01	0.10	-0.12 [+0.50]

Table 7: The metrics and ranks of the teams for follow-up entries to the special issue on the twelve-lead version of the full test set, including teams' papers and changes in score ("±") from their original entries. "AUROC" is area under the receiver operating characteristic curve, and "AUPRC" is area under the precision recall curve. 'Voting' indicates the voting algorithm described in Section 3.2, which used a subset of the algorithms in the Challenge, rather than any follow-up entries. "N/A" denotes "not available": N/A_a indicates that the voting model did not report numerical outputs, N/A_b indicates that the voting model was run only once, and N/A_a indicates failed original entries on the test set during the Challenge.

5. Discussion

While the 2021 Challenge sought to assess the diagnostic potential of reduced-lead ECGs, the real-world issues of using clinical data proved to be more of an obstacle to the automatic detection of cardiac abnormalities than the different choices of lead sets, highlighting the challenge of generalizing to datasets from new settings with different data collection procedures and populations. However, this common issue represents a diversity of approaches to automatically identifying cardiac abnormalities from reduced-lead ECGs.

Table S1 summarizes the 39 algorithms submitted by teams that satisfied all of the requirements of the 2021 Challenge. It shows that deep learning (DL) approaches were common for the 2021 Challenge, which was also true of the 2020 Challenge and reflects recent trends in ECG signal processing and arrhythmia detection. Some participants adopted algorithms from other applications, but they did not necessarily perform better than custom-made machine learning algorithms. The performance of these algorithms showed that the custom model architectures, custom optimization techniques, and deliberate attempts to

Rank	Team [Reference]	AUROC	AUPRC	Accuracy	F-measure	Challenge metric [± original score]
1	CeZIS [39]	0.87	0.53	0.34	0.50	0.61 [+0.09]
-	Voting	N/A_a	N/A_a	0.32	0.47	0.60 [N/A_b]
2	ISIBrnoAIMT [40]	0.87	0.44	0.28	0.40	0.60 [+0.02]
3	HeartBeats [41]	0.93	0.50	0.21	0.46	0.58 [+0.05]
4	Dr_Cubic [42]	0.93	0.51	0.26	0.45	0.56 [+0.05]
5	AIRCAS_MEL1 [43]	0.91	0.45	0.23	0.41	0.51 [+0.08]
6	iadi-ecg [44]	0.87	0.44	0.29	0.39	0.46 [-0.01]
7	Revenger [48]	0.86	0.45	0.38	0.42	0.40 [+0.07]
8	skylark [46]	0.83	0.29	0.02	0.26	0.39 [-0.06]
9	itaca-UPV [47]	0.84	0.33	0.04	0.28	0.38 [+0.08]
10	SMS+1 [45]	0.82	0.33	0.18	0.30	0.36 [-0.14]
11	Medics [49]	0.79	0.29	0.08	0.27	0.32 [N/A_c]
12	Biomedic2ai [50]	0.79	0.33	0.29	0.27	0.19 [-0.10]
13	WEAIT [51]	0.52	0.07	0.01	0.10	-0.12 [+0.50]

Table 8: The metrics and ranks of the teams for follow-up entries to the special issue on the three-lead version of the full test set, including teams' papers and changes in score ("±") from their original entries. "AUROC" is area under the receiver operating characteristic curve, and "AUPRC" is area under the precision recall curve. 'Voting' indicates the voting algorithm described in Section 3.2, which used a subset of the algorithms in the Challenge, rather than any follow-up entries. "N/A" denotes "not available": N/A_a indicates that the voting model did not report numerical outputs, N/A_b indicates that the voting model was run only once, and N/A_a indicates failed original entries on the test set during the Challenge.

generalize to new databases can help to provide better diagnostic outcomes. The algorithms developed by the Challenges teams widely used convolutional neural networks (CNN) and ResNet deep neural networks with different architectures and customized models. An example of a customized model was a channel self-attention-based model developed by team `cardiochallenger`, which used an ensemble inception and residual architecture with a genetic algorithm to optimize thresholds for each class for maximizing the Challenge evaluation metric [52].

For handling different signal characteristics across different datasets in the training set, including different sampling rates, gains, signal quality, and signal lengths of the ECG recordings, many algorithms applied preprocessing steps to the ECG signals. The preprocessing steps included resampling, normalization, peak correction, filtering, noise reduction, and/or discarding the beginnings and ends of the signals [53]–[55]. Different normalization techniques were applied, including z-score normalization [54], [56] and min-max scaling [57]. `snu_adsl` reported that standardization did not necessarily improve the

Rank	Team [Reference]	AUROC	AUPRC	Accuracy	F-measure	Challenge metric [± original score]
-	Voting	N/A_a	N/A_a	0.30	0.46	0.60 [N/A_b]
1	CeZIS [39]	0.87	0.52	0.33	0.49	0.59 [+0.07]
1	ISIBrnoAIMT [40]	0.87	0.43	0.27	0.39	0.59 [+0.00]
3	HeartBeats [41]	0.92	0.50	0.20	0.42	0.57 [+0.04]
4	Dr_Cubic [42]	0.92	0.50	0.25	0.44	0.55 [+0.07]
5	AIRCAS_MEL1 [43]	0.89	0.44	0.22	0.40	0.50 [+0.12]
5	SMS+1 [45]	0.86	0.36	0.26	0.32	0.50 [+0.01]
7	iadi-ecg [44]	0.87	0.42	0.27	0.38	0.45 [-0.01]
8	skylark [46]	0.81	0.27	0.03	0.26	0.39 [-0.10]
9	Medics [49]	0.78	0.28	0.08	0.27	0.38 [N/A_c]
10	itaca-UPV [47]	0.82	0.31	0.04	0.26	0.37 [+0.03]
11	Revenger [48]	0.84	0.43	0.35	0.40	0.35 [+0.02]
12	Biomedic2ai [50]	0.80	0.33	0.28	0.29	0.26 [-0.08]
13	WEAIT [51]	0.62	0.17	0.01	0.17	-0.08 [+0.54]

Table 9: The metrics and ranks of the teams for follow-up entries to the special issue on the two-lead version of the full test set, including the teams' papers and changes in score ("±") from their original entries. "AUROC" is area under the receiver operating characteristic curve, and "AUPRC" is area under the precision recall curve. 'Voting' indicates the voting algorithm described in Section 3.2, which used a subset of the algorithms in the Challenge, rather than any follow-up entries. "N/A" denotes "not available": N/A_a indicates that the voting model did not report numerical outputs, N/A_b indicates that the voting model was run only once, and N/A_a indicates failed original entries on the test set during the Challenge.

performance of their algorithm, but they implemented it in their preprocessing step in case the unseen dataset had unexpected characteristics.

One of the common preprocessing steps of the algorithms (implemented by 25 teams, 64% of the official algorithms) was filtering the signals using different techniques such as the Butterworth bandpass filter with a bandwidth between 1-45 Hz. [58] used a finite impulse response bandpass filter with a bandwidth between 3-45 Hz [59].

Some of the algorithms investigated the quality of the signals or used data augmentation. For instance, `HaoWan_AIeC` assessed the quality of each lead and created a mask for low-quality leads [60]. They also applied data augmentation by randomly cropping signals and randomly generating masks [60]. `HeartlyAI` applied different augmentation techniques such as cut-out, adding different types of noise, and allowing dropout of individual or groups of ECG channels [61].

Many algorithms segmented the signals into windows during preprocessing. For example, `Biomedic2ai` segmented the signals into 5-second windows with a stride of one

second for a 4-second overlap for adjacent signals [62]. snu_adsl selected a random window with a width of 13.3 seconds and zero-padded ECG signals shorter than 13.3 seconds at the end of the signal [58]. prna set a fixed window width of 15.36 seconds, allowing the signal to be split into divisible segments sizes and zero-padding the ends of the signals as needed [59]. iadiecg extracted the middle of recordings that were longer than 10 seconds and zero-padded both sides of recordings that were shorter than 10 seconds and normalized the signals so that they had zero mean and unit variance [63]. Although these algorithms were not the highest performing algorithms, they were among the top half of the entries and obtained a Challenge evaluation metric between 0.44 and 0.46 on the two-lead full test set.

Some participants decided not to train their models on some of the training databases. The scores of the winning algorithms shows that inclusion of all of the available data may lead to better generalization on the unseen test data [52], [54], [64].

For addressing differences in data collection practices from different sources, teams applied different methods to improve generalization. For example, HaoWan_AIeC adopted MixStyle to use feature-based augmentation to generalize to different domains [60]. Team DSAIL_SNU used the WRN model architecture with 14 convolution/dense layers and a widening factor of 1 and attempted to improve generalization by using constant-weighted cross-entropy loss, additional features, MixUp augmentation, a squeeze/excitation block, and a OneCycle learning rate scheduler [65]. Another team, NIMA, whose entry was among the top three algorithms, used spatial dropouts and average pooling between each layer of two separate deep CNNs to reduce overfitting and model complexity [64].

Eleven entries (about 30% of the official teams) used a binary cross-entropy loss function for multi-label classification, but custom loss functions for this problem also helped to improve classification performance. Team ISIBrno-AIMT, the winning algorithm, optimized a ResNet architecture with a multi-head attention mechanism using a mixture of a binary cross-entropy loss function, a custom loss function that provided a differentiable approximation of Challenge evaluation metric, and an evolutionary optimization loss function that attempted to estimate the optimal probability threshold for each class [54]. BUTTeam noticed that using the Challenge evaluation metric as a loss function seemed to be unstable and lead to sub-optimal results [66]. Their approach was to first train their model with weighted cross-entropy (WCE) loss and then retrain it with the Challenge evaluation metric and a decaying learning rate [66]. Another example of a custom loss-function was a weighted, generalized softmax loss function with quadratic differences by AADAConglomerate [53].

Although 90% of the 39 official entries used DL models, about 40% of the algorithms combined hand-crafted features with their DL models [62], [67]. Team PhysioNauts was among the unofficial teams that used a ResNet model with a squeeze and excitation module with handcrafted and DL features and used a grid search and the Nelder-Mead method to optimize the Challenge evaluation metric [68]. UMCU used an adaptive pooling layer to combine the features over the temporal dimension, after which a linear layer created the final

output [69].

Class imbalance was another significant issue for classification, and the larger number and varying prevalence rates of the diagnoses from different sources represented the real-world problem of clinical diagnosis. Many algorithms applied different methods to correct for class imbalance. `UMCU` weighted each class by dividing the maximum number of positive samples from any class by the number of positive samples from the weighted class and used a threshold of 0.5 for prediction; all values greater than 0.5 were positive, and all below 0.5 were negative [69]. Most teams performed best on the CPSC dataset, but it was the least representative dataset due to having fewer and more balanced diagnoses than the other datasets.

Almost all teams performed similarly on the different lead combination with average scores change in the Challenge evaluation metric of less than 2% from the twelve-lead to two-lead versions of the test data, which could be interpreted as responding to the Challenge question of "Will two do?" with the answer "Yes, two can do!". Although the average scores change between different lead combinations is relatively negligible, performances varied across different diagnoses and different data sources, suggesting that better data processing and generalization techniques are required for better performance on unseen datasets.

6. Conclusions

This article explores the diagnostic potential of automated approaches for reading standard twelve-lead and various reduced-lead ECG recordings. While most algorithms had similar overall performance across the different lead combinations considered during the Challenge, including two-lead ECGs and standard twelve-lead ECGs, performance varied more widely across different diagnoses and on data from different institutions. We found no definitive evidence that over-complete lead systems provided any additional diagnostic power, although without underlying knowledge of the electronic circuitry of the individual systems used to capture the ECGs, it is impossible to be sure if inter-database differences in performances as a function of lead combinations are due to variations in: population phenotypes; clinical practice behaviors; hardware filters; manufacturer software pre-processing before data storage; or circuit configuration differences (such as derivation of the WCT). Moreover, the lack of consistent differences between lead configurations across all data is confined to the rhythms explored in this work. Other conditions, such as much more subtle ST changes perhaps, may yield different results.

Most importantly, this article describes the world's largest open-access database of twelve-lead ECG recordings along with large hidden validation and test databases of twelve-lead and various reduced-lead recordings, providing unbiased and repeatable research on the diagnostic potential of automated approaches for reduced-lead diagnoses. The data were drawn from several countries across three continents with diverse and distinctly

different populations, encompassing 133 diagnoses with 30 diagnoses of special interest for the Challenge. Additionally, we introduced a novel scoring matrix that captures the similarities between and risks of different diagnostic outcomes. Finally, we supported the development of a large corpus of open-source, repeatable, and diverse algorithmic approaches for classifying full-lead and reduced-lead ECG recordings. The algorithms and data can provide benchmarks for the field and help push beyond the current theme of applying machine learning to large single-center databases, which in our experience are unlikely to generalize across populations.

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. GC has financial interests in Alivecor, LifeBell AI and Mindchild Medical. GC also holds a board position in LifeBell AI and Mindchild Medical. AE receives financial support from the Spanish Ministerio de Ciencia, Innovacion y Universidades through grant RTI2018-101475-BI00, jointly with the Fondo Europeo de Desarrollo Regional (FEDER), and by the Basque Government through grant IT1229-19. 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] S. S. Virani, A. Alonso, H. J. Aparicio, E. J. Benjamin, M. S. Bittencourt, C. W. Callaway, A. P. Carson, A. M. Chamberlain, S. Cheng, F. N. Delling, M. S. V. Elkind, K. R. Evenson, J. F. Ferguson, D. K. Gupta, S. S. Khan, B. M. Kissela, K. L. Knutson, C. D. Lee, T. T. Lewis, J. Liu, M. S. Loop, P. L. Lutsey, J. Ma, J. Mackey, S. S. Martin, D. B. Matchar, M. E. Mussolino, S. D. Navaneethan, A. M. Perak, G. A. Roth, Z. Samad, G. M. Satou, E. B. Schroeder, S. H. Shah, C. M. Shay, A. Stokes, L. B. VanWagner, N.-Y. Wang, C. W. Tsao, and American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee, "Heart disease and stroke statistics – 2021 update: A report from the American Heart Association", *Circulation*, vol. 143, no. 8, e254–e743, 2021.

[2] P. Kligfield, "The centennial of the Einthoven electrocardiogram", *J. Electrocardiol.*, vol. 35, no. 4, pp. 123–129, 2002.

[3] 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.

[4] J. L. Willems, C. Abreu-Lima, P. Arnaud, J. Vanbemmel, C. Brohet, R. Degani, B. Denis, J. Gehring, I. Graham, G. vanHerpen, H. C. Machado, P. W. Macfarlane, J. Michaelis, S. Moulopoulos, P. Rubel, and C. Zywietz, "The diagnostic performance of computer programs for the interpretation of electrocardiograms", *The New England Journal of Medicine*, vol. 325, pp. 1767–1773, 1991.

[5] A. Shah and S. Rubin, "Errors in the computerized electrocardiogram interpretation of cardiac rhythm", *Journal of Electrocardiology*, vol. 40, pp. 385–90, Sep. 2007. DOI: 10.1016/j.jelectrocard.2007.03.008.

[6] 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.

[7] B. Mohammadzadeh-Asl and S. K. Setarehdan, "Neural network based arrhythmia classification using heart rate variability signal", in *2006 14th European Signal Processing Conference*, 2006, pp. 1–4.

[8] S. L. Oh, E. Y. Ng, R. S. Tan, and U. R. Acharya, "Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats", *Computers in Biology and Medicine*, vol. 102, pp. 278–287, 2018, ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2018.06.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482518301446.

[9] S. G, S. K P, and V. R, "Automated detection of cardiac arrhythmia using deep learning techniques", *Procedia Computer Science*, vol. 132, pp. 1192–1201, 2018, International Conference on Computational Intelligence and Data Science, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2018.05.034. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S187705091830766X.

[10] V. D. Nagarajan, S.-L. Lee, J.-L. Robertus, C. A. Nienaber, N. A. Trayanova, and S. Ernst, "Artificial intelligence in the diagnosis and management of arrhythmias", *European Heart Journal*, vol. 42, no. 38, pp. 3904–3916, Aug. 2021, ISSN: 0195-668X. DOI: 10.1093/eurheartj/ehab544. eprint: https://academic.oup.com/eurheartj/

article-pdf/42/38/3904/40526393/ehab544.pdf. [Online]. Available: https://doi.org/10.1093/eurheartj/ehab544.

[11] H. R. Aldrich, N. B. Hindman, T. Hinohara, M. G. Jones, J. Boswick, K. L. Lee, W. Bride, R. M. Califf, and G. S. Wagner, "Identification of the optimal electrocardiographic leads for detecting Acute Epicardial Injury in Acute Myocardial Infarction", *Am. J. Cardiol.*, vol. 59, no. 1, pp. 20–23, 1987.

[12] B. J. Drew, M. M. Pelter, D. E. Brodnick, A. V. Yadav, D. Dempel, and M. G. Adams, "Comparison of a new reduced lead set ECG with the standard ECG for diagnosing cardiac arrhythmias and Myocardial Ischemia", *J. Electrocardiol.*, vol. 35, no. 4, Part B, pp. 13–21, 2002.

[13] M. Green, M. Ohlsson, J. Lundager Forberg, J. Björk, L. Edenbrandt, and U. Ekelund, "Best leads in the standard electrocardiogram for the emergency detection of acute coronary syndrome", *J. Electrocardiol.*, vol. 40, no. 3, pp. 251–256, 2007.

[14] E. A. P. Alday, A. Gu, A. J. Shah, C. Robichaux, A.-K. I. Wong, C. Liu, F. Liu, A. B. Rad, A. Elola, S. Seyedi, Q. Li, A. Sharma, G. D. Clifford, and M. A. Reyna, "Classification of 12-lead ECGs: The PhysioNet/Computing in Cardiology Challenge 2020", *Physiological Measurement*, vol. 41, no. 12, p. 124 003, Jan. 2021. DOI: 10.1088/1361-6579/abc960. [Online]. Available: https://doi.org/10.1088/1361-6579/abc960.

[15] M. A. Reyna, N. Sadr, E. A. Perez Alday, A. Gu, A. Shah, C. Robichaux, B. A. Rad, A. Elola, S. Seyedi, S. Ansari, Q. Li, A. Sharma, and G. D. Clifford, "Will two do? varying dimensions in electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

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

[17] G. D. Clifford, C. Liu, B. Moody, L.-w. H. Lehman, I. Silva, Q. Li, A. E. Johnson, and R. G. Mark, "AF classification from a short single lead ECG recording: the PhysioNet/Computing in Cardiology Challenge 2017", in *2017 Computing in Cardiology (CinC)*, IEEE, vol. 44, 2017, pp. 1–4.

[18] J. Zheng, J. Zhang, S. Danioko, H. Yao, H. Guo, and C. Rakovski, "A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients", *Sci. Data*, vol. 7, no. 48, pp. 1–8, 2020.

[19] F. Liu, C. Liu, L. Zhao, X. Zhang, X. Wu, X. Xu, Y. Liu, C. Ma, S. Wei, Z. He, J. Li, and E. N. Y. Kwee, "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.

[20] 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.

[21] J. Zheng, H. Cui, D. Struppa, J. Zhang, S. M. Yacoub, H. El-Askary, A. Chang, L. Ehwerhemuepha, I. Abudayyeh, A. Barrett, G. Fu, H. Yao, D. Li, H. Guo, and C. Rakovski, "Optimal multi-stage arrhythmia classification approach", *Sci. Data*, vol. 10, no. 2898, pp. 1–17, 2020.

[22] R. Bousseljot, D. Kreiseler, and A. Schnabel, "Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet", *Biomedizinische Technik*, vol. 40, no. S1, pp. 317–318, 1995.

[23] 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", *Sci. Data*, vol. 7, no. 1, pp. 1–15, 2020.

[24] L. Bacharova, R. H. Selvester, H. Engblom, and G. S. Wagner, "Where is the central terminal located?: In search of understanding the use of the wilson central terminal for production of 9 of the standard 12 electrocardiogram leads", *Journal of Electrocardiology*, vol. 38, no. 2, pp. 119–127, 2005, ISSN: 0022-0736. DOI: https://doi.org/10.1016/j.jelectrocard.2005.01.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022073605000178.

[25] N. Miyamoto, Y. Shimizu, G. Nishiyama, S. Mashima, and Y. Okamoto, "The absolute voltage and the lead vector of Wilson's central terminal", *Jpn Heart J.*, vol. 37, no. 2, pp. 203–214, 1996, ISSN: 00214868. DOI: 10.1536/ihj.37.203.

[26] G. Gargiulo, A. Thiagalingam, A. McEwan, M. Cesarelli, P. Bifulco, J. Tapson, and A. van Schaik, "True unipolar ECG leads recording (without the use of WCT)", *Hear Lung Circ.*, vol. 22, S102, Jan. 2013, ISSN: 14439506. DOI: 10.1016/j.hlc.2013.05.243.

[27] G. D. Gargiulo, "True unipolar ECG machine for Wilson Central Terminal measurements", *Biomed Res Int.*, vol. 2015, p. 586 397, 2015, ISSN: 23146141. DOI: 10.1155/2015/586397.

[28] H. Moeinzadeh, P. Bifulco, M. Cesarelli, A. L. McEwan, A. O'Loughlin, I. M. Shugman, J. C. Tapson, A. Thiagalingam, and G. D. Gargiulo, "Minimization of the Wilson's Central Terminal voltage potential via a genetic algorithm", *BMC Research Notes*, vol. 11, no. 1, pp. 1–5, Dec. 2018, ISSN: 17560500. DOI: 10.1186/S13104-018-4017-Y/FIGURES/3. [Online]. Available: https://bmcresnotes.biomedcentral.com/articles/10.1186/s13104-018-4017-y.

[29] B. E. Jin, H. Wulff, J. H. Widdicombe, J. Zheng, D. M. Bers, and J. L. Puglisi, "A simple device to illustrate the Einthoven triangle", *Advances in Physiology Education*, vol. 36, no. 4, p. 319, 2012, ISSN: 15221229. DOI: 10.1152/ADVAN.00029.2012. [Online].

Available: /pmc/articles/PMC3776430/%20/pmc/articles/PMC3776430/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3776430/.

[30] J. Malmivuo and R. Plonsey, "Bioelectromagnetism. 15. 12-lead ECG system", in. Jan. 1975, pp. 277–289, ISBN: 978-0195058239.

[31] R. E. Gregg, T. Yang, S. W. Smith, and S. Babaeizadeh, "ECG reading differences demonstrated on two databases", *Journal of Electrocardiology*, 2021.

[32] 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.

[33] 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, pp. 210–217, 2 2019.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[35] P. S. Ignacio, "Leveraging period-specific variations in ECG topology for classification tasks", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[36] A. Hammer, M. Scherpf, H. Ernst, J. Weiß, D. Schwensow, and M. Schmid, "Automatic classification of full- and reduced-lead electrocardiograms using morphological feature extraction", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[37] J. van Prehn, S. Ivanov, and G. Nalbantov, "Pathologies prediction on short ECG signals with focus on feature extraction based on beat morphology and image deformation", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[38] S. Krivenko, A. Pulavskyi, L. Kryvenko, O. Krylova, and S. Krivenko, "Using mel-frequency cepstrum and amplitude-time heart variability as XGBoost handcrafted features for heart disease detection", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[39] L. Antoni, E. Bruoth, P. Bugata, P. B. Jr., D. Gajdos, S. Horvat, D. Hudak, V. Kmecova, R. Stana, M. Stankova, A. Szabari, and G. Vozarikova, "Automatic ECG classification and label quality in training data", *Under Review*,

[40] P. Nejedly, A. Ivora, R. Smisek, I. Viscor, Z. Koscova, P. Jurak, and F. Plesinger, "Classification of ECG using ensemble of Residual CNNs with or without attention mechanism", *Under Review*,

[41] Z. Xu, Y. Guo, T. Zhao, Y. Zhao, Z. Liu, and X. Sun, "Abnormality classification from electrocardiograms with various lead combinations", *Under Review*,

[42] X. Li, C. Li, X. Xu, Y. Wei, J. Wei, Y. Sun, B. Qian, and X. Xu, "Towards generalization of cardiac abnormality classification using reduced-lead multi-source ECG signal", *Under Review*,

[43] P. Xia, Z. He, Y. Zhu, Z. Bai, X. Yu, Y. Wang, F. Geng, L. Du, X. Chen, P. Wang, and Z. Fang, "A novel multi-scale 2-D convolutional neural network for arrhythmias detection on varying-dimensional ECGs", *Under Review*,

[44] P. Aublin, J. A. Behar, J. Fix, and J. Oster, "Cardiac abnormality detection based on single lead classifier voting", *Under Review*,

[45] C. G. Vazquez, A. Breuss, O. Gnarra, J. Portmann, and G. D. Poian, "Label noise and self-learning label correction in cardiac abnormalities classification", *Under Review*,

[46] D. Nankani and R. D. Baruah, "Feature fused multichannel ECG classification using channel specific dynamic CNN for detecting and interpreting cardiac abnormalities", *Under Review*,

[47] S. Jimenez-Serrano, M. Rodrigo, C. J. Calvo, F. Castells, and J. Millet, "Multiple cardiac disease detection from distinct ECG leads sets using a hybrid supervised and unsupervised machine learning approach", *Under Review*,

[48] J. Kang and H. Wen, "A study on several critical problems on arrhythmia detection using varying-dimensional electrocardiography", *Under Review*,

[49] N. K. Sawant and S. Patidar, "Identification of cardiac abnormalities applying Fourier-Bessel expansion and LSTM on ECG signals", *Under Review*,

[50] M. Heydarian and T. E. Doyle, "Two-dimensional convolutional neural network model for classification of ECG", *Under Review*, 2022.

[51] B. Puszkarski, K. Hryniow, and G. Sarwas, "Comparison of N-BEATS and SotA RNN architectures for heart dysfunction classification", *Under Review*, 2022.

[52] A. Srivastava, A. Hari, S. Pratiher, S. Alam, N. Ghosh, N. Banerjee, and A. Patra, "Channel self-attention deep learning framework for multi-cardiac abnormality diagnosis from varied-lead ECG signals", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[53] O. Linschmann, M. Rohr, K. S. Leonhardt, and C. H. Antink, "Multi-label classification of cardiac abnormalities for multi-lead ECG recordings based on auto-encoder features and a neural network classifier", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[54] P. Nejedly, A. Ivora, R. Smisek, I. Viscor, Z. Koscova, P. Jurak, and F. Plesinger, "Classification of ECG using ensemble of Residual CNNs with attention mechanism", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[55] A. K. Cornely, A. Carrillo, and G. M. Mirsky, "Reduced-lead electrocardiogram classification using wavelet analysis and deep learning", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[56] P. Xia, Z. He, Y. Zhu, Z. Bai, X. Yu, Y. Wang, F. Geng, L. Du, X. Chen, P. Wang, and Z. Fang, "A novel multi-scale convolutional neural network for arrhythmia classification on reduced-lead ECGs", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[57] N. Seki, T. Nakano, K. Ikeda, S. Hirooka, T. Kawasaki, M. Yamada, S. Saito, T. Yamakawa, and S. Ogawa, "Reduced-lead ECG classifier model trained with DivideMix and model ensemble", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[58] J. Suh, J. Kim, E. Lee, J. Kim, D. Hwang, J. Park, J. Lee, J. Park, S.-Y. Moon, Y. Kim, M. Kang, S. Kwon, E.-K. Choi, and W. Rhee, "Learning ECG representations for multi-label classification of cardiac abnormalities", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[59] A. Natarajan, G. Boverman, Y. Chang, C. Antonescu, and J. Rubin, "Convolution-free waveform transformers for multi-lead ECG classification", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[60] H.-C. Yang, W.-T. Hsieh, and T. P.-C. Chen, "A mixed-domain self-attention network for multi-label cardiac irregularity classification using reduced-lead electrocardiogram", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[61] P. F. Sodmann, M. Vollmer, and L. Kaderali, "Segment, perceive classify multitask learning of the electrocardiogram in a single neural network", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[62] R. Clark, M. Heydarian, K. Siddiqui, S. Rashidiani, A. Khan, and T. E. Doyle, "Detecting cardiac abnormalities with multi-lead ECG signals: A modular network approach", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[63] P. Aublin, M. B. Ammar, N. Achache, M. Benahmed, A. E. Hichami, M. Barret, J. Fix, and J. Oster, "Cardiac abnormality detection based on an ensemble voting of single-lead classifier predictions", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[64] N. L. Wickramasinghe and M. Athif, "Multi-label cardiac abnormality classification from electrocardiogram using deep convolutional neural networks", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[65] H. Han, S. Park, S. Min, H.-S. Choi, E. Kim, H. Kim, S. Park, J. Kim, J. Park, J. An, K. Lee, W. Jeong, S. Chon, K. Ha, M. Han, and S. Yoon, "Towards high generalization performance on electrocardiogram classification", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[66] T. Vicar, P. Novotna, J. Hejc, O. Janousek, and M. Ronzhina, "Cardiac abnormalities recognition in ECG using a convolutional network with attention and input with an adaptable number of leads", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[67] M. Alkhodari, G. Apostolidis, C. Zisou, L. J. Hadjileontiadis, and A. H. Khandoker, "Swarm decomposition enhances the discrimination of cardiac arrhythmias in varied-lead ECG using ResNet-BiLSTM network activations", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[68] G. Garcia-Isla, F. Muscato, A. Sansonetti, S. Magni, V. Corino, and L. Mainardi, "Ensemble classification combining ResNet and hand-crafted features with three-steps training", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[69] H. Jessen, R. van de Leur, P. Doevendans, and R. van Es, "Automated diagnosis of reduced-lead electrocardiograms using a shared classifier", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[70] F. Chollet *et al.* (2015). "Keras", [Online]. Available: `https://github.com/fchollet/keras`.

[71] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy", *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. DOI: `10.1038/s41586-020-2649-2`. [Online]. Available: `https://doi.org/10.1038/s41586-020-2649-2`.

[72] T. pandas development team, *Pandas-dev/pandas: Pandas*, version latest, Feb. 2020. DOI: `10.5281/zenodo.3509134`. [Online]. Available: `https://doi.org/10.5281/zenodo.3509134`.

[73] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library", in *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

[74] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python", *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[75] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro,

F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental algorithms for scientific computing in Python", *Nature Methods*, vol. 17, pp. 261–272, 2020. DOI: 10.1038/s41592-019-0686-2.

[76] Martın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Y. Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015. [Online]. Available: https://www.tensorflow.org/.

[77] L. Antoni, E. Bruoth, P. Bugata, P. B. Jr., D. Gajdos, S. Horvat, D. Hudak, V. Kmecova, R. Stana, M. Stankova, A. Szabari, and G. Vozarikova, "A two-phase multilabel ECG classification using one-dimensional convolutional neural network and modified labels", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[78] H. Ren, M. Xiong, and B. Hooi, "Robust and task-aware training of deep residual networks for varying-lead ECG classification", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[79] W. Cai, F. Liu, X. Wang, B. Xu, and Y. Wang, "Classifying different dimensional ECGs using deep residual convolutional neural networks", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[80] C. G. Vazquez, A. Breuss, O. Gnarra, J. Portmann, and G. D. Poian, "Two will do: CNN with asymmetric loss, self-learning label correction, and hand-crafted features for imbalanced multi-label ECG data classification", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[81] X. Li, C. Li, X. Xu, Y. Wei, J. Wei, Y. Sun, B. Qian, and X. Xu, "Towards generalization of cardiac abnormality classification using ECG signal", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[82] F. M. Muscato, V. D. Corino, and L. T. Mainardi, "Ensemble learning of modified residual networks for classifying ECG with different set of leads", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[83] G. Bortolan, "3-D ECG images with deep learning approach for identification of cardiac abnormalities from a variable number of leads", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[84] W. P. Lebing Pan and, M. Li, Y. Guan, and Y. An, "MTFNet: A morphological and temporal features network for multiple leads ECG classification", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[85] B. U. Demirel, A. H. Dogan, and M. A. A. Faruque, "Two might do: A beat-by-beat classification of cardiac abnormalities using deep learning with domain-specific features", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[86] H. J. Crocker and A. W. Costall, "An InceptionTime-Inspired convolutional neural network to detect cardiac abnormalities in reduced-lead ECG data", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[87] S. Jimenez-Serrano, M. Rodrigo, C. J. Calvo, F. Castells, and J. Millet, "Multiple cardiac disease detection from minimal-lead ECG combining feed-forward neural networks with a one-vs-rest approach", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[88] M. Bodini, M. W. Rivolta, and R. Sassi, "Classification of ECG signals with different lead systems using AutoML", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[89] T. Uhlemann, J. Prim, N. Gumpfer, D. Grun, S. Wegener, S. Krug, J. Hannig, T. Keller, and M. Guckert, "An ensemble learning approach to detect cardiac abnormalities in ECG data irrespective of lead availability", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[90] P. Warrick, V. Lostanlen, M. Eickenberg, M. N. Homsi, A. Rodriguez, and J. Anden, "Arrhythmia classification of reduced-lead electrocardiograms by scattering-recurrent networks", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[91] N. Osnabrugge, K. Keresztesi, F. Rustemeyer, C. Kaparakis, F. Battipaglia, P. Bonizzi, and J. Karel, "Multi-label classification on 12, 6, 4, 3 and 2 lead electrocardiography signals using convolutional recurrent neural networks", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[92] M. Baumgartner, M. Kropf, L. Haider, S. Veeranki, D. Hayn, and G. Schreier, "ECG classification combining conventional signal analysis, random forests and neural networks – a stacked learning scheme", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[93] B. Puszkarski, K. Hryniow, and G. Sarwas, "N-beats for heart dysfunction classification", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[94] B.-J. Singstad, E. M. Muten, and P. H. Brekke, "Multi-label cardiac abnormality classification from electrocardiogram using deep convolutional neural networks", *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

Appendices

Table 1: A summary of the developed algorithms by the official teams ranked on the scores of the two-lead version of the hidden test data. The following abbreviations are used in the table: Kr: Keras [70], NP: NumPy [71], PD: pandas [72], PT: PyTorch [73], SKL: scikit-learn [74], SP: SciPy [75], TF: TensorFlow [76], ASL: asymmetric loss, BCE: binary cross-entropy, CM: Challenge metric, CNN: convolutional neural network, CRNN: convolutional recurrent neural networks, DNN: deep neural network, LSTM: long short-term memory, MLP: multi-layer perceptron, PCA: principal component analysis, ResNet: Residual Networks, SE: squeeze-and-excitation, SQI: signal quality index.

Team and paper	**Score**	**Libraries**	**Model**	**Methods**
`ISIBrno-AIMT` [54]	0.58	NP, PT, SKL, SP	custom variant ResNets, multi-head attention mechanism	preprocessing, BCE, custom challenge loss and custom sparsity loss functions
`DSAIL_SNU` [65]	0.57	PT	WRN with SE block	BCE, ASL, demographic encoded features plus flags, *Adam* optimizer
`NIMA` [64]	0.56	Kr, TF	CNN, Spatial dropouts, average pooling	signal preprocessing, mixure BCE loss
`cardiochallenger` [52]	0.53	NP, SKL	channel self-attention architecture	preprocessing and inception and residual neural model
`CeZIS` [77]	0.52	NP, PD, PT	ResNet50	modified labels, BCE
`DataLA_NUS` [78]	0.52	NP, SKL	1D-Resnet18 and 1D-Efficientnet	task-aware training of deep ResNet, weight loss for imbalanced class
`USST_Med` [79]	0.50	Kr, SP, TF	CNN layer, SENet Residual blocks	preprocessing, relabeled, merged label, focal loss
`SMS+1` [80]	0.49	NP, PT	1D CNN	combined hand-crafted features, self-learning label correction, ASL, , *Adam* optimizer
`Dr_Cubic` [81]	0.48	NP, PT, SKL	SE-ResNet	peak detection, ensemble learning, *AdamW* optimizer
`ami_kagoshima` [57]	0.47	NP, PT, SKL, SP	Model ensemble, DNN, EffientNet, MLP	DivideMix label refinement to clean and noisy samples with GMM, manifold-MixUp
`BUTTeam` [66]	0.46	NP, PT	modified ResNet CNN, attention layer	weighted cross-entropy, CM loss, manually adjusted learning rate, *Adam* optimizer, data augmentation

Continued on next page

Table 1 – continued from previous page

Team and paper	Score	Libraries	Model	Methods
`iadi-ecg` [63]	0.46	NP, PT, SKL	CNN, SE modules	preprocessing, post-processing by adjusting decision thresholds, combined BCE and soft-dice loss
`snu_adsl` [58]	0.45	NP, PT	EfficientNet-B3	preprocessing, feature extraction, pre-training, label masking, threshold optimization
`Polimi_1` [82]	0.45	NP, PD, SKL	three modified ResNet ensemble, SE block	preprocessing, trained on three subsets, channel attention, majority voting
`prna` [59]	0.44	NP, PD, PT, SKL	waveform transformer, MLP head	linear projection, pre-trained weights
`UMCU` [69]	0.41	NP, PT	3-layer shared classifier, CNN, MLP, Adaptive Pooling layer	combined features, cross-entropy loss, manual tuning of linear layers number, additional Euclidean loss, pre-training, *Adam* optimizer
`ibmtPeakyFinders` [36]	0.40	MATLAB R2020b	adaptive logistic regression algorithm (LogitBoost)	preprocessing, SQI, Hilbert QRS detection and correction, hand-crafted features
`Gio_new_img` [83]	0.39	MATLAB R2020b	pre-trained image CNN, *GoogleNet*	preprocessing, ECG images, random oversampling (ROS) for class imbalance
`csu_anying` [84]	0.38	NP, PT	1D-CNN, bidirectional LSTM	preprocessing, BCE loss, *AdamW* optimizer
`AIRCAS_MEL1` [56]	0.38	Kr, NP, SKL, TF	multiple branch 2-D CNN with residual connection, SE	preprocessing, BCE loss, *Adam* optimizer
`METU-19` [85]	0.34	Kr, PD, SKL, SP	CNN with residual blocks	preprocessing, peak correction, feature extraction, *Adam* optimizer
`UoB_HBC` [86]	0.34	Kr, NP, TF	*InceptionTime* CNN	max pooling, hyperparameter tuning, *AdamW* optimizer
`HaoWan_AIeC` [60]	0.34	NP, PD, PT	mixed-domain self-attention ResNet, multi-head attention, SE	preprocessing, lead quality check, data augmentation, excluded two training sets, *Adam* optimizer
`Biomedic2ai` [62]	0.34	neurokit, NP, PD, SKL, TF	1D-dCNN, shallow perceptron network, wide deep model	preprocessing, feature extraction, BCE loss, *Adam* optimizer
`itaca-UPV` [87]	0.34	MATLAB R2020b	feed-forward neural network	preprocessing, features extraction, correction and selection, One-vs-Rest classification

Continued on next page

Table 1 – continued from previous page

Team and paper	Score	Libraries	Model	Methods
`Eagles` [55]	0.30	MATLAB R2021a	*SqueezeNet*, wavelet analysis	preprocessing, limiting patients records, scalograms, transfer learning, stochastic gradient descent with momentum optimizer
`HeartlyAI` [61]	0.25	NP, PT	U-Net, residual convolutional blocks	preprocessing, manual annotation modification, pre-trained model, Tversky, BCE, ASL and Ranger Loss, augmentation
`AADAConglomerate` [53]	0.23	NP, PT, SKL	CNN, LSTM, linear layer	preprocessing, weighted, generalised Softmax loss
`DSC` [37]	0.23	MATLAB R2021a	XGBoost binary classifiers	feature extraction, one-versus-rest, Bayesian optimization, image deformation
`BiSP_Lab` [88]	0.23	NP, PD, SKL, SP	AutoML, auto-SKL, AutoKeras, Tree-Based Pipeline Optimization Tool	preprocessing, optimized ML by genetic programming
`Care4MyHeart` [67]	0.20	MATLAB R2021a	ResNet Bi-directional LSTM	preprocessing, novel swarm decomposition algorithm, deep-activated plus hand-crafted features, *Adam* optimizer
`Sunset` [38]	0.20	NP, PD, SKL	XGBoost	preprocessing, data augmentation, hand-crafted features, balanced data groups
`CardioIQ` [89]	0.13	NeuroKit2, NP, PT	hybrid ensemble of CNN sub-models	soft voting, rule-based inference
`BitScattered` [90]	0.10	NP	scattering transform and LSTM	preprocessing, depth-wise separable convolution, transfer learning by canonical correlation analysis
`Cordi-Ak` [35]	-0.10	NP, SKL, SP	random forest classifier	PCA segmentation, ECG segments as high-dimensional cloud embeddings, feature extraction
`heartMAASters` [91]	-0.46	Kr, NP, TF	CRNN	preprocessing, wavelet-based ECG segmentation, feature extraction, hyperparameter tuning
`easyG` [92]	-0.60	NP, PD, SKL	regression random forest models, MLP	feature extraction, *Adam* optimizer
`WEAIT` [93]	-0.62	NP, PT	MLP, LSTM, N-BEATS	preprocessing, wavelet, PCA
`CardiOUS` [94]	-0.63	NeuroKit2, NP, SKL	CNN	preprocessing, classified Fourier transformed ECG and mean heart rate, *Adam* optimizer