Leveraging Period-Specific Variations in ECG Topology for Classification Tasks

Paul Samuel Ignacio

University of the Philippines Baguio, Baguio City, Philippines

Abstract

We explore whether specific time-varying shape characteristics of electrocardiograms can be tapped to inform computational approaches in classifying cardiac abnormalities. In particular, we train a random forest classifier on features derived from relative differences between algebraically-computable topological signatures of consecutive segments within ECGs. We convert segments of ECGs as point cloud embeddings in high-dimensional space, extract their topological summaries, and compare these via statistical descriptors and different metrics. As part of the PhysioNet/Computing in Cardiology Challenge 2021, we (Team Cordi-Ak) test this approach across full- and reduced-lead ECGs. Using the Challenge's evaluation metric, our classifiers received scores of -0.06, -0.07, -0.08, -0.08, and -0.10 (consistently ranked 35th out of 39 official entries) for the 12-lead, 6-lead, 4-lead, 3-lead, and 2-lead versions of the hidden test set.

1. Introduction

Access to early and accurate diagnosis of cardiac conditions affords options to ensure survivability or promote tailor-fit treatment programs. As such, there abound many initiatives that explore automation of classifying cardiac abnormalities based on readily available, non-intrusive, and inexpensive sources of data by tapping fast-developing machine and deep learning technologies [1–7]. Needless to say, the success of these initiatives entails far-reaching influence on the future of Cardiology as it promotes a paradigm shift in the way cardiac conditions are studied by significantly cutting down on temporal and highly specialized skill requirements for obtaining accurate diagnoses.

In this work, we continue our inquiry on the utility and viability of topological information embedded within electrocardiogram (ECG) readings in informing methods for classification tasks. We build on our work in [8] where, inspired by the 2017 Physionet Challenge [1], we used topological features induced from time-delayed embeddings of single-lead ECGs to detect Atrial Fibrillation (AF), and our previous participation in the 2020 Physionet Challenge [6] where we used statistical moments from topological signatures extracted from point cloud embeddings representing segments of collections of ECG leads to train a two-level random forest model in a multi-class classification task. The primary deviations in the current approach are: (i) we use the groupings of leads provided by the 2021 Physionet Challenge [7] instead of the proposed groupings that we constructed in [9] to collectively span the full ECG recording and incorporate specific collections of leads described in the literature as references for diagnosing specific cardiac conditions; and (ii) we compute differences in the extracted topological signatures between consecutive segments instead of the actual signatures themselves as base for feature engineering. For a comprehensive description of the Challenges, including the data and specific rules that shaped our approaches, we refer the reader to [6, 7].

2. Methods

Our pipeline follows the standard approach in TDA-informed machine learning: model the data via abstract algebraic objects, extract homology-based signatures for feature engineering, and employ machine learning algorithms and strategies for classification and evaluation.

2.1. Segmentation and Representation

We begin by partitioning each full and reduced-lead ECG recordings into segments, each covering 750 time points to represent roughly three periods of a regular cardiac cycle recorded in the ECG. The choice of including three periods per segment is so that each segment will hopefully capture at least one non-abnormal period of the cardiac rhythm that can be used as basis for anomaly detection in the adjacent periods. The distribution in the lengths of all recordings in the training data reveals that a significant majority of all recordings contain fewer than 5000 time points. Thus, to ensure uniformity and comparability of topological features observed across ECGs, we consider the first five consecutive segments in every recording.

Where possible, to minimize the inherent multicollinearity among the leads, we perform Principal Component Analysis (PCA) and take the first four principal components accounting for about 98% of explained variance. Otherwise, we treat actual values in reduced-lead

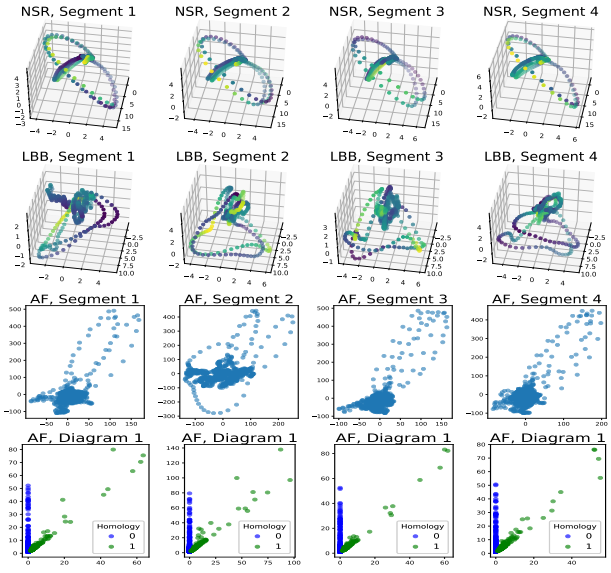

Figure 1. Consecutive segments of ECGs are transformed as point clouds objects using four principal components (top 2 rows; points are colored using the fourth principal component). For 2- and 3-lead recordings, actual values are used as coordinates to generate the point clouds (third row). Persistence diagrams are then computed to extract topological signatures of the point clouds (bottom row).

recordings as components. We use these components as coordinates to represent each segment as point cloud objects (see Figure 1). Finally, we use simplicial complexes as combinatorial models that encode the topology of the point clouds as well as the local similarity profiles of points in the high-dimensional representation with respect to the euclidean distance, and persistent homology to capture evolving topological signatures of the combinatorial model recorded as persistence diagrams (see Figure 1). We refer the interested reader to [10] for a quick introduction to this approach in the context of time series classification.

2.2. Feature Extraction and Engineering

Using the computed persistence diagrams from different segments, we craft features by taking standard statistical descriptors on observed differences between diagrams from consecutive segments. We only focus on topological signatures in dimensions 0 and 1 corresponding to clustering and periodicity information embedded in the point cloud representation. We summarize the features below.

1. Minimum, range, mean, standard deviation, skewness, and kurtosis of the collection of dimension 0 bottleneck distance between consecutive segments computed using the LUMÁWIG algorithm [11].
2. Minimum, maximum, mean, standard deviation, skewness, and kurtosis of the collection of dimension 1 persistent entropy [12] within segments.
3. Minimum, maximum, mean, standard deviation, skewness, and kurtosis of the collection of absolute differences in dimension 1 persistent entropy between segments.
4. Minimum, maximum, mean, standard deviation, skewness, and kurtosis of the collection of differences in persistence landscape [13] features (see [9]) between segments.
5. Minimum, maximum, mean, and standard deviation of the collection of differences in persistence diagram (barcode) [14] features (see [9]) between segments.

To evaluate the utility and practicality of our crafted features, we include demographic data (age and sex), statistical moments of RR intervals for each lead extracted using Vest et al.'s toolbox [15], and some naive baseline features (root mean square (RMS) per lead). We include the static topological features listed second in the list above to benchmark the performance of variation-based features against non-variation-based topological features.

2.3. Classifier Training

We train a random forest classifier over the public training data on 200 trees with a maximum of 20 leaf nodes per tree, where leaf nodes contain no fewer than 5 samples, and are split when a minimum of 10 samples is reached. Due to running time and memory usage limits, we cap to 5000 the number of ECG recordings used in training for classes (Normal Sinus Rhythm (NSR) and Sinus Bradycardia (SB)) that significantly outnumber other classes.

To examine how variation-based features fare against other features, we rank all features by importance using scikit-learn [16], determine the type distribution of the top 100 features, and re-train the classifier using the top 100 features. The distribution of the top 100 features used for final training is summarized in Figure 2.

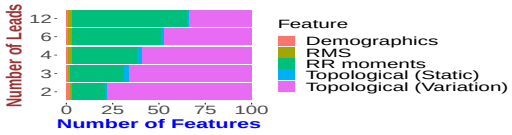

Figure 2. Breakdown of features in the top 100 features ranked by importance and used in the final training.

Unfortunately we are not able to optimize the classifier's parameters based on validation performance due to failure of all but one of our submissions to complete both training and testing at the Challenge's computing system.

3. Results

We report the average Challenge scores as well as other validation metrics obtained over a 5-fold stratified cross validation on the entire public training data in Table 1.

We also report using the heatmap in Figure 3 the 5-fold

Leads	Training	Validation	Test	Ranking
12	0.23 ± 0.00	0.17	-0.06	35/39
6	0.22 ± 0.00	0.15	-0.07	35/39
4	0.20 ± 0.01	0.15	-0.08	35/39
3	0.19 ± 0.01	0.15	-0.08	35/39
2	0.19 ± 0.01	0.13	-0.10	35/39

Table 1. Challenge scores for our final selected entry (team Cordi-Ak) using 5-fold cross validation on the public training set, repeated scoring on the hidden validation set, and one-time scoring on the hidden test set as well as the ranking on the hidden test set among official entries.

cross validation average for the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC) across all cardiac diagnosis classes. We omit the obtained F_1-scores due to very high correlation with AUPRC across all models.

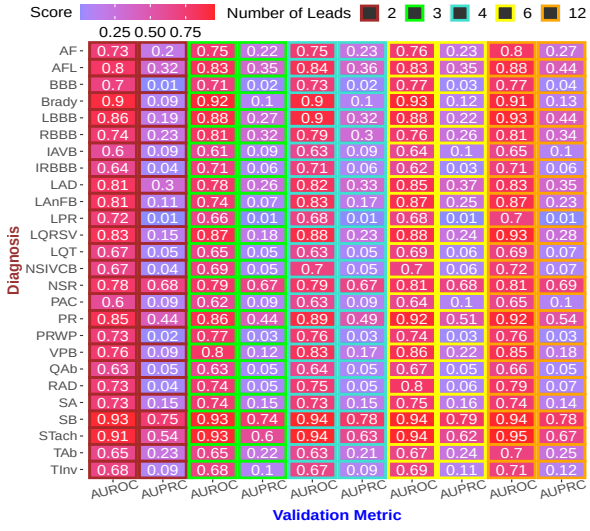

Figure 3. Pairs of columns color-coded by the number of leads used, report the average AUROC and AUPRC obtained from a 5-fold cross validation over the training set.

Finally, we test the 5-fold cross validation models on the training data from different Challenge databases (CPSC and CPSC-Extra [2], INCART [17], PTB [18] and PTB XL [19], Georgia [6, 7], Chapman-Shaoxing [20] and Ningbo [21]) and report this in Table 2. The organizer-provided Challenge scores on hidden test sets (CPSC, G12EC, Undisclosed, and UMich [6, 7]) also appear in Table 2.

4. Discussion and Conclusions

As with our experience during the 2020 Physionet Challenge, our topology-based approach proved computationally expensive. As such, we implemented some interventions to manage computational costs. For example, while each segment considered covers 750 time points, we skip every other point in the sequence, thereby halving the number of uniformly sampled points that generate the point cloud. While, in principle, this method should still preserve the global structure of the resulting representation, it does introduce an unexamined loss of information on the local dynamics of the time series data and may result to missing important portions of an anomaly event.

We also note that the trained classifiers exhibit symptoms that reflect our non-optimization of tuning parameters. For example, despite training with balanced weighting across classes, capping the number of NSR or SB used for training still introduced notable improvements in the classification accuracy of other less represented classes at the moderate expense of these two overly represented classes. This poses an intriguing question on the robustness of topology variation-based features from artefacts

Database	Validation Metric	Number of Leads 2	3	4	6	12
5-fold cross validation models applied to databases						
CPSC	AUROC	0.90	0.91	0.90	0.90	0.90
	AUPRC	0.74	0.75	0.74	0.75	0.76
	F1-Score	0.17	0.18	0.17	0.17	0.18
	Challenge	0.32	0.33	0.32	0.35	0.35
CPSC Extra	AUROC	0.95	0.94	0.95	0.94	0.93
	AUPRC	0.77	0.78	0.77	0.77	0.78
	F1-Score	0.60	0.57	0.60	0.55	0.56
	Challenge	0.78	0.80	0.78	0.79	0.79
INCART	AUROC	0.94	0.97	0.97	0.93	0.95
	AUPRC	0.86	0.87	0.90	0.86	0.85
	F1-Score	0.83	0.79	0.75	0.71	0.61
	Challenge	0.74	0.75	0.70	0.65	0.63
PTB	AUROC	0.88	0.89	0.89	0.89	0.92
	AUPRC	0.69	0.74	0.75	0.75	0.81
	F1-Score	0.26	0.28	0.37	0.36	0.56
	Challenge	-1.38	-1.71	-1.79	-1.81	-1.83
PTB XL	AUROC	0.91	0.91	0.91	0.92	0.92
	AUPRC	0.71	0.71	0.72	0.72	0.73
	F1-Score	0.61	0.60	0.60	0.57	0.60
	Challenge	0.50	0.45	0.43	0.43	0.44
Georgia	AUROC	0.94	0.94	0.94	0.94	0.95
	AUPRC	0.80	0.81	0.81	0.81	0.83
	F1-Score	0.69	0.69	0.68	0.68	0.71
	Challenge	0.54	0.60	0.60	0.58	0.64
Chapman-Shaoxing	AUROC	0.91	0.93	0.93	0.93	0.94
	AUPRC	0.76	0.78	0.78	0.79	0.80
	F1-Score	0.56	0.57	0.57	0.56	0.58
	Challenge	0.54	0.56	0.58	0.63	0.63
Ningbo	AUROC	0.92	0.92	0.93	0.93	0.94
	AUPRC	0.77	0.78	0.78	0.79	0.80
	F1-Score	0.67	0.67	0.68	0.69	0.70
	Challenge	0.56	0.58	0.61	0.65	0.66
Organizer-provided scores on hidden test sets						
CPSC	AUROC	0.64	0.65	0.65	0.63	0.65
	AUPRC	0.23	0.27	0.26	0.24	0.26
	F1-Score	0.07	0.08	0.08	0.08	0.08
	Challenge	0.18	0.17	0.16	0.17	0.18
G12EC	AUROC	0.69	0.70	0.71	0.70	0.69
	AUPRC	0.14	0.15	0.16	0.16	0.16
	F1-Score	0.07	0.07	0.08	0.08	0.08
	Challenge	0.13	0.15	0.15	0.15	0.18
Undisclosed	AUROC	0.64	0.66	0.66	0.65	0.64
	AUPRC	0.17	0.18	0.18	0.18	0.18
	F1-Score	0.02	0.02	0.02	0.03	0.02
	Challenge	-0.41	-0.41	-0.41	-0.40	-0.43
UMich	AUROC	0.68	0.69	0.69	0.70	0.69
	AUPRC	0.15	0.16	0.16	0.17	0.16
	F1-Score	0.06	0.07	0.07	0.07	0.08
	Challenge	-0.04	-0.01	0.00	0.00	0.03

Table 2. Validation and test scores on different databases.

of model architecture or data imbalance. Meanwhile, the highly contrasting average 5-fold cross validation metric scores (see Table 1) against those from different databases (see Table 2) points to a considerable degree of overfitting.

With respect to Figure 3, the relatively high AUROC and AUPRC class scores, consistent across variable lead inclusions, suggest the presence of signal in the topological variations across segments within full and reduced-lead ECGs for diagnosing *specific* conditions. However, this must be balanced with the observation from Figure 2 that the number of ranking variation-based features decreases as the feature pool grows. This suggests the need to amplify this signal to maintain its model influence.

The current model seems to provide acceptable accuracy levels for classifying the normal sinus rhythm, sinus Bradycardia and Tachycardia correctly. It is interesting to note that the set of classes which the current model is able to classify at acceptable accuracy levels is almost completely disjoint from that of our earlier topology-informed model in [9] that uses actual topological signatures within segments rather than observed variations between them. This suggests that the two approaches indeed capture different information suited for classifying different classes, prompting a further examination on the complementarity, and the degree of such, between the two approaches.

References

[1] Clifford GD, Liu C, Moody B, Lehman LwH, Silva I, Li Q, et al. AF Classification from a Short Single Lead ECG Recording: the PhysioNet/Computing in Cardiology Challenge 2017. In 2017 Computing in Cardiology (CinC), volume 44. IEEE, 2017; 1–4.

[2] Liu F, Liu C, Zhao L, Zhang X, Wu X, Xu X, et al. An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection. Journal of Medical Imaging and Health Informatics 2018;8(7):1368—1373.

[3] Gao H, Liu C, Wang X, Zhao L, Shen Q, Ng EYK, et al. An Open-Access ECG Database for Algorithm Evaluation of QRS Detection and Heart Rate Estimation. Journal of Medical Imaging and Health Informatics 2019;9(9):1853–1858.

[4] Cai Z, Liu C, Gao H, Wang X, Zhao L, Shen Q, et al. An Open-Access Long-Term Wearable ECG Database for Premature Ventricular Contractions and Supraventricular Premature Beat Detection. Journal of Medical Imaging and Health Informatics 2020;10(11):2663–2667.

[5] Reyna MA, Alday EAP, Gu A, Liu C, Seyedi S, Rad AB, et al. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. In 2020 Computing in Cardiology, volume 47. IEEE, 2020; 1–4.

[6] Perez Alday EA, Gu A, Shah A, Robichaux C, Wong AKI, Liu C, et al. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. Physiological Measurement 2020;41.

[7] Reyna MA, Sadr N, Perez Alday EA, Gu A, Shah A, Robichaux C, et al. Will Two Do? Varying Dimensions in Electrocardiography: the PhysioNet/Computing in Cardiology Challenge 2021. Computing in Cardiology 2021;48:1–4.

[8] 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.

[9] Ignacio PS, Bulauan JA, Manzanares JR. A Topology Informed Random Forest Classifier for ECG Classification. In 2020 Computing in Cardiology. 2020; 1–4.

[10] Karan A, Kaygun A. Time Series Classification Via Topological Data Analysis. Expert Systems with Applications 2021;183:115326.

[11] Ignacio PS, Bulauan JA, Uminsky D. Lumáwig: An Efficient Algorithm for Dimension Zero Bottleneck Distance Computation in Topological Data Analysis. Algorithms 2020;13(11).

[12] 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.

[13] Bubenik P. Statistical Topological Data Analysis Using Persistence Landscapes. J Mach Learn Res 2015;16:77–102.

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

[15] Vest AN, Li Q, Liu C, Nemati S, Da Poian G, Shah AJ, et al. An Open Source Benchmarked Toolbox for Cardiovascular Waveform and Interval Analysis. Physiol Meas 2018;39(10):105004.

[16] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;12:2825–2830.

[17] Tihonenko V, Khaustov A, Ivanov S, Rivin A, Yakushenko E. St Petersburg INCART 12-lead Arrhythmia Database. PhysioBank PhysioToolkit and PhysioNet 2008;Doi: 10.13026/C2V88N.

[18] Bousseljot R, Kreiseler D, Schnabel A. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. Biomedizinische Technik 1995;40(S1):317–318.

[19] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, et al. PTB-XL, a Large Publicly Available Electrocardiography Dataset. Scientific Data 2020;7(1):1–15.

[20] Zheng J, Zhang J, Danioko S, Yao H, Guo H, Rakovski C. A 12-lead Electrocardiogram Database for Arrhythmia Research Covering More Than 10,000 Patients. Scientific Data 2020;7(48):1–8.

[21] Zheng J, Cui H, Struppa D, Zhang J, Yacoub SM, El-Askary H, et al. Optimal Multi-Stage Arrhythmia Classification Approach. Scientific Data 2020;10(2898):1–17.

Address for correspondence:

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