A Topology Informed Random Forest Classifier for ECG Classification

Paul Samuel Ignacio, Jay-Anne Bulauan, John Rick Manzanares

University of the Philippines Baguio, Baguio City, Philippines

Abstract

This paper accompanies Team Cordi-Ak's entry to the classification of 12-lead ECGs for the PhysioNet/Computing in Cardiology Challenge 2020. Our approach leverages mathematically computable topological signatures of 12-lead ECGs as proxy for features informed by medical expertise to train a two-level random forest model in a multi-class classification task. We view ECGs as multivariate time series data and convert different segments and groupings of leads to point cloud embeddings. This stores both local and global structures of ECGs, and encodes periodic information as attractor cycles in high-dimensional space. We then employ topological data analysis on these embeddings to extract topological features based on different summaries available in the literature. We supplement these features with demographic data and statistical moments of RR intervals based on the Pan-Tompkins algorithm for each lead to train the classifier. Our classifier achieved a challenge validation score of 0.304, and a final score of -0.113 on the full hidden test data, placing us 37th out of 41 officially ranked teams that participated in this year's Challenge.

1. Introduction

Cardiovascular diseases lead the causes of death worldwide [1]. Early and accurate diagnosis of cardiac conditions are prerequisite to the development of an appropriate and personalized treatment program [2]. This in turn increases the chances of survivability or successful management of the specific cardiac condition. The diagnosis of cardiac conditions relies on the rigorous and manual analysis of a patient's 12-lead electrocardiogram reading as different cardiovascular diseases have different causes and require different interventions [3]. Evidently, this is time-consuming and requires interpretation provided by highly skilled personnel with similarly high degree of training.

The PhysioNet/CinC Challenge 2020 focused on automated, open-source approaches for classifying cardiac abnormalities from 12-lead ECGs [4, 5]. Our entry to this challenge leverages mathematically computable topological signatures of 12-lead ECGs as proxy for features informed by medical expertise to train a random forest model in a multi-class classification task. As has been shown for detecting Atrial Fibrillation using single-lead ECGs [6], this approach verifies the existence and viability of signal in the topology of ECGs for improving diagnosis of cardiac conditions. Upscaling this to the use of all 12 leads of a standard ECG to diagnose multiple heart conditions improves accessibility to automated diagnostics by reducing expert-dependent input in feature extraction.

2. Methods

We use a standard pipeline for examining time series data using topological data analysis. We first generate point cloud embeddings from the ECG data, then extract topology-based features using tools from algebraic topology, and finally employ these features to train a two-level random forest classifier. Figure 1 illustrates this pipeline.

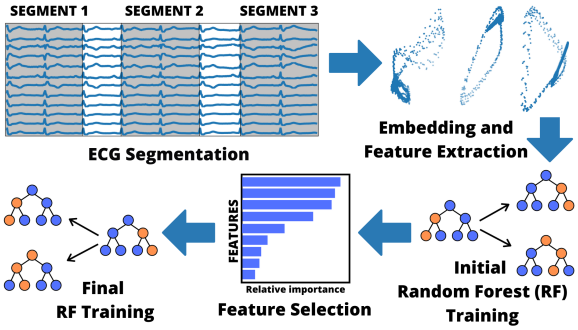

Figure 1. Pipeline. We convert time series data to point cloud embeddings from which topological signatures can be extracted via persistent homology, and used for machine learning.

2.1. Point Cloud Embedding

To generate point cloud embeddings, we consider each ECG reading as a sequence of multi-dimensional vectors where leads correspond to coordinates. Each time-slice of a 12-lead ECG represents a vector in high-dimensional Euclidean space, and each periodic signal in an ECG is embedded as an attractor cycle.

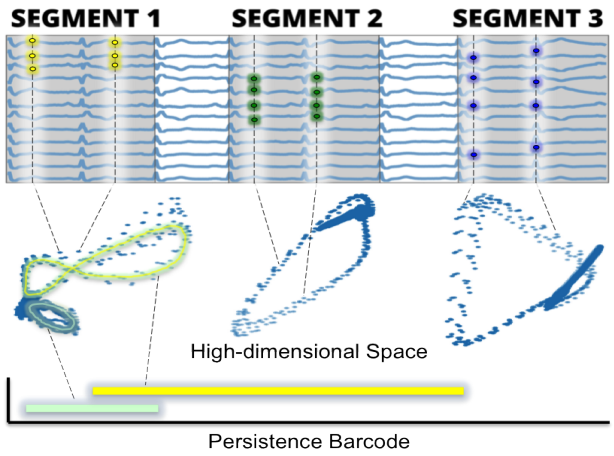

Figure 2. Each lead grouping from an ECG segment is sampled, then mapped to a point cloud whose topological features are computed and recorded as a barcode.

We unwrap characteristics of ECG readings by considering point cloud embeddings generated from different groupings of lead segments. The idea is to record characteristics that become more pronounced after filtering out excessive or redundant information from multiple leads. We consider three segments consisting of 1800 time points respectively from the beginning, middle, and ending portions of the ECG from which different lead groupings will be extracted to generate point clouds. See Figure 2. It is worth noting that due to the variable lengths in the ECG readings, the middle and end segments across ECG readings may refer to different time points. However, we argue that because of the overall periodic behavior of the cardiac cycle, and since the length of the segments being considered spans multiple periods, the point cloud embeddings generated from the captured middle and end segments provide good representations of the topology of the corresponding portions of the ECG readings. All things considered, the inclusion of the middle and end segments provides additional topological information about the ECGs.

Within each segment, we then consider 7 different groupings of leads to generate several point cloud embeddings. To reduce computational costs, we represent each group of leads by 300 equally spaced time slices within the segments. The first grouping uses all 12 leads in a standard ECG and represents the overall topology of the ECG segment. The other groupings are constructed based on two criteria: i. groups must collectively span all 12 leads; ii. some groups may represent collections of leads described in the literature as references for diagnosing specific cardiac conditions belonging to the original 8 classes identified in the unofficial phase of the challenge. Table 1 provides the different lead groupings considered in every segment.

	1	2	3	4	5	6	7	8	9	10	11	12
1	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
2	✔	✔	✔									
3				✔	✔	✔	✔					
4								✔	✔	✔	✔	✔
5	✔						✔				✔	✔
6			✔	✔		✔			✔			
7		✔			✔					✔	✔	

Table 1. Leads are grouped either to span all 12 leads of an ECG, or to represent collections of leads used in practice as references for diagnosing specific cardiac conditions.

2.2. Feature Extraction

We examine each point cloud embedding using topological data analysis, particularly via *persistent homology* [7]. A quick introduction to this approach with accompanying similar application is found in [6]. Succinctly, we induce a distance-parameterized monotonic sequence of abstract simplicial complexes revealing a multi-scale record of the evolving topological signatures of the underlying point cloud. In this application, these signatures pertain to connectivity and periodicity information about the corresponding ECG segment. We track topological information that persist across different parameter values and record it as a collection of bars, called *persistence barcodes*. In this work, we are only concerned with persistent homological features from dimensions 0 and 1, and use the python package RIPSER [8] to compute the barcodes. We derive statistical features from persistence barcodes and other persistence-based summaries such as *landscapes* [9], and *entropy* [10]. We summarize these below, where $\overline{x}$ refers to the average, SD the standard deviation, Sk the skewness, and $Kurt$ the kurtosis for the collection of values.

1. Dimension 0 and 1 Barcode Statistics.
   (a) $\overline{x}, SD, Sk$, and $Kurt$ for dimension 0 persistence, birth and death time, and dimension 1 persistence.
   (b) Sums of persistence in dimensions 0 and 1.
2. Dimension 0 and 1 Truncated Barcode Statistics where the most persistent bar is removed.
   (a) $\overline{x}, SD, Sk$, and $Kurt$ for dimension 0 persistence, birth and death time, and dimension 1 persistence.
   (b) Sum of dimension 1 persistence.
3. Dimension 1 Persistence Landscape Statistics.
   (a) Number of layers in the landscapes.
   (b) $\overline{x}, SD, Sk$, and $Kurt$ for the number of peaks and number of valleys per layer, and percentage of area under a layer relative to that above it.
4. Dimension 0 and 1 persistence entropy.

This feature set is extracted from each point cloud generated from the lead groupings in every segment. Finally, we include demographic data and statistical moments of RR intervals for each lead extracted using the algorithm supplied by the Challenge organizers [11]. A total of 1238 features are considered per ECG.

2.3. Classifier Training

We use a two-level random forest classifier. A first level random forest is trained to classify between *scored* and *unscored* classes based on the Challenge-provided lists. The second level uses two random forests, one each for scored and unscored classes. All forests have a maximum depth of 20 and use the square root function for the maximum number of features.

We scan the labels of each ECG in the training set and determine if at least one label belongs to the scored classes. In such case, the first label of an ECG that belongs to the scored classes is taken as its assigned label, and the ECG is included in the training subset for the scored classes. Otherwise, it is placed in the training subset for the unscored classes. Using a binary list, we keep track of which ECGs belong to either scored and unscored classes, and use this list as training labels for the first level random forest classifier. We emphasize that across all levels, we only use the features described in the previous section for training. Features are ranked by importance using Scikit-learn's [12] built-in functions and a shortlist is used for re-training.

Each second-level forest is trained using the appropriate subset, and features are again ranked and shortlisted for re-training. Table 2 shows the non-optimized parameter values for each random forest. We note that parameters for the random forest that classifies the unscored classes have smaller values to control computational costs due to significantly larger number of unscored classes.

RF Level	# Features	# Trees
1	400	1000
2-1	400	600
2-2	200	200

Table 2. Parameters for Random Forest Training.

3. Results

Table 3 reports the scores of our classifier on a 3-fold cross validation over the full training set.

Fold	AUPRC	AUROC	Acc.	F_1	Challenge
1	0.319	0.845	0.208	0.257	0.220
2	0.327	0.848	0.214	0.256	0.219
3	0.321	0.845	0.205	0.253	0.217
Average	0.322	0.846	0.209	0.255	0.219
Std. Dev.	0.004	0.002	0.005	0.002	0.002

Table 3. 3-fold cross validation scores on the training set.

Figure 3 shows the metric scores by class on the validation dataset as provided by the Challenge organizers, and the test scores across different databases are reported in Figure 4.

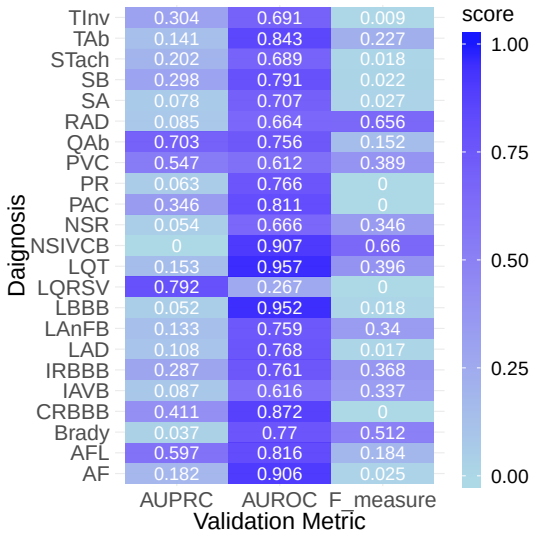

Figure 3. Scores by class on the validation dataset.

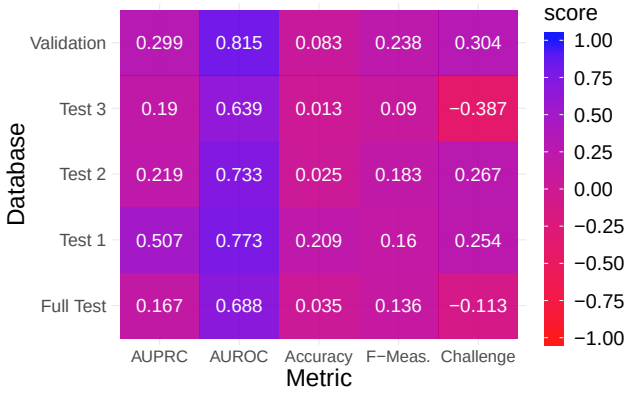

Figure 4. Classifier scores across different databases.

Our final ranking (Cordi-Ak) is 37th out of 41 officially ranked teams that participated in this year's Challenge.

4. Discussions

We employed a random forest classifier as we believe it mimics what is practiced at large in the community of experts in coming up with a collective and community-accepted diagnosis of cardiac conditions. However, as the significant portion of our work is about extracting mathematically computable topological features from 12-lead ECGs, other machine or deep learning algorithms may be trained using these features to develop new models.

Computational obstructions, some leading to the eventual abortion of some of our Challenge submissions, abound in our approach due to limited available technology in extracting topological information from point cloud data. Our response to this challenge was a two-fold reduc-

tion of computational costs by first segmenting the ECGs instead of using the entire recordings, then sampling within these segments. It must be noted however that this choice in turn ignores a large amount of information that could significantly affect the method.

Our moderately high AUROC scores across all validations and tests suggest that the topology-informed classifier performs relatively well in discriminating condition-positive from condition-negative ECGs, that is, the classifier is more likely to produce correct positive diagnosis than to classify an ECG as having a condition that it has not. This reveals that our topology-informed classifier is indeed able to capture crucial information for correct positive diagnosis. However, we see that the classifier is prone to make incorrect negative diagnosis as evidenced by the significantly smaller F_1 scores. To put the Challenge score on the validation dataset in perspective of our underlying objective, the Challenge-released baseline random forest classifier, trained on features similar to our non-topological features, received a validation Challenge score of 0.076. Finally, while the classifier performed consistently on the first two test databases, particularly on the positive Challenge scores, the reduced performance and larger negative Challenge score on the third database, which is matched to the second database in terms of demographics and prevalence of classes, reveal some generalizability concerns for the trained classifier. This trend is inherited in the full test set as the third database dominates the full test set constitution, and is not unique to our classifier.

5. Conclusion and Future Plans

We examined whether mathematically computable topological information embedded within 12-lead ECG readings contain signal that can be tapped to improve cardiac diagnosis. This is important as it improves accessibility to automated diagnostics by reducing expert-dependent input in feature extraction. Despite the computational obstructions we encountered that limited our opportunities to properly test and tune our model, we find positive evidences to support this hypothesis such as the ability of the topology-based features to train a classifier model with high true positive rates, and the marked positive difference in the performance of such classifier relative to baseline models trained on features similar to our non-topological features. On the other hand, we also learned about specific issues on generalizability and other weaknesses of the topology-informed classifier. We plan to further investigate these in the next PhysioNet/CinC Challenge.

Acknowledgments

We thank Prof. David Uminsky for the useful discussions to improve our approach. We also acknowledge support from the University of the Philippines Baguio.

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;e56–e528.

[2] Melamed RJ, Tillmann A, Kufleitner HE, Thürmer U, Dürsch M. Evaluating the efficacy of an education and treatment program for patients with coronary heart disease. Deutsches Aerzteblatt Online 2014;111(47):802–808.

[3] 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. J Amer Coll Cardiol 2007;49(10):1109–1127.

[4] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–e220.

[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. Physiol Meas ;(In Press).

[6] Ignacio PS, Dunstan C, Escobar E, Trujillo L, Uminsky D. Classification of single-lead electrocardiograms: TDA informed machine learning. In 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA). 2019; 1241–1246.

[7] Zomorodian A, Carlsson G. Computing persistent homology. Discret Comp Geom 2005;33(2):249–274.

[8] Tralie C, Saul N, Bar-On R. Ripser.py: A lean Persistent Homology library for Python. J Open Source Softw Sep 2018;3(29):925.

[9] Bubenik P. Statistical topological data analysis using persistence landscapes. J Mach Learn Res 2015;16:77–102.

[10] Atienza N, Gonzalez-Diaz R, Soriano-Trigueros M. A new entropy based summary function for topological data analysis. Electron Notes Discret Math 2018;68:113–118.

[11] Vest AN, Li Q, Liu C, Nemati S, Da Poian G, Shah AJ, Clifford GD. An open source benchmarked toolbox for cardiovascular waveform and interval analysis. Physiol Meas 2018;39(10):105004.

[12] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;12:2825–2830.

Address for correspondence:

Paul Samuel Ignacio, ppignacio@up.edu.ph
University of the Philippines Baguio,
Governor Pack Road, Baguio City, Philippines 2600