Icentia11K: An Unsupervised Representation Learning Dataset for Arrhythmia Subtype Discovery

Shawn Tan[1,3], Guillaume Androz[4], Satya Ortiz-Gagné[3], Ahmad Chamseddine[2,3],
Pierre Fecteau[4], Aaron Courville[1,3], Yoshua Bengio[1,3], & Joseph Paul Cohen[1,3]

[1]Université de Montréal, [2]Polytechnique Montréal
[3]Mila, [4]Icentia

Abstract

We release a public electrocardiogram (ECG) dataset of continuous raw signals for representation learning containing 11 thousand patients and 2 billion labelled beats. The signals were recorded with a 16-bit resolution at 250Hz with a fixed chest mounted single lead probe for up to 2 weeks. The average age of the patient is 62.2±17.4 years. 20 technologists annotated each beat's type (Normal, Premature Atrial Contraction, Premature Ventricular contraction) and rhythm (Normal Sinusal Rhythm, Atrial Fibrillation, Atrial Flutter).

To analyse this data we evaluate existing supervised classification methods to replicate their results. We also explore unsupervised representation learning methods to both improve classification performance at small numbers of labelled samples as well as identify arrhythmia subtypes. We present a semi-supervised evaluation framework to evaluate the quality of representation learning methods.

We achieve over 80% accuracy on beat and rhythm classification tasks using supervised models when training using large numbers of samples. In a low data setting these supervised methods do not work as well (achieving around 40% accuracy) and the semi-supervised methods we explore only slightly improve performance. This presents an open challenge to develop better ECG representation learning algorithms and the dataset we release is well suited to develop such a method.

1. Introduction

Arrhythmia detection is presently performed by cardiologists or technologists familiar with ECG readings. Recently, supervised machine learning has been successfully applied to perform automated detection of many arrhythmias [1–4].

However, there may be ECG anomalies that warrant further investigation because they do not fit the morphology of presently known arrhythmia. We seek to use a data driven approach to finding these differences that cardiologists have anecdotally observed, which motivates the representation learning potential of this data.

Our data is collected by the CardioSTAT^{TM}, a single-lead heart monitor device from Icentia[5]. The raw signals were recorded with a 16-bit resolution and sampled at 250Hz with the CardioSTAT^{TM}in a modified lead 1 position. The wealth of data this provides us can allow us to improve on the techniques currently used by the medical industry to process days worth of ECG data, and perhaps to catch anomalous events earlier than currently possible. All data is made public[1].

The ethics institutional review boards at the Université de Montréal approved the study and release of data CERSES-19-065-D.

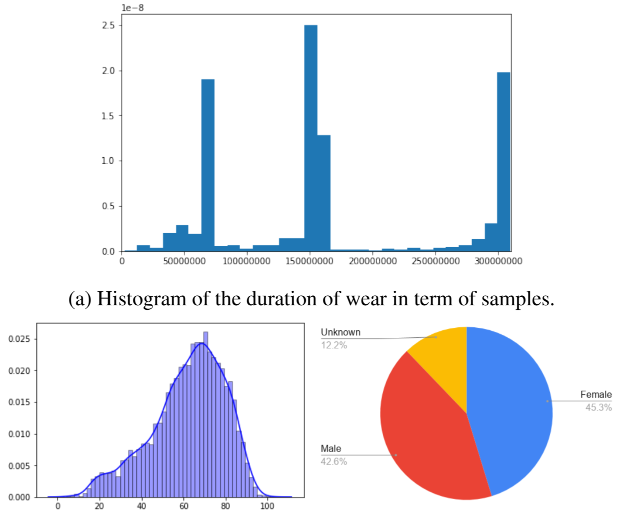
2. Related Work

ECG (or sometimes known as EKG) signals are collected by electrocardiograph machines. These machines traditionally have 10 electrodes, resulting in 12-lead ECG data. These can be thought of as a 12 channel signal that provides additional data about the heartbeat, but allows for only short periods of data capture due to the cumbersome nature of these machines, and are not sufficient for capturing rarer events that happen over time.

One of the first open dataset of ECG signals was the MIT-BIH dataset, created in 1979 [6]. They "expected that the availability of a common database would foster rapid and quantifiable improvements in the technology of automated arrhythmia analysis." This dataset, with just 47 subjects, is still in use today.

The MIMIC-III Waveform Database [7] contains 67,830 waveform records from 30,000 ICU patients. These samples are at a higher sampling rate and with more leads. However, they are only recorded for short periods of time. The ECG-ViEW II dataset [8] aims to be a freely avail-

[1] Data available: https://academictorrents.com/details/af04abfe9a3c96b30e5dd029eb185e19a7055272



(a) Histogram of the duration of wear in term of samples.

(b) Histogram of age (years).

(c) Patient sex.

Figure 1: (a) Demographics of the patients in the data and statistics

able dataset of ECG records together with clinical data for 461,178 patients. Instead of raw signals, only beat information is included: RR interval, PR interval, QRS duration, etc.

More recently, single-lead wearable devices provided much larger amounts of data than before. As these devices could be worn for throughout the day, over a period of a couple of weeks, machine learning had much more data to work with. [9] created an annotated training dataset of ECG signals consisting of 30,000 patients [10]. The authors' approach, and the follow up work claim that their automated models perform at the level of trained cardiologists [1]. However, their data has not been made publicly available.

3. Icentia11k Dataset

The dataset is processed from data provided by 11,000 patients who used the CardioSTAT$^{\text{TM}}$device predominantly in Ontario, Canada, from various medical centers. While the device captures ECG data for up to two weeks, the majority of the prescribed duration of wear was one week. Figure 1a shows the distribution over duration of wear in the unprocessed data.

It should be noted that since the people who wear the device are patients, the dataset does not represent a true random sample of the global population. For one, the average age of the patient is 62.2 ± 17.4 years of age. Furthermore, whereas the CardioSTAT$^{\text{TM}}$can be worn by any patient, it is mostly used for third line exam[2], so the majority of records in the dataset exhibit arrhythmias. No particular effort has been done on patient selection except data collection has been conducted over years 2017 and 2018. Figure 1c shows the distribution over age and gender.

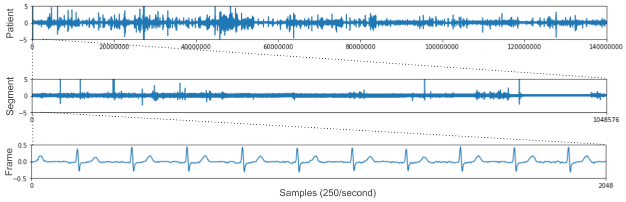


Figure 2: ECG data at different levels of the hierarchy. From top to bottom, a full patient record, a segment, and a frame.

The data is analysed by Icentia's team of 20 technologists who performed annotation using proprietary analysis tools. Initial beat detection is performed automatically and then a technologist analyses the record labelling beat and rhythm types (these will be further elaborated in §3) performing a full disclosure analysis (i.e. they see the whole recording). Finally the analysis is approved by a senior technologist before making it to the dataset.

To further prepare the data for our purposes, we segment each patient record into segments of $2^{20}+1$ signal samples (≈ 70 minutes). This longer time context was informed by discussions with technologists: the context is useful for rhythm detection. We made it a power of two with a middle sample to allow for easier convolution stack parameterisation. From this, we randomly select 50 of the segments and their respective labels from the list of segments. The goal here is to reduce the size of the dataset while maintaining a fair representation of each patient. In the training data, we remove the labels for 80% of the patients. For the remaining 20%, half will be kept for the semi-supervised task, while another half will remain as test data for evaluation. Further details of nomenclature and statistics of the unprocessed and processed data can be found in Table 1.

We describe in further detail the different levels of hierarchy we have separated the data into:

Patient level (3-14 days) At this level, the data can capture features which vary in a systematic way and not isolated events, like the placement of the probes or patient specific noise.

Segment level (approximately 1 hour) A cardiologist can look at a specific segment and identify patterns which indicate a disease while ignoring noise from the signal such as a unique signal amplitude. Looking at trends in the segment help to correctly identify arrhythmia as half an hour provides the necessary context to observe the stress of a specific activity.

Frame level (approximately 8 seconds) At this level, the data can capture features about the beat as well as the rhythm.

While we have provided baseline results only for frame-

[2] Most patients were prescribed CardioSTAT$^{\text{TM}}$by a tertiary referral hospital or care centre

Table 1: Dataset Statistics

Statistic	# (units)
Number of patients	11,000
Number of labeled beats	2,774,054,987
Sample Rate	250Hz
Frame size	$2^{11} + 1 = 2,049$
Segment size	$2^{20} + 1 = 1,048,577$
Total number of frames	1,084,314
Total number of segments	542,157
Dataset Size	271.27GB

level features in this paper, we believe that processing the data with these levels of hierarchy results in some grouping information that could be leveraged to attain better results.

4. Baseline Methods

We replicate the convolutional ResNet neural network of [9] on our public dataset as well as classical baselines which operate on the raw signal. Here the models take a frame as input and are trained to predict beat and rhythm. The number of labelled examples in the evaluation dataset is shown in Table 2.

We also evaluate semi-supervised classification models which utilize the representations described in the next section §4.1. These representations take in a frame and produce a lower dimension representation which is used by a k-nearest neighbor classifier to make predictions. This is done to see if these methods can utilize fewer training examples than the supervised methods as this would be desirable to avoid collecting data for new tasks. The pipeline used for these evaluations is shown in Table 3 and the results of this evaluation are shown in Table 3. Further details, code, and models have been made public in order to facilitate reproducibility and future work[3].

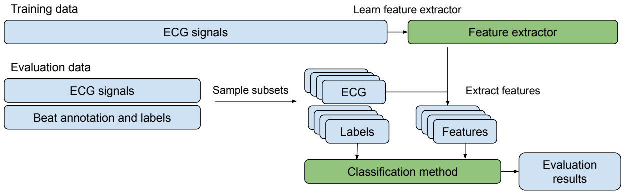


Figure 3: Diagram detailing the training and evaluation pipeline for the representation learning task. We have vary methods in our evaluation for the blocks colored in green.

Table 2: Label counts in the evaluation subset (patients 9000-10999). Each frame is labelled. Only 2 types of labels are provided. Only these meaningful labels are used for evaluation and presented to the classifier.

Beat labels	Count
Normal	174,249
Premature Atrial Contractions	58,780
Premature Ventricular contractions	44,835

(a) Beat labels in the evaluation set

Rhythm Labels	Count
NSR (Normal Sinusal Rhythm)	261,377
AFib (Atrial Fibrillation)	13,056
AFlutter (Atrial Flutter)	3,330

(b) Rhythm labels in the evaluation set

4.1. Unsupervised Representation Learning

While the processed data includes labelled beat and arrhythmia information, we propose an *unsupervised* representation learning challenge to the community.

The goal of this data is to develop unsupervised representations of the ECG signal which can aid in two aspects: 1) Improving the performance of supervised tasks by using the learned representations, and 2) Identifying unknown subtypes of disease by studying the clustering of the representations. While the first aspect is an immediate consequence of a larger dataset, we hope the second aspect will be of interest to researchers as well.

However, being able to find structure in the given raw data can be subjective, so we propose a quantitative semi-supervised classification task as a proxy for evaluating the usefulness of extracted features. We benchmark some common unsupervised algorithms in a semi-supervised setting to establish base quality.

The evaluation consists of predicting the beat and rhythm for each frame in a hold out set (samples id's $> 10,000$). The beat task is to predict if a frame contains all normal beats or contains at least one premature ventricular contractions (PVC) or premature atrial contraction (PAC) anywhere in a frame.

Another task is to predict the rhythm type given a frame. For a given frame the classification method must predict if the rhythm is normal, atrial fibrillation (AFib)[4], or atrial flutter, based on the input representation.

[3] https://github.com/shawntan/icentia-ecg

[4] AFib is a controversal rhythm as cardiologists do not agree on the minimum duration. 8 second frames might not be sufficient to make such a decision.

Table 3: Performance using different learning models and feature representation learning methods. $N/2$ is the number of labelled training examples seen and $N/2$ is the number of test examples evaluated. The proportion of classes in each set is sampled to be equal. This is varied to figure out which models work at low numbers of training examples. Each experiment is repeated using 10 different random seeds for the data split and model initialization. The mean accuracy is over all classes is presented.

		Beat Classification Accuracy		Rhythm Classification Accuracy	
Classification Model	Feature Representation	N=120	N=12000	N=120	N=12000
ConvResNet (48 layers) [9]	Raw signal	0.34 ± 0.04	**0.85 ± 0.03**	0.33 ± 0.06	**0.88 ± 0.01**
Basic ConvNet (5 layer)	Raw signal	0.40 ± 0.08	0.80 ± 0.01	0.35 ± 0.04	0.62 ± 0.11
Multilayer perceptron (1000 units)	Raw signal	0.39 ± 0.08	0.69 ± 0.01	0.39 ± 0.09	0.52 ± 0.01
k-nearest neighbor (k=3)	Raw signal	0.37 ± 0.04	0.65 ± 0.01	0.32 ± 0.04	0.45 ± 0.01
k-nearest neighbor (k=3)	BioSPPy mean beat	0.33 ± 0.08	0.40 ± 0.01	0.36 ± 0.03	0.61 ± 0.01
k-nearest neighbor (k=3)	ConvAE	**0.42 ± 0.06**	0.68 ± 0.01	0.33 ± 0.05	0.47 ± 0.01
k-nearest neighbor (k=3)	FFT	0.39 ± 0.07	0.57 ± 0.01	0.37 ± 0.08	0.61 ± 0.01
k-nearest neighbor (k=3)	PCA (50 dim)	0.38 ± 0.08	0.67 ± 0.01	0.34 ± 0.07	0.47 ± 0.01
k-nearest neighbor (k=3)	Periodogram	0.36 ± 0.06	0.50 ± 0.01	**0.39 ± 0.07**	0.56 ± 0.01

5. Conclusion

Single-lead heart monitors like the CardioSTAT^{TM}are increasingly common, and have the potential for cardiologists to learn much more about arrhythmia and related heart diseases. However, this amount of data means that manual analysis is no longer practical.

Supervised learning serves well as an assistant in medical field; however, it hardly provides information beyond human knowledge. Additionally, certain human body signals can be very complex and imply features that cannot be easily identifiable manually. At present, representation learning methods have a potential in disentangling complex features, and potentially, unveil new signal structures of certain diseases which can correlate with clinical presentations. By releasing this dataset, we believe that we can leverage unsupervised representation learning expertise to not only help to enable training models with lower number of samples, but potentially find new diseases and identify patterns associated with them.

We also propose an evaluation pipeline for learning a feature extractor and evaluating extracted features using known arrhythmia as a proxy to measure the usefulness of the features, providing baseline results for *frame*-level representations under different feature extraction methods. Our data preparation makes a three level hierarchy available — the *segment* and *patient* level grouping of data. While we did not provide baselines that exploit all of these levels, future work that can take advantage of this context to extract better representations, and perhaps, find more interesting structure in the representation space. We also believe that this dataset can serve as a benchmark in other areas of machine learning, such as anomaly and outlier detection, and hierarchical sequence modelling.

References

[1] Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, Ng AY. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine 2019.

[2] Yıldırım Ö, Pławiak P, Tan RS, Acharya UR. Arrhythmia detection using deep convolutional neural network with long duration ecg signals. Computers in biology and medicine 2018.

[3] Mincholé A, Rodriguez B. Artificial intelligence for the electrocardiogram. Nature Medicine 1 2019.

[4] Porumb M, Iadanza E, Massaro S, Pecchia L. A convolutional neural network approach to detect congestive heart failure. Biomedical Signal Processing and Control 2020.

[5] Paquet P, Levesque D, Fecteau P. Adhesive extender for medical electrode anduse thereof with wearable monitor, June 20 2019. US Patent App. 16/093,151.

[6] Moody GB, Mark RG. The impact of the MIT-BIH arrhythmia database. IEEE engineering in medicine and biology magazine the quarterly magazine of the Engineering in Medicine Biology Society 2001.

[7] Johnson AE, Pollard TJ, Shen L, Lehman LwH, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG. MIMIC-III, a freely accessible critical care database. Scientific Data 2016.

[8] Kim YG, Shin D, Park MY, Lee S, Jeon MS, Yoon D, Park RW. ECG-ViEW II, a freely accessible electrocardiogram database. PloS one 2017.

[9] Rajpurkar P, Hannun AY, Haghpanahi M, Bourn C, Ng AY. Cardiologist-level arrhythmia detection with convolutional neural networks. arXiv preprint arXiv170701836 2017.

[10] Turakhia MP, Hoang DD, Zimetbaum P, Miller JD, Froelicher VF, Kumar UN, Xu X, Yang F, Heidenreich PA. Diagnostic utility of a novel leadless arrhythmia monitoring device. The American journal of cardiology 2013.