Benchmarking Deep Learning Methods for Arrhythmia Detection

E. Merdjanovska*,† and A. Rashkovska*
* Department of Communication Systems, Jožef Stefan Institute, Ljubljana, Slovenia
† Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
E-mails:{elena.merdjanovska, aleksandra.rashkovska}@ijs.si

***Abstract*—Automatic arrhythmia detection methods are a very significant area of computational ECG analysis. This field has been researched for a long time, however, there are various challenges still faced. Some of the main flaws in current ECG-based arrhythmia classification research are limited variety of datasets used and varying experimental setups, which makes it difficult to directly compare different methods. Most often, a method is evaluated on a specific dataset and task (set of arrhythmia classes). By placing these methods under unified evaluation setup (one umbrella), we can apply (evaluate) them on a wider range of datasets and tasks than they were originally proposed for. To address these challenges, in this paper, we perform benchmarking of some of the most significant deep-learning based methods for arrhythmia detection. These methods are compared on four datasets, considering the most significant state-of-the-art arrhythmia classification tasks. Included are the data from the CinC2017 and CPSC2018 challenges, as well as two recently published large-scale ECG arrhythmia datasets: the PTB-XL and the Shaoxing Hospital Database. The analyses cover a wide range of both morphological and rhythmic arrhythmias, all while focusing on methods suitable for single-lead analysis. In addition, the classification performance on 12-lead data and single-lead data is compared and discussed.**

***Keywords*—deep learning; arrhythmia; ECG; benchmarking;**

I. Introduction

Arrhythmia detection is the most significant and widely researched ECG application area. Various heartbeat abnormalities are known as arrhythmias under one name. These abnormalities are detected by medical professionals using ECG due to its simplicity and non-invasive nature. There are two main categories of arrhythmias. The first type is called morphological arrhythmias and are characterized by the irregularity of a single heartbeat. The second type are the so-called rhythmic arrhythmias, characterized by a set of irregular heartbeats. The corresponding arrhythmia detection tasks are known as form and rhythm tasks, respectively. The development of automatic ECG-based heartbeat classification and arrhythmia detection methods represents a large portion of the research involving computational methods for ECG analysis. The research on this topic has involved some standard methods in the past, such as frequency analysis [1], wavelet transform [2] and template matching [3]. In recent years, the focus has started to shift towards machine learning methods [4], with the majority of state-of-the-art arrhythmia detection studies now using deep learning (DL) techniques. Convolutional neural networks (CNNs) are commonly used for all learning tasks related to images and signals, including ECG. CNNs have been used for arrhythmia detection on a wide range of datasets, evaluation scenarios, as well as various target class groupings [5, 6, 7, 8].

The research in the area of arrhythmia detection has been mostly focused on classifying the heartbeats in the MIT-BIH Arrhythmia database [9] in the 5 groups of arrhythmias established by the Association for Advancement of Medical Instruments (AAMI) [10]. Some studies report almost perfect results for this specific problem, for example, the study in [11] reports overall precision and recall of around 96-97%. However, this is achieved using the intra-patient evaluation paradigm and does not reflect a realistic scenario. Due to this variability in the evaluation procedures employed, some of which are highly flawed, as well as the limited number of test subjects in this public database, there is still a need for further research before employing automatic machine learning models for detecting arrhythmias in clinical practice. Standardization of the evaluation procedure, as well as including representative heartbeats from a variety of data sources, instead of only one database, is necessary to further advance the research area of heartbeat classification for arrhythmia detection.

In recent years, besides the MIT-BIH database, several new databases have been established as highly notable regarding the number of subjects involved or the type of arrhythmia included. Some of them have been the subject of the recent prominent CinC/Physionet [12] or China Physiological Signal challenges [13]. Recently made public databases have drawn much attention with the high number of patient measured for 12-lead ECG, like the PTB-XL database [14] and the Chapman University and the Shaoxing People's Hospital database [15]. Therefore, benchmarking these datasets with the currently most significant deep-learning based methods for arrhythmia detection under unified evaluation setup (one umbrella) is the main challenge in this paper. The analyses will cover a wide range of both heartbeat form and heart rhythm classes, with the focus on tasks for single-lead analysis.

The rest of this paper is structured as follows. Section II describes the ECG databases and defines the investigated arrhythmia tasks. Section III presents the experimental design,

We acknowledge the financial support of the Slovenian Research Agency under Grant P2-0095.

including the preprocessing pipeline, the selected DL architecture for the benchmarking tasks, and the unified evaluation procedure. The obtained results are presented and analyzed in Section IV. Finally, Section V concludes the paper.

II. ECG Databases and Arrhythmia tasks

In this paper, we use 4 databases that contain annotations for arrhythmia types: CinC/Physionet Challenge 2017 Database [12], China Physiological Signal Challenge 2018 Database [13], Chapman University and Shaoxing People's Hospital Arrhythmia Database [15], and PTB-XL Database [14]. Each of this databases contains short-term ECG measurements, ranging from a few seconds up to a few minutes, while the majority of the recordings is 10 seconds long. Common for these databases is that each short measurement is from a different person and each measurement is annotated with one or a few arrhythmia labels, referring to the entire measurement. These databases were mainly chosen due to their wide use in arrhythmia detection literature and their size. Included here are one single-lead database and three 12-lead databases, however in all cases we perform benchmarking for single-lead classification methods. For this purpose, we always choose lead II from 12-lead databases to simulate the use of single-lead sensor, since it has been shown as the best-performing lead [7].

Arrhythmia detection is commonly treated as a supervised machine learning task, more specifically as a classification task. Each of these databases is associated with one or two arrhytmhia tasks (set of arrhythmia classes), which will be used here for benchmarking. An overview of the most common arrhythmia types (classes) and their abbreviations, as well as whether they are related to heartbeat morphology or heart rhythm, is given in Table I. The classes in each task are given in Table II, while more details about each dataset and corresponding task(s) are given in the subsections below.

A. CinC/Physionet Challenge 2017 (CinC2017)

The dataset used in the Atrial Fibrilation (AFIB) Classification Challenge in 2017 [12], organized by PhysioNet and CinC, is one of the most widely used datasets for the development of AFIB detection methods, establishing itself as an AFIB benchmark dataset after the challenge [16]. The data for the challenge consisted of a collection of 8528 recordings, lasting from 9 seconds to 60 seconds. Each recording included one non-invasive ECG signal sampled at 300 Hz, which was obtained using a mobile ECG recording device – Alivecor KardiaMobile [12]. The recordings were obtained while the users held the two electrodes of the device in one hand each, creating a lead I (LA - RA) equivalent ECG. Many of the ECGs in the dataset were inverted (RA - LA) since the device did not require the user to have the electrodes in any particular orientation. This dataset contains only 4 labels: Normal, AFIB, Other and Noise. The class "Other" covers a wide range of abnormal non-AFIB rhythms. This four-class classification task for AFIB detection, from the 2017 CinC/Physionet Challenge, is the first task considered in this paper. In Table II, it is referred to as CinC2017 task.

B. China Physiological Signal Challenge 2018 (CPSC2018)

The China Physiological Signal Challenge [13] has been organized every year since 2018, with the dataset of the first year being continuously used ever since for validation of arrhythmia detection methods, especially those attempting to utilize all 12 ECG leads [17, 8]. This database was collected from 11 hospitals and contains 12-lead ECG recordings lasting from 6 to 60 seconds. The recordings were sampled at 500 Hz and were taken from 3178 female and 3699 male patients. This dataset covers a wide range of significant arrhythmia types, including bundle branch blocks (LBBB and RBBB), premature beats (PVC and PAC), atrioventricular blocks (AVB1) and atrial fibrillation (AFIB). All 9 classes comprising the CPSC2018 task can be found in Table II.

C. PTB-XL Database

The PTB-XL database [14] is a large dataset by Physikalisch-Technische Bundesanstalt (PTB) in Germany, collected between 1989 and 1996, but it was made publicly available in 2020. It consists of 21,837 12-lead ECG recordings from 18,885 distinct patients, which were sampled at 500 Hz, each lasting 10 seconds. The database contains 71 different statements, whereas at least one statement is assigned to each recording as a recording-level annotation. The statements are divided into three categories: diagnostic, form and rhythm. The authors propose multiple classification tasks on this database [17], however we choose the form and rhythm tasks, since they contain finest granularity of arrhythmia types. The 15 form classes and 12 rhythm classes in the PTB-XL benchmark tasks are given in Table II.

D. Chapman University and Shaoxing People's Hospital Arrhythmia Database (ARR10000)

This database is among the first to include a very large number of individual subjects (more than 10,000), which is significant for many computational ECG analysis applications, including arrhythmia detection. It was collected by the Chapman University and the Shaoxing People's Hospital [15], and contains 12-lead ECG signals sampled at 500 Hz. The signals are short, each 10 seconds in duration. The database includes 11 heart rhythms, significantly covering a few types of not very common supraventricular tachycardia (SVTA). In addition, it also includes a wide range of form labels, however the authors propose the use of this database foremost for rhythm classification tasks [15]. They present two options: a finer-grained 11-class rhythm task and 4-class merged rhythm task. Both of them are given in Table II, where the subclasses of the merged arrhythmia task are given in brackets.

III. Experimental Design

In this paper, we perform benchmarking of state-of-the-art deep learning pipelines for arrhythmia classification. Each pipeline consists of data preparation procedure and is aimed

TABLE I. ARRHYTHMIA TYPES DICTIONARY AND CATEGORIZATION

Heartbeat morphology	Rhythm
NORM - normal sinus beat	SR - sinus rhythm
∼(Noise) - signal noise	SI - sinus irregularity
AVB1 - first-degree atrioventricular block	SARRH - sinus arrhythmia
LBBB - left bundle branch block beat	STACH - sinus tachycardia
RBBB - right bundle branch block beat	SBRAD - sinus bradycardia
PAC - premature atrial contraction	PACE - normal functioning artificial pacemaker
PVC - premature ventricular contraction	SVARR - supraventricular arrhythmia
STD - ST segment depression	BIGU - bigeminal pattern
STE - ST segment elevation	SVTAC - supraventricular tachycardia
ABQRS - abnormal QRS	PSVT - paroxysmal supraventricular tachycardia
VCLVH - voltage criteria (QRS) for left ventricular hypertrophy	TRIGU - trigeminal pattern
QWAVE - Q wave present	AFIB - atrial fibrillation
LOWT - low amplitude T-waves	AFLT - atrial flutter
NT_ - non-specific T-wave changes	SVT - supraventricular tachycardia
LPR - prolonged PR interval	AT - atrial tachycardia
INVT - inverted T-waves	AVNRT - atrioventricular node reentrant tachycardia
LVOLT - low QRS voltages in the frontal and horizontal leads	AVRT - atrioventricular reentrant tachycardia
HVOLT - high QRS voltage	SAAWR - sinus atrium to atrial wandering rhythm
TAB_ - T-wave abnormality	
PRC(S) - premature contraction(s)	

TABLE II. ARRHYTHMIA DETECTION TASKS; MULTIPLE LABELS COMPOSING ONE CLASS ARE GIVEN IN BRACKETS

Task	No. Classes	Classes
CinC2017	4	NORM, AFIB, OTHR, ∼(Noise)
CPSC2018	9	NORM, AFIB, AVB1, LBBB, RBBB, PAC, PVC, STD, STE
PTB-XL Form	15	ABQRS, PVC, STD, VCLVH, QWAVE, LOW, NT_, PAC, LPR, INVT, LVOLT, HVOLT, TAB_, STE, PRC(S)
PTB-XL Rhythm	12	SR, AFIB,STACH, SARRH, SBRAD, PACE, SVARR, BIGU, AFLT, SVTAC, PSVT, TRIGU
ARR10000 Rhythm	7	NORM, SI, SBRAD, AFIB, AFLT, STACH, SVT
ARR10000 Rhythm Merged	4	(NORM, SI), (SBRAD), (AFIB, AFLT), (STACH, SVT, AT, AVNRT, AVRT, SAAWR)

for a specific frequency. The most important part of the pipeline is the deep learning architecture. Each of these pipelines is originally proposed for specific dataset and task. This paper aims to apply the methods in these pipelines to multiple datasets and tasks, under a unified evaluation setup. The methods in the pipeline, more specifically the data preparation and the neural network architecture, as well as the evaluation setup, are given in the sections below.

A. Data Preparation

Since we are focusing on deep learning methods, raw data is fed to the neural network, which means that complex data preprocessing or feature engineering is not necessary. We still need to modify the ECG data to a format suitable to serve as neural network input. This means that the ECG signals need to be normalized and re-sampled to the frequency the neural network was designed for. In addition, each input has to be of the same length, which is achieved by either cropping longer ECG segments or zero-padding the shorter ones. The actual frequency used in the neural network input, as well as the length of the signal, varies in each pipeline.

In most cases, the entire ECG segment, sometimes cropped or padded, serves as input. This type of input vector generation is referred to as sequence segmentation in our experiments. Alternatively, a sliding window approach can be used, where smaller windows are fed in the network. A prediction is obtained for each window separately, and then those predictions are aggregated to get a single prediction for the entire measurement. In these cases, a sliding window of 2.5 seconds is used, with 50% overlap.

B. Deep Learning Architectures

We use three main neural network architectures for the experiments in this paper. First is the residual network (ResNet), whose variations have been used in multiple arrhythmia detection works [5, 6, 18]. Next, we include the winning model of the 2018 China Physiological Signal Challenge (CPSCWinnerNet) [8], and finally, a novel architecture with a residual-based temporal attention block (RTA-CNN) [16], proposed for the CinC2017 dataset. We do not apply each architecture to each task, since our main aim is to try to confirm the reported state-of-the-art results under a unified evaluation setup.

The ResNet we use consists of residual blocks with convolutional layers. Different works have used very similar variations of the ResNet architecture [5, 17], however we use the implementation described in [18]. This architecture was intended for single-lead ECG with a sampling frequency of 250 Hz and uniform input length of 60 seconds.

CPSCWinnerNet [8] consists of convolutional blocks (CNNs), gated recurrent units (GRUs) and an attention layer. It is designed to be used for 12-lead ECG data, however it can be modified to work with a single-lead ECG as well.

The proposed architectures consists of filters intended for 500 Hz input data. The input signal should be 144 seconds long, which is the longest recording length in CPSC2018 dataset. The winning model scores are obtained with an ensemble of 12-lead and single-lead models in order to get the final predictions, however here we only experiment with the architecture itself applied to single-lead ECG.

The third architecture used in this paper, RTA-CNN architecture with exponential nonlinearity loss (EN-loss), has been proposed for the CinC2017 challenge task and dataset. It is a residual network with RTA blocks utilizing temporal attention mechanisms. The authors also propose a novel EN-loss. It was originally used on single-lead CinC2017 data, sampled at 300 Hz and limited to 30 seconds input length. Recordings shorter than 30 seconds are expanded by replicating, and for those longer than 30 seconds, a random 30-second segment is chosen. This method is different to the simple cropping and zero padding used in the other methods and is referred to as repeatcrop later in the results.

All networks were trained for a maximum of 300 epochs, using early stopping on a validation set. We used a batch size of 128 and Adam optimizer with a learning rate of 0.0001. The implementation[1] was done in Python 3.8 with Tensorflow 2, and trained on an Nvidia GeForce RTX 3090 GPU.

C. Evaluation Procedure

The aim of this paper is to compare arrhythmia classification methods under a unified setting. In order to achieve this, most important is a fair and unified evaluation procedure, which gives a realistic estimate of the model performance. In this paper, we use a stratified 10-fold cross-validation. In each iteration, 8 folds are used for training, one is used for validation and one for testing. The scores presented in the results section are average scores over all folds on the test set. The validation sets are used for early stopping of the neural network training. This procedure is applied to each dataset and each task.

D. Evaluation Metrics

In order to measure the performance of classifiers quantitatively, multiple metrics can be calculated. When an imbalanced dataset is in question, as is the case of most diagnostic classification tasks, the accuracy is not very indicative of model performance. More important are the true positive rate (TPR) and positive predictive value (PPV) of each class, which are combined in the F1 score. In addition, the area under ROC curve (AUC) is another strong classification metric. Macro-averaged scores are preferred in imbalanced classification tasks, so in this paper we compare the methods using macro F1 score and macro AUC. It should be noted that in some cases only one of these metrics is reported in the reference paper and only that one can be used for comparison.

[1]The code is available on: https://github.com/elenamer/ecg_classification_DL

IV. Results and Discussion

In this section, we will present and discuss the results from the benchmarking experiments described in this paper. An overview of the classification settings included in this paper is given in Table III. The results for the CinC2017 Challenge Dataset are given in Table VI, for the CPSC2018 Database in Table IV, for PTB-XL in Table V, and for ARR10000 in Table VII, including scores for different classification tasks relevant for each dataset. AUC and F1 scores are shown, in addition to ACC included for comparison, in the cases where ACC is reported in the original paper. As described in the previous sections, the methods are compared under a unified training and evaluation setup, using only a single ECG lead, while the reported state-of-the-art results are sometimes referring to models using all 12 leads.

The results for CinC/Physionet Challenge 2017 are given in Table VI. The reference scores are given in the last two rows, where the last row is referring to the overall winner of the challenge [19]. This challenge-best method used a complex classification pipeline including expert feature extraction, abductive interpretation and extensive data preprocessing, including data relabeling and lead inversion. As a result, it achieves an F1 score of 0.831, which is significantly higher than all of the methods that we chose to benchmark. The method that we want to compare with is RTA-CNN [16] (given in the second last row), which achieves accuracy of 0.83 and does not report the corresponding F1 or AUC scores. We were able to achieve the same accuracy using the CPSCWinnerNet architecture, however our experiments did not reproduce the reported result with the RTA-CNN architecture and repeatcrop data preparation technique, as described in [16].

The China Physiological Signal Challenge 2018 results are given in Table IV. In this table, we can observe that with the same CPSCWinnerNet architecture [8] an F1 score of 0.7287 is achieved. This is comparable with their best single-lead reported result of F1 score (0.75). The winning score is much higher, F1 of 0.837, however it is obtained with an ensamble of many 12-lead and single-lead models. Since our goal is to benchmark the methods under the same set of conditions, we only perform experiments with the architectures without additional ensemble or complex data preprocessing techniques.

The PTB-XL benchmark paper [17] includes the PTB-XL form and PTB-XL rhythm tasks. The reported results are on 12-lead data using the ResNet architecture, with the scores on both tasks reported in Table V. Our benchmarking experiments with both traditional ResNet and CPSCWinnerNet showed that with ResNet we were able to achieve better scores on the PTB-XL Rhythm task than those reported (F1 of 0.4673 as opposed to 0.4190). However, we did not come close to the state-of-the-art for the PTB-XL form task, with an F1 of 0.1795, which is significantly lower than the reported 0.2823. It should also be noted that the best-performing ResNet scores are obtained with full-sequence input, as opposed to the sliding window segmentation used in the PTB-XL benchmarking paper.

The fourth dataset, Chapman University and Shaoxing Peo-

TABLE III. OVERVIEW OF THE CLASSIFICATION SETTINGS USED IN THIS PAPER

Model	ResNet	CPSCWinner	RTA-CNN
Task	CPSC2018, PTB-XL Rhythm, PTB-XL Form, ARR10000 Rhythm and ARR10000 Rhythm Merged	CPSC2018, CINC2017, PTB-XL Rhythm, PTB-XL Form, ARR10000 Rhythm and ARR10000 Rhythm Merged	CINC2017
Segmentation	sequence and sliding window	sequence	repeatcrop and sequence
Frequency	250Hz	500Hz	300Hz
Reference	[18, 17]	[8]	[16]

ple's Hospital Arrhythmia Database (ARR10000), is benchmarked on two tasks proposed by the database's authors [15]: ARR10000 Rhythm and ARR10000 Rhythm Merged. The results on these two tasks are given in Table VII. The results from the same two models and pipelines on single-lead ECG, same as for the other datasets, are shown. The results reported in one of the first papers using this dataset are also given in this table for reference. We can see that both CPSCWinnerNet and ResNet achieve higher scores than the reference reported result on the extended rhythm task, with a significantly higher F1 of 0.8834. On the reduced ARR10000 Rhythm Merged task, our benchmark methods achieve similar results to those reported in literature.

The most important findings from this paper are summarized in Fig. 1, where the reported and obtained F1 scores for each dataset and task can be compared. We can see that on both tasks in the ARR10000 database, as well as for rhythm classification on PTB-XL, our state-of-the-art methods, chosen due to their performance on other datasets and tasks, achieve higher scores than those currently reported in literature. Furthermore, we achieved lower, but comparable results both on the CinC2017 and CPSC2018 challenge datasets. Regarding the PTB-XL form task, our setup resulted in significantly lower scores than those obtained in the PTB-XL benchmarking paper. Since this task covers a very wide range of fine-grained ECG form abnormalities, not found in any of the other tasks, this low score could indicate that in order to classify these arrhythmia types, all 12 leads are necessary. In all other tasks, mainly consisting of rhythm classes, single-lead models have performed well, with scores comparable to 12-lead models.

When comparing the performances of the different deep learning architectures, ResNet and CPSCWinnerNet result in similar scores on the PTB-XL rhythm task, both ARR10000 tasks and the CPSC2018 task. This proves that these architectures are able to successfully capture both heartbeat morphology and rhythm abnormalities. RTA-CNN, on the other hand, is the significantly weaker architecture, which indicates that the novel RTA block is not as robust as the well-established standard residual blocks, found in ResNet, and the recurrence-based networks with attention, like CPSCWinnerNet.

V. CONCLUSION

We have benchmarked four most prominent ECG arrhythmia databases (CinC2017, CPSC2018, PTB-XL, ARR1000) with the currently most significant deep-learning methods for arrhythmia detection (ResNet, CPSCWinnerNet, RTA-CNN) by placing them under unified evaluation setup (one umbrella). While doing so, we consider the most significant state-of-the-art arrhythmia classification tasks: form and rhythm detection. The results have shown that some of the selected deep learning architectures can achieve even better performance on some datasets when compared to the results reported in the original benchmark paper. This confirms that standardization of the evaluation procedure should be seen as a necessity to further advance the research area of arrhythmia detection.

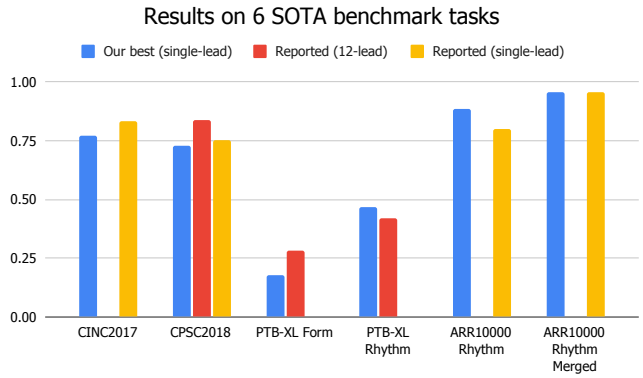

Figure 1. Comparison of the best F1 scores obtained with the benchmarking approach and the state-of-the-art (SOTA) F1 scores reported in literature

REFERENCES

[1] K. Minami, H. Nakajima, and T. Toyoshima, "Real-time discrimination of ventricular tachyarrhythmia with Fourier-transform neural network," *IEEE Transactions on Biomedical Engineering*, vol. 46, no. 2, pp. 179–185, 1999.

[2] O. T. Inan, L. Giovangrandi, and G. T. A. Kovacs, "Robust Neural-Network-Based Classification of Premature Ventricular Contractions Using Wavelet Transform and Timing Interval Features," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 12, pp. 2507–2515, 2006.

[3] V. Krasteva and I. Jekova, "QRS Template Matching for Recognition of Ventricular Ectopic Beats," *Annals of Biomedical Engineering*, vol. 35, pp. 2065–76, 01 2008.

[4] E. Luz, W. Schwartz, G. Chávez, and D. Menotti, "ECG-based Heartbeat Classification for Arrhythmia Detection: A Survey," *Computer Methods and Programs in Biomedicine*, vol. 127, pp. 144–164, 2015.

[5] A. Hannun, P. Rajpurkar, M. Haghpanahi, G. Tison, C. Bourn, M. Turakhia, and A. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Medicine*, vol. 25, pp. 65–69, 2019.

TABLE IV. BENCHMARKING RESULTS ON THE CPSC2018 DATASET

Model	Segmentation	CPSC2018 AUC	CPSC2018 F1
CPSCWinnerNet	slidingwindow2.5s	0.9137	0.6777
	sequence30s	0.8499	0.5141
	sequence10s	0.8496	0.4899
	sequence144s (500 Hz)	0.941	**0.7287**
ResNet	slidingwindow2.5s	0.9133	0.6709
	sequence10s	0.9297	0.7103
Reported (12-lead ensemble) [8]		/	0.837
Reported (single-lead) [8]		/	**0.75**

TABLE V. BENCHMARKING RESULTS ON THE PTB-XL DATASET

Model	Segmentation	PTB-XL Form AUC	PTB-XL Form F1	PTB-XL Rhythm AUC	PTB-XL Rhythm F1
CPSCWinnerNet	slidingwindow2.5s	0.7989	0.1492	0.9522	0.4533
	sequence10s	0.7399	0.1169	0.951	0.4054
ResNet	slidingwindow2.5s	0.7951	0.1753	0.9367	0.461
	sequence10s	0.7822	**0.1795**	0.921	**0.4673**
Reported (12-lead) [17]		0.89	**0.2823**	0.957	**0.4190**

TABLE VI. BENCHMARKING RESULTS ON THE CINC2017 DATASET

Model	Segmentation	CinC2017 AUC	CinC2017 F1	CinC2017 ACC
CPSCWinnerNet	sequence30s	0.9205	**0.7699**	**0.8305**
	sequence10s	0.8537	0.6058	0.7462
	sequence144s (300 Hz)	0.9038	0.7299	0.8087
	sequence144s (500 Hz)	0.9212	0.76	0.827
RTA-CNN	repeatcrop30s	0.8421	0.6705	0.7724
	sequence10s	0.7934 (0.028)	0.455	0.7088
	sequence30s	0.8311	0.6037	0.7575
Reported [16]		/	/	**0.83**
Reported challenge-best [19]		/	**0.831**	/

TABLE VII. BENCHMARKING RESULTS ON THE ARR10000 DATASET

Model	Segmentation	ARR10000 Rhythm AUC	ARR10000 Rhythm F1	ARR10000 Rhythm ACC	ARR10000 Rhythm Merged AUC	ARR10000 Rhythm Merged F1	ARR10000 Rhythm Merged ACC
CPSCWinnerNet	slidingwindow2.5s	0.9883	0.8762	0.9422	0.9951	**0.9563**	0.9614
	sequence10s	0.9866	0.8546	0.94	0.9955	0.956	0.9612
ResNet	slidingwindow2.5s	0.9889	0.8723	0.9372	0.9947	0.9515	0.9567
	sequence10s	0.9887	**0.8834**	0.9447	0.9936	0.9435	0.9499
Reported (single-lead) [7]		/	**0.8004**	0.9224	/	**0.9557**	0.9613

[6] A. Sellami and H. Hwang, "A robust deep convolutional neural network with batch-weighted loss for heartbeat classification," *Expert Systems with Applications*, vol. 122, pp. 75–84, 2019.

[7] O. Yildirim, M. Talo, E. J. Ciaccio, R. S. Tan, and U. R. Acharya, "Accurate deep neural network model to detect cardiac arrhythmia on more than 10,000 individual subject ECG records," *Computer Methods and Programs in Biomedicine*, vol. 197, 2020, article no. 105740.

[8] 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, 2020, article no. 100886.

[9] G. Moody and R. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2001.

[10] "Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms," Association for the Advancement of Medical Instrumentation, Standard, 2012.

[11] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, M. Adam, A. Gertych, and R. S. Tan, "A deep convolutional neural network model to classify heartbeats," *Computers in Biology and Medicine*, vol. 89, pp. 389–396, 2017.

[12] G. D. Clifford, C. Liu, B. Moody, L. 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)*, 2017, pp. 1–4.

[13] 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.

[14] P. Wagner, N. Strodthoff, R. Bousseljot, D. Kreiseler, F. Lunze, W. Samek, and T. Schaeffter, "PTB-XL: A Large Publicly Available electrocardiography Dataset," *Scientific Data*, vol. 7, 2020, article no. 154.

[15] 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," *Scientific Data*, vol. 7, 2020, article no. 48.

[16] Y. Gao, H. Wang, and Z. Liu, "An end-to-end atrial fibrillation detection by a novel residual-based temporal attention convolutional neural network with exponential nonlinearity loss," *Knowledge-Based Systems*, vol. 212, 2021, article no. 106589.

[17] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, "Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1519–1528, 2021.

[18] K. Weimann and T. O. F. Conrad, "Transfer learning for ECG classification," *Scientific Reports*, vol. 7, 2021, article no. 6067.

[19] T. Teijeiro, C. A. García, D. Castro, and P. Félix, "Arrhythmia classification from the abductive interpretation of short single-lead ecg records," *2017 Computing in Cardiology (CinC)*, pp. 1–4, 2017.