DeepQ Arrhythmia Database: A Large-Scale Dataset for Arrhythmia Detector Evaluation

Meng-Hsi Wu
HTC Research & Healthcare
jake.mh_wu@htc.com

Edward Y. Chang
HTC Research & Healthcare
eyuchang@gmail.com

ABSTRACT

DeepQ Arrhythmia Database, the first generally available large-scale dataset for arrhythmia detector evaluation, contains 897 annotated single-lead ECG recordings from 299 unique patients. DeepQ includes beat-by-beat, rhythm episodes, and heartbeats fiducial points annotations. Each patient was engaged in a sequence of lying down, sitting, and walking activities during the ECG measurement and contributed three five-minute records to the database. Annotations were manually labeled by a group of certified cardiographic technicians and audited by a cardiologist at Taipei Veteran General Hospital, Taiwan. The aim of this database is in three folds. First, from the scale perspective, we build this database to be the largest representative reference set with greater number of unique patients and more variety of arrhythmic heartbeats. Second, from the diversity perspective, our database contains fully annotated ECG measures from three different activity modes and facilitates the arrhythmia classifier training for wearable ECG patches and AAMI assessment. Thirdly, from the quality point of view, it serves as a complement to the MIT-BIH Arrhythmia Database in the development and evaluation of the arrhythmia detector. The addition of this dataset can help facilitate the exhaustive studies using machine learning models and deep neural networks, and address the inter-patient variability. Further, we describe the development and annotation procedure of this database, as well as our on-going enhancement. We plan to make DeepQ database publicly available to advance medical research in developing outpatient, mobile arrhythmia detectors.

CCS CONCEPTS

• **Applied computing → Health informatics**;

KEYWORDS

DeepQ, arrhythmia database, Tricorder XPRIZE

1 INTRODUCTION

Arrhythmia is a very common heart condition in which the heart beats too slowly, too fast or irregularly. This condition is often preceded by the events of aberrant heartbeats. Interpreting electrocardiograms (ECGs) is an inexpensive and noninvasive way for cardiologists to assess the cardiac conduction system and diagnose arrhythmias. There are two major arrhythmia categories: the beat-level (morphological) arrhythmia for a single heartbeat being irregular and the rhythm-level (rhythmic) arrhythmia for a series of heartbeats being abnormal. A physician can determine the abnormal cardiac activities at each part of the heart by measuring the time intervals between fiducial points on the ECG. Measuring the amount of a wave that travels through the heart muscle can help infer which part of the heart being hypertrophic, and precisely which type of arrhythmia has occurred. However, the process of analyzing each heartbeat of an entire ECG recording can be very demanding for a human, especially for a long-term ECG. The variability of a patient's underlying wave morphology, cardiac rhythms, and artifacts conspire to make the analysis difficult. This difficulty is further elevated by the inter-patient diversity. Designs for automated arrhythmia detection system must address these confounding influences.

The evaluation of the arrhythmia detector is categorized into two standards, namely the intra-patient and inter-patient. In the intra-patient paradigm, the entire ECG database is subdivided into a training set and a test set based on a fraction of each heartbeat class. The specification on which patients or heartbeats to be included in the training phase or the test phase is not guaranteed. Both training and test sets will most likely comprise parts of an ECG recording from the same patient. As the ECG characteristics within a person are less variable, using heartbeats from the same patient for both training and testing will potentially leak information about the test data into the design of the model and lead to optimistic bias to the performance evaluation. In spite of the drawback, intra-patient approach has been widely used in many publications for arrhythmia classification evaluation [2, 6–8, 10, 11, 14]. In contrast to the previous method, the more clinically realistic inter-patient evaluation paradigm requires the classifier to detect heartbeat abnormalities of unseen patients. With this approach, inter-patient variation is considered and the entire ECG dataset is partitioned into the training and testing subsets based on patients. All data from the same record will only be presented in either the training phase or the test phase and arrhythmia detectors trained from this scheme are expected to provide more realistic estimates.

The MIT-BIH Arrhythmia Database (MITDB) is the first generally available standard test material for arrhythmia detection analysis [9]. The availability of this database has inspired extensive research and publication in arrhythmia detection over the past two decades [8]. It is regarded as the most representative database for developing automated arrhythmia detectors and is listed as one of the evaluation standards by the Association for the Advancement of Medical Instrumentation (AAMI). While the MITDB database has been an invaluable benchmark, the small number of unique

MMHealth'17, October 23–27, 2017, Mountain View, CA, USA

ACM ISBN 978-1-4503-5504-9/17/10…$15.00
https://doi.org/10.1145/3132635.3132647

individuals in this database characterizes the limited variability and insufficiency for exhaustive studies. A widely referenced division scheme, proposed by [1], separates the MITDB database into two divisions to be more coherent with the reality; each division contains only 22 records and exhibits some underlying sample bias issues. Specifically, one single patient contributed as much as 40% of the total supraventricular ectopic beats (SVEB) in the training set and two patients contributed nearly 90% of the SVEB beats in the test set. As long as the automated classifiers are specifically designed to correctly classify those significant SVEB contributors, irrespective of how they might perform on those SVEB minorities, pretty-looking results can still be produced. Furthermore, the MITDB database suffers from two significant limitations. First, 60% of its ECG recordings were obtained from inpatients in which subjects were usually in a controlled environment with 12-lead ECGs attached. Outpatient ECG measures are obtained from patients in less controlled environments and usually contain countless motion artifacts and data loss. The deterioration of signal quality usually impose significant difficulty on reliable arrhythmia detection. Ideally, first-class automated detectors should perform well on both inpatients and outpatients. For those arrhythmia detection algorithms developed solely based on the MITDB database, their precise performance on outpatients' data is uncertain. Second, the MITDB database does not always provide the same pair of ECG leads in each recording. The principal lead is a modified limb lead II (MLII) and the supporting lead can be V1, V2, V4, or V5 depending on patients. Provided that the input domains are morphologically different for the second channel, training detectors using morphologically derived features from MITDB's second channel is controversial. The amount of information gained from using MITDB's second channel is questionable and remains unexplained. The work done by [13] demonstrates that a cardiologist-level arrhythmia detector can be achieved with single-lead ECGs (we will discuss the shortcoming of this work in Section 3.4). In the other two public arrhythmia datasets — MIT-BIH Supraventricular Arrhythmia Database (SVDB) [5] and St. Petersburg Institute of Cardiological Technics 12-lead Arrhythmia Database (INCART) [4] — the number of unique patients is also in the two-digit scale size and disease annotations are unaudited by a cardiologist. Importantly, outpatient ECG measures are not considered in these two databases as well.

Approaches from machine learning models and deep neural networks have consistently achieved the state-of-the-art performance and outperformed their human counterpart with the help of seizable labeled datasets. We strongly believe the limited impact on improving the disease classifier performance is due to the lack of variability and availability of large-scale data, with the largest datasets having about 78 unique subjects. This assumption concurs with our successful experience in the Qualcomm Tricorder XPRIZE competition [12]. In contemplating these weaknesses, in 2015, we began to construct a new dataset that has reached about 10 folds the size of the MITDB database to fit three main purposes:

(1) *Scale*: Large-scale data in terms of unique patients
(2) *Diversity*: Inpatient and outpatient ECG measures with detailed beat-by-beat, rhythm and heartbeat fiducial points annotations.
(3) *Quality*: Complement the MITDB database in the exhaustive development and evaluation of the arrhythmia detector

In this work, we make the first step towards a large-scale arrhythmia dataset recorded using a wearable ECG monitoring patch. We intentionally consider three different activities and motion intensities, namely lying down, sitting and walking, that both inpatients and outpatients are most likely to experience during their long-term recordings. This inspiration can facilitate the development and evaluation of automated disease detectors to account for more differentiation in the ECG morphology. Our new arrhythmia dataset provides more examples and variety of arrhythmia events than other publicly available databases of its kinds.

2 THE DEEPQ ARRHYTHMIA DATABASE

The DeepQ Arrhythmia Database (DeepQ) is being developed with Taipei Veteran General Hospital, Taiwan. We attempt to include a plethora of realistic arrhythmic ECG recordings that are observed in clinical practice and outpatients. The entire data collection process complies with the human participants guideline and regulation of the Institutional Review Board (IRB)[1]. During a clinical visit or cardiac examination, patients were asked if they were willing to participate in the data collection. Participating patients were explained through the guidance and the IRB informed consents were obtained before their ECG examinations. We used a water-resistant, non-invasive single-lead ECG patch adopted from our award-winning DeepQ Tricorder vital sense module in the Qualcomm Tricorder XPRIZE competition [12]. DeepQ ECG patch was worn on the participant's left chest in the modified lead II configuration. We also placed the Philips 1810 series holter on the participant along with our DeepQ ECG recording device for comparison throughout the collection process. All patients were instructed to rest for at least five minutes during the initial setup and between each activity session. Throughout the ECG measurement, each patient was engaged in a sequence of three five-minute activities, namely lying down, sitting and walking and contributed three recordings to the database. This protocol ensures a smooth transition between different activity intensities. In the walking session, participants were allowed to walk freely around the facility to mimic the recordings in outpatient situations.

The ECG data was sampled at a frequency of 250 Hz with with 24-bit resolution and wirelessly transferred via BLE to a remote smartphone receiver for temporary storage. The recorded ECG signals were then uploaded to our private server and later compared with Philips 1810 series holters for quality assurance by the certified cardiographic technicians prior to the annotation phase. Figure 1 shows three ECG excerpts from the DeepQ dataset. All records were annotated using a web interface tool purposely designed for this task. Disease labeling works were initially done by a group of certified cardiographic technicians and then verified by a cardiologist. To ensure the annotation quality and consistency, annotation rules were devised; the cardiographic technicians were guided through the use of web tool and supervised by a senior technician. There are three labeling categories in this database: beat-by-beat, rhythm episodes and heartbeat fiducial points. The beat-by-beat annotation protocol is compatible with the AAMI recommendations. Along with the beat-by-beat class annotations, each heartbeat's P, QRS, and T fiducial points are also marked, if present. A strip is marked

[1]IRB reference number: 2015-03-001A

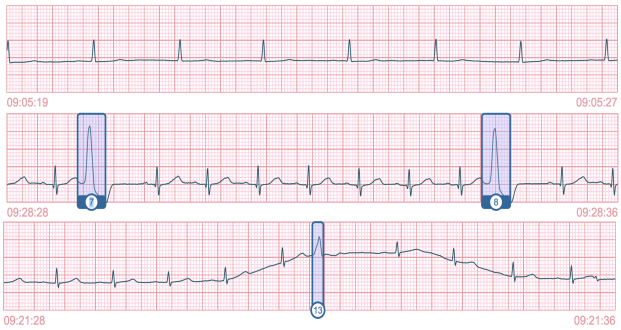

Figure 1: Three ECG excerpts from the DeepQ dataset. From top to bottom: normal sinus rhythm, two PVC beats, a PVC event during a walking session.

for rhythm-level labels in which the beginning and the end of the strip correspond to the onset and offset of an abnormality segment.

Currently the DeepQ database contains 897 annotated records from 299 unique patients. Each record is about five minutes in duration, from one activity and includes a compact clinical summary with technical information about the recording. Comparing the DeepQ dataset to the other three public databases (MITDB, SVDB, and INCART), DeepQ's number of unique subjects and records outnumber them. DeepQ not only has at least twice the amount of unique patients but also has more heartbeats in the AAMI supraventricular ectopic beats (SVEB) and ventricular ectopic beats (VEB) classes compared to the MITDB database. Table 1 summarizes this comparison. In addition to the beat-level classes, the rhythm-level arrhythmia collection includes 48 unique atrial fibrillation (AF) or atrial flutter (AFL) cases, four first degree atrioventricular block (1°AVB), three second degree atrioventricular block (2°AVB Type I and Type II), one paroxysmal supraventricular tachycardia (PSVT), six supraventricular tachycardia (SVT), one sinoatrial block (SAB), two ventricular tachycardia (VT), and one junctional rhythm. Table 2 summarizes the quantity of overlapped rhythm-level arrhythmias between DeepQ and the other three representative databases.

3 OTHER ARRHYTHMIA DATABASES

In this section, we start by discussing three important open databases on ECG arrhythmia — the MIT-BIH Arrhythmia Database, the MIT-BIH Supraventricular Arrhythmia Database, and the St. Petersburg Institute of Cardiological Technics 12-lead Arrhythmia Database — that are available to researchers and comparing them with our DeepQ database. Later, we will mention a recent work done by Stanford [13] and discuss the differences between the their work and ours.

3.1 The MIT-BIH Arrhythmia Database

This database contemplates the AAMI recommendations [3] for class labeling and is widely referenced in many literatures. It contains 48 two-lead ECG recordings from 47 different patients (recordings 201 and 202 were obtained from the same patient). Each record is sampled at 360 Hz for approximately 30 minutes. The first 23 recordings (100-series) were obtained from a random selection of the routine holter recordings and the remaining 25 records (200-series) were obtained to include the uncommon but clinically important arrhythmia cases. In majority of the recordings, the principal lead is a modified limb lead II (MLII) and the other lead is usually lead V1. In some cases, the second channel is lead V2, V4, or V5. The annotations for beat classification information are verified by at least two independent experts. In general, the MLII lead shows more prominent QRS complex and is often used to detect heartbeats. Ectopic beats are more discernible than the normal beats in the second lead source (V1, V2, V4, or V5). The MITDB database is by far the most rigorous dataset among all freely available databases. More details on this database are shown in Tables 1 and 2.

3.2 The MIT-BIH Supraventricular Arrhythmia Database

This database supplements the examples of supraventricular arrhythmias in the MITDB database and comprises 78 two-lead ECG recordings. Each record is about 30 minutes in duration and is sampled at 128 Hz. A Marquette Electronics 8000 holter scanner was initially used for automatic annotation. A medical student later reviewed the annotations and performed editing if needed. Its development was primarily based on records that are moderately difficult and contained a mixture of mixed supraventricular and ventricular ectopic beats. Since the annotations were not audited by any cardiologist, the lack of rigorousness placed this database in a less representative position. More details on this database are presented in Table 1.

3.3 St. Petersburg Institute of Cardiological Technics 12-lead Arrhythmia Database

This database contains 75 twelve-standard-lead ECG recordings from 32 holter records. The original records were collected from 32 patients undergoing tests for coronary artery disease. Each recording is half-hour in length and is sampled at 257 Hz. Beat annotations were initially done by an automatic algorithm and then corrected manually, following the standard PhysioBank beat annotation definitions [4]. However, the annotation locations have not been manually corrected and as a result, there might be occasional misalignment in the records. Most patients in this database had ventricular ectopic beats and none had pacemakers. In constructing the database, preference was given to patients with ECGs consistent with ischemia, coronary artery disease, conduction abnormalities, and arrhythmias. Without a verification from a cardiologist, the lack of rigorousness also placed this database in a less representative position. More details of this database are presented in Tables 1 and 2.

3.4 Stanford Arrhythmia Database[2]

Stanford ML Group proposed a model that outperforms a cardiologist in diagnosing a great variety of arrhythmias from single-lead ECGs. The key to this accomplishment is the combination of a deep convolutional network and a novel dataset two orders of magnitude larger than previous datasets of its kind. The ECG data is sampled at a frequency of 200 Hz and is collected from the single-lead ECG

[2]The Stanford database remains unreleased by the time of submission

Table 1: Database details

Database	#Records	#Subjects	#SVEB Records	#VEB Records	AAMI Heartbeat Class			Sample Rate (Hz)	Duration (min)	#Channels
					N	SVEB	VEB			
DeepQ	**897**	**299**	**71**	**125**	**303233**	**4158**	**8616**	**250**	**5**	**1**
MITDB	48	47	32	37	90125	2781	7009	360	30	2
SVDB	78	78	73	67	161902	12083	9897	128	30	2
INCART	75	32	18	69	153517	1958	19991	257	30	12

Table 2: Number of records for overlapped rhythmic classes

Diagnosis	#Patients		
	DeepQ	MITDB	INCART
AF/AFL	48	11	3
AV block (AVB)	7	1	1
SVT	6	7	-
VT	2	13	-
Junctional rhythm	1	3	-

patch. There are 12 arrhythmias, sinus rhythm and noise classes included in the database. The training dataset contains 64,121 annotated ECG records from 29,163 patients. Each ECG record is 30 seconds in duration and is annotated by a certified cardiographic technician. The test set consists of 336 records from 328 unique patients. Ground truth annotations for each record in the test set were obtained by a committee of three board-certified cardiologists. The cardiologists discussed each individual record as a group and came to a consensus labeling.

Different to our DeepQ's approach, Stanford database only contains rhythmic labels; beat-by-beat and heartbeat fiducial points annotations are not included. The lack of beat-by-beat annotations is a major shortcoming of the Stanford database. In compliance with the AAMI standards, all arrhythmia algorithms must provide the aggregate statistics (sensitivities and positive predictive values) for beat-level comparisons in a record-by-record basis [3]. Without beat-by-beat labels, automated arrhythmia detecting algorithms trained on the Stanford dataset cannot be evaluated in the AAMI assessment.

4 CONCLUSIONS AND FUTURE WORK

So far, the largest freely available dataset merely has 78 unique subjects. We strongly believe the impact on improving the disease classifier's performance is limited by the factor of lacking sizable annotated data and diversity. DeepQ Arrhythmia Database (DeepQ), a novel large-scale arrhythmia database that contains 897 records of fully annotated beat-by-beat and rhythm episodes, is a marked milestone achieved by our two year annotation effort. The DeepQ dataset is also the largest database with heartbeat fiducial points fully annotated. Different to all existing public databases, we have included ECGs that are influenced by different activities to mimic the outpatient situations. What we anticipate to accomplish is to include more diversity and representative arrhythmic ECG data from both inpatient and outpatient scenarios and support the medical research in developing reliable and mobile automatic arrhythmia detectors for outpatients. DeepQ also opens a new door to the exhaustive studies using most advanced algorithms such as machine learning models and deep neural networks.

In future work, we plan to expand our database to at least three orders of magnitude larger than the current size to include more rare arrhythmic cases and greater diversity. When the dataset size is at a scale of tens of thousands, data can be fully exploited, data-driven models can be fully utilized and precision can be much improved.

ACKNOWLEDGMENTS

We thank Shih-Lin Chang MD for his helpful feedback and contribution to the data collection process.

REFERENCES

[1] Philip De Chazal, Maria O'Dwyer, and Richard B Reilly. 2004. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. *IEEE Transactions on Biomedical Engineering* 51, 7 (2004), 1196–1206.

[2] Lorena SC de Oliveira, Rodrigo V Andreão, and Mario Sarcinelli-Filho. 2011. Premature ventricular beat classification using a dynamic Bayesian network. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. IEEE, 4984–4987.

[3] ANSI-AAMI EC57. 2012. Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms. *Association for the Advancement of Medical Instrumentation, Arlington, VA* (2012).

[4] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet. *Circulation* 101, 23 (2000), e215–e220.

[5] Scott David Greenwald, Ramesh S Patil, and Roger G Mark. 1990. Improved detection and classification of arrhythmias in noise-corrupted electrocardiograms using contextual information. In *Computers in Cardiology 1990, Proceedings*. IEEE, 461–464.

[6] Xing Jiang, Liqing Zhang, Qibin Zhao, and Sahin Albayrak. 2006. ECG arrhythmias recognition system based on independent component analysis feature extraction. In *TENCON 2006. 2006 IEEE Region 10 Conference*. IEEE, 1–4.

[7] Martin Lagerholm, Carsten Peterson, Guido Braccini, Lars Edenbrandt, and Leif Sornmo. 2000. Clustering ECG complexes using Hermite functions and self-organizing maps. *IEEE Transactions on Biomedical Engineering* 47, 7 (2000), 838–848.

[8] Eduardo José da S Luz, William Robson Schwartz, Guillermo Cámara-Chávez, and David Menotti. 2016. ECG-based heartbeat classification for arrhythmia detection: A survey. *Computer methods and programs in biomedicine* 127 (2016), 144–164.

[9] George B Moody and Roger G Mark. 2001. The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine* 20, 3 (2001), 45–50.

[10] Stanislaw Osowski, Linh Tran Hoai, and Tomasz Markiewicz. 2004. Support vector machine-based expert system for reliable heartbeat recognition. *IEEE transactions on biomedical engineering* 51, 4 (2004), 582–589.

[11] G Krishna Prasad and JS Sahambi. 2003. Classification of ECG arrhythmias using multi-resolution analysis and neural networks. In *TENCON 2003. Conference on Convergent Technologies for the Asia-Pacific Region*, Vol. 1. IEEE, 227–231.

[12] Qualcomm. 2017. XPRIZE Tricorder Winning Teams. http://tricorder.xprize.org/teams. (2017).

[13] Pranav Rajpurkar, Awni Y Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y Ng. 2017. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. *arXiv preprint arXiv:1707.01836* (2017).

[14] Jimena Rodriguez, Alfredo Goni, and Arantza Illarramendi. 2005. Real-time classification of ECGs on a PDA. *IEEE Transactions on Information Technology in Biomedicine* 9, 1 (2005), 23–34.