

# Multi-label Arrhythmia Classification From 12-Lead Electrocardiograms

Po-Ya Hsu<sup>1</sup>, Po-Han Hsu<sup>1</sup>, Tsung-Han Lee<sup>1</sup>, Hsin-Li Liu<sup>2</sup>

<sup>1</sup>UC San Diego, CA, USA

<sup>2</sup>Central Taiwan University of Science and Technology, Taichung City, Taiwan

## Abstract

*In participation of the PhysioNet/Computing in Cardiology Challenge 2020, we developed a novel computational approach for efficiently identifying cardiac abnormalities from 12-lead electrocardiogram (ECG) data. The developed methodology is composed of three processes: selecting representation, generating features, and predicting outcomes. We proposed a cache-inspired method to select a 12-lead ECG heartbeat representation. Moreover, we devised a physiologically interpretable feature generator for segmented 12-lead ECG signals. For multi-label arrhythmia outcome prediction, we innovated an efficient arrhythmia outcome prediction procedure that is adaptable to ECG data of variant lengths. Our team, JuJuRock, received a score of 0.402 using 5-fold cross-validation on the full training data and a score of 0.244 on the final full test data. Team JuJuRock ranked 16<sup>th</sup> out of the 41 teams that participated in this year's Challenge.*

## 1. Introduction

Arrhythmia is a substantial cardiovascular disease since it has been reported to correlate with high prevalence and associated mortality [1]. Different arrhythmia types have different mechanisms and require the appropriate interventions for successful treatments. The electrocardiogram (ECG) is an essential tool for screening and diagnosing cardiac electrical abnormalities [2]. Online cardiac activity ECG monitoring and early diagnosis of arrhythmia are critical in clinical practice. To reduce the time-consuming manual arrhythmia labeling, several computer-aided-diagnosis (CAD) tools have been proposed [3, 4].

However, a majority of these CAD tools have not been verified on multiple datasets. Another drawback of these CAD models is their power in diagnosis - they focus on single-label diagnosis instead of multi-label cardiac abnormalities identification.

To develop an automated program that addresses the aforementioned issues, we participate the PhysioNet/Computing in Cardiology Challenge 2020, which focused on automated, open-source approaches for clas-

sifying multi-label cardiac abnormalities from 12-lead ECGs [5]. Our best entry in the Challenge applied boosting classifier to identify the cardiac abnormalities. Moreover, we deliver the computational approach that contributes to:

- Generating 12-lead ECG heartbeat representation;
- Producing physiologically reasonable feature maps;
- Making efficient cardiac abnormalities identification.

## 2. Methods

We formulate the multi-label arrhythmia classification problem into diagnosing each arrhythmia class with a binary classifier. The binary classifier's training process is illustrated in Figure 1. First, we select the appropriate datasets to include in the training. Next, we process and segment a representative heartbeat from each patient. Subsequently, we generate the amplitude and timing feature maps from each heartbeat segment for classifier training. Following the feature generation, we train the binary classifiers with different deep learning (DL) and machine learning (ML) models. Once the classifiers are trained, we experiment with these models and optimize the parameters in our devised arrhythmia decision algorithm.

### 2.1. Dataset Selection

Six datasets are generously provided as the training candidate, and four out of six were chosen to be included in our model training. The four selected datasets are Georgia, CPSC, CPSC\_2, and PTB-XL 12-Lead ECG databases. Such dataset selection is made based on three reasons: **1) sample size**, these four datasets cover the majority of the cases and could possibly represent the data distribution in the testing set; **2) data length**, variant data length, from five seconds to several minutes, are lying within the chosen datasets; **3) signal quality**, the signal-to-noise ratio (SNR) is relatively high in the selected four datasets compared to the two unselected ones. Furthermore, given the 27 evaluated classes in the challenge, only the data from the patient diagnosed as at least one of the 27 classes is sifted.

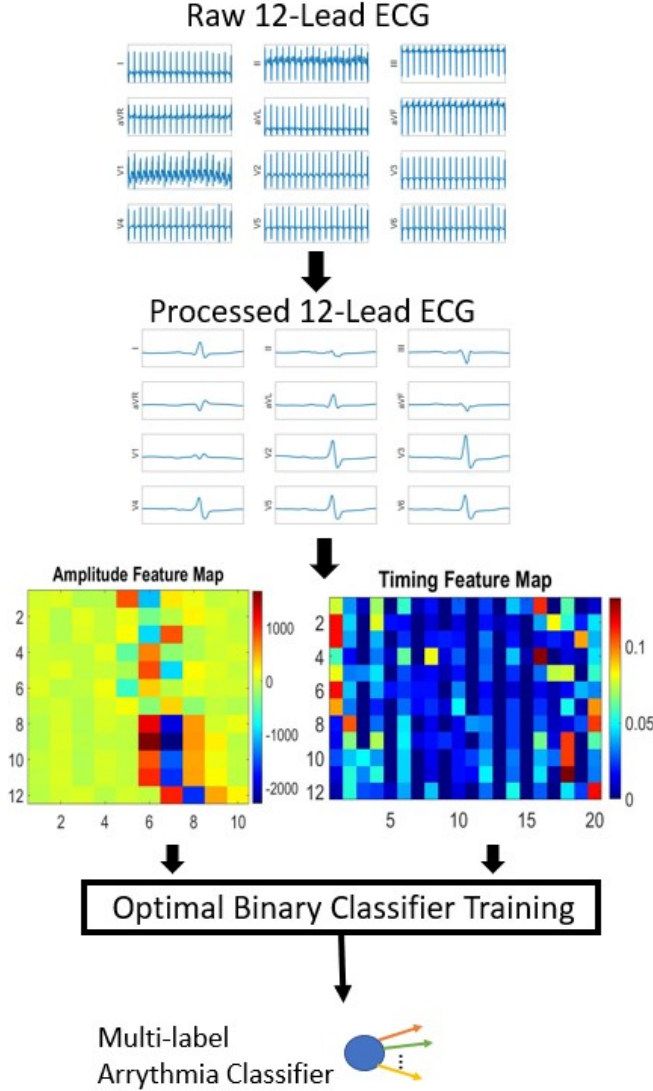


Figure 1. Flowchart of training the proposed multi-label arrhythmia classifier.

## 2.2. Data Segmentation

We developed a heartbeat segmentation algorithm to transform each patient’s raw ECG signal into one representative heartbeat data. Such algorithm consists of four steps: data cleaning, R-peak detection, heartbeat segmentation, and representation selection. In Figure 1, we demonstrate the raw and the processed 12-Lead ECG data.

First, we clean the raw ECG data with a Butterworth low-pass filter and a smoothing function channel-wisely. The Butterworth filter has an order of 12 and 50Hz cut-off frequency. The smoothing function adopts the moving average strategy with 10ms window span.

In the second step, we detect the R-peaks in the cleaned ECG data. All the R-peaks detection is carried out on Lead II data. For the majority of patient samples, we run the famous Pan-Tompkins algorithm to identify the R-peaks; for the few corner cases, we take on the heartbeat detection method targeted at noisy physiological signal as applied in Hsu *et al.*’s work [6] to resolve the issue.

Once we finish labeling the R-peaks of one patient, we chop down every ECG heartbeat into 1-second long frame and concatenate the frames into a tensor. Every 1-second frame has the channel-wise R-peak located at the center. Supposed that  $N$  number of heartbeats is considered, then the tensor will have the dimension of  $12 \times f_s \times N$ , where  $f_s$  symbolizes the ECG sampling rate. To be more specific, we do not consider the head and the tail of the heartbeats into the tensor. Because the SNR is typically low at the endpoints of the signal and the duration of the ECG heartbeat are likely to be less than one second.

At the last step, we employ the clustering method and similarity metric to select the representative one-second ECG heartbeat. For each patient, we conglomerate the ECG frames into at most three groups and recognize the frame lying within the largest group as the representative.

The clustering approach is inspired by the cache updated rule. At the beginning, we construct a cache of three entries, and each entry contains five blocks. Next, we follow the least-recently-used rule to update our cache, which is a frequently utilized technique in computer architecture. We assign a newly visited ECG frame to an existing cache block if the ECG frame has sufficiently high similarity with the frames in the entry, or if there is an empty entry to be filled with; otherwise, we move on to the next ECG frame. Eventually, we select the block data that stores the most recent ECG frame in the largest group as the representation.

As for the similarity metric, we borrow the structural similarity index (SSIM) defined in Wang *et al.*’s study [7]. The SSIM expression is

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \quad (1)$$

where  $c_1 = (0.01L)^2$ , and  $c_2 = (0.03L)^2$ . Inputs  $x, y$  are the two ECG frames;  $\mu_x, \mu_y$  are the averages of the two input frames;  $\sigma_x, \sigma_y$ , and  $\sigma_{xy}$  are the variances of frame  $x$ , frame  $y$ , and the covariance of the two frames  $x, y$ .  $L$  is the range of the ECG voltage values. In this work, frames  $x$  and  $y$  represent the two ECG frames to be compared with, and we empirically set the SSIM threshold as 0.3 to consider two ECG frames belonging to the same group.

## 2.3. Feature Map Generation

We invented a novel physiology-inspired feature generator that is able to efficiently produce the feature maps of

an arbitrary ECG frame. We incorporate the knowledge of saliency into our feature generation model to quantify the P-wave, QRS-complex, and T-wave relevant geometry on two feature maps. One feature map renders the amplitude features, while the other characterizes the timing information. We present our amplitude and timing feature generation algorithms [1,2] as follows:

---

**Algorithm 1** Amplitude Feature Generation

---

**Input:** 1-second 12-Lead ECG Data,  $K$

**Output:** Amplitude Feature Map  $Amp\_Map$

```

Initialize  $Amp\_Map$  as a matrix of size  $12 \times K$ 
for  $ch = 1 \rightarrow 12$  do
   $x \leftarrow$  ECG data of channel  $ch$ 
   $P \leftarrow \frac{dx}{dt} = 0$  // valleys and peaks in the data
   $Q \leftarrow$  array of size  $|P| - 1$ 
  for  $i = 1 \rightarrow |P| - 1$  do
     $Q[i] = P[i + 1] - P[i]$ 
  end for
   $M \leftarrow |Q|$ 
   $I \leftarrow$  Indices of top  $K$  largest values in  $M$ 
  //  $I$  is in non-decreasing order
   $Amp\_Map[ch, :] \leftarrow Q[I]$ 
end for

```

---



---

**Algorithm 2** Timing Feature Generation

---

**Input:** 1-second 12-Lead ECG Data,  $f_s, K$

**Output:** Timing Feature Map  $Time\_Map$

```

Initialize  $Time\_Map$  as a matrix of size  $12 \times 2K$ 
for  $ch = 1 \rightarrow 12$  do
  Same procedure as Algorithm 1 until obtaining  $I$ 
  for  $i = 1 \rightarrow K - 1$  do
     $Time\_Map[ch, 2i - 1] = I[i + 1] - I[i]$ 
     $Time\_Map[ch, 2i] = k - I[i]$ 
    //  $I[i] < k \leq I[i + 1], k \leftarrow \operatorname{argmin} P - I[i]$ 
  end for
   $Time\_Map[ch, 2K - 1] = f_s - I[K]$ 
   $Time\_Map[ch, 2K] = k - I[K]$ 
  //  $I[K] < k \leq f_s, k \leftarrow \operatorname{argmin} f_s - I[i]$ 
end for
 $Time\_Map \leftarrow Time\_Map / f_s$ 

```

---

Both amplitude and timing generation algorithms take the 1-second 12-Lead ECG and the assumed fiducial point number  $K$  as inputs, and output the feature maps of size  $12 \times K$  and  $12 \times 2K$ , respectively. The algorithms compute the features of each channel independently, and then project them onto the output feature maps. For the amplitude map, the salient magnitudes are assessed; for the timing map, the durations between the salient points are taken into account. Under the assumption of existing P, Q,

R, S, and T waves, we set  $K = 10$  and exhibit an example in Figure 1.

## 2.4. Classifier Training

We treat solving the multi-label classification problem as training binary classifiers for each evaluated class. Our heuristics are that each arrhythmia type bears its own unique waveform and is reflected in our generated amplitude or timing maps. Based on the reasoning, we implement the following two experiments to train the binary classifiers for each evaluated class.

**1. Training Models:** Basic DL and ML models are the candidates. Regarding DL strategy, convolutional neural network (CNN) and recurrent neural network (RNN) models are nominated. Referring to CNN, we borrow the AlexNet architecture with the input size being an image of 12 rows. As for RNN, we employ the long-short term memory units with inputs having a dimension of 12. Concerning ML methods, the models examined include support vector machine, logistic regression, boosting, k-nearest neighbor, decision tree, and random forest.

**2. Training Features:** Three experiments are carried out for feature selection. One includes only the amplitude feature in the model training. Another takes only the timing feature into consideration. To be more specific, we utilize not only the timing map but also the averaged heart rate and heart rate variation in this experiment. The other leverages both amplitude and timing feature maps for training.

To determine the best model for each evaluated class, we run the cross-validation tests on all the designed experiments. We split the data into 80% for training and 20% for testing. In addition, to address the issue of under-represented classes, or named the imbalanced classes, we randomly pick the samples from the class that have larger sample size to match the size of the smaller group.

## 2.5. Arrhythmia Decision

We proposed a strategy to diagnose the arrhythmias of a patient that is adaptable to the ECG data length. If the given data length is shorter than 20s, then we construct one representative ECG following the procedure in 2.2, 2.3 and determine the existence of each arrhythmia type by running through the best trained binary classifiers of each class.

For data length exceeding 20s, we randomly select  $N$  number of data segments to quantify  $N$  representations. Given the original data length  $L$  in the unit of seconds, the number  $N$  is derived from the proposed expression:

$$N = \begin{cases} \lceil \frac{L}{10} \rceil & \text{if } \frac{L}{10} < 100, \\ 100 & \text{if } \frac{L}{10} \geq 100. \end{cases}$$

For  $N$  representations, we compute the arrhythmia labels across all evaluated classes, and consider the scores of the classifier as the indicator. For each arrhythmia type, if the maximum score of these  $N$  representations is greater than 0.1, then we optimistically assume that the patient has the corresponding arrhythmia type diagnosed.

### 3. Results

We elaborate the outcomes of model training and feature selection experiments in 3.1 and display the scores of cross-validation and official tests in 3.2.

#### 3.1. Training Experiments

From the results of our cross-validation tests, boosting classifiers have the most robust and top-notch performance for all the evaluated classes. The second place goes to the random forest method, whereas deep learning models are inclined to overfitting.

With respect to the feature selection experiments, we discovered that the timing-deviated arrhythmia types showcase the best performance with solely the timing features. Also, we found that the abnormal waveform-based arrhythmias favor the amplitude feature maps. Combining amplitude and timing feature maps does not significantly improve the performance of the classifiers.

#### 3.2. Scores

Dataset	Runtime (hour:minute:second)	Score
CPSC	≈0:20:00	0.455
CPSC_2	≈0:30:00	0.402
Georgia	≈1:00:00	0.456
PTB	≈0:10:00	-2.589
PTB-XL	≈ 1:30:00	0.173
St. Petersburg	≈ 0:30:00	0.340
Official	1:55:00	0.406

Table 1. Performance of the proposed model.

We exhibit the cross-validation results of each dataset and the current official score in Table 1. Judging from the runtime, we show that our proposed model is competitive (compared to other challengers’ models). Assessing the scores, we believe that the proposed model has successfully learned the features since the weighting scores of the cross-validation tests is similar to the current official score.

### 4. Discussion and Conclusions

Our team has delivered an interpretable multi-label arrhythmia classifier. The classifier is built of our devised

12-Lead ECG heartbeat segmentation and feature generation algorithms. We demonstrate that different types of arrhythmias favor the corresponding features, either the amplitude or the timing maps. For future work, we aim at optimizing the arrhythmia prediction and representation learning algorithms.

### References

- [1] Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Das SR, et al. Heart Disease and Stroke Statistics – 2019 Update: a report From the American Heart Association. *Circulation* 2019;.
- [2] Kligfield P, Gettes LS, Bailey JJ, Childers R, Deal BJ, Hancock EW, Van Herpen G, Kors JA, Macfarlane P, Mirvis DM, 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* 2007;49(10):1109–1127.
- [3] Hsu PY, Cheng CK. Arrhythmia classification using deep learning and machine learning with features extracted from waveform-based signal processing. In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2020; 292–295.
- [4] Acharya UR, Oh SL, Hagiwara Y, Tan JH, Adam M, Gertych A, San Tan R. A deep convolutional neural network model to classify heartbeats. *Computers in biology and medicine* 2017;89:389–396.
- [5] Perez Alday EA, Gu A, Shah A, Robichaux C, Wong AKI, Liu C, Liu F, Rad BA, Elola A, Seyedi S, Li Q, Sharma A, Clifford GD, Reyna MA. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. *Journal of Physiological Measurements* 2020;.
- [6] Chang E, Cheng CK, Gupta A, Hsu PH, Hsu PY, Liu HL, Moffitt A, Ren A, Tsaur I, Wang S. Cuff-less blood pressure monitoring with a 3-axis accelerometer. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2019; 6834–6837.
- [7] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 2004;13(4):600–612.

Address for correspondence:

Po-Ya Hsu  
9500 Gilman Drive 0404, La Jolla, CA, USA  
p8hsu@eng.ucsd.edu