Shapelet Discovery for Atrial Fibrillation Detection

Saman Parvaneh*, Yale Chang*

Philips Research North America, Cambridge, MA, USA

Abstract

The main goal of this study is to develop an automatic classification algorithm for normal sinus rhythm (NSR) versus Atrial Fibrillation (AF) from a single channel short ECG segment. For this purpose, AF and NSR from PhysioNet/Computing in Cardiology Challenge 2017 training dataset were used in this study. RR intervals were extracted using GQRS algorithm, and RR time series with less than 30 beats were excluded from the analysis. The stratified split was applied to create a training set (NSR: 1521 and AF: 239) and test set (NSR: 1527 and AF:234). Shapelets were extracted by scanning RR time series in the training set and identifying statistically significant patterns. For classification, an XGBoost model was trained using the presence or absence of the top 100 shapelets. Using the top 100 significant shapelets (Shapelet length between 5 and 29), we achieved the area under the ROC curve (AUC) and the area under the precession recall curve (AUPRC) of 0.94 and 0.77 in discrimination between AF and NSR. Among the top 100 significant shapelets, we used all shapelets with length no greater than a certain threshold (maximum acceptable Shapelet length) for training different models. Increasing the number of shapelet features by varying the threshold from 5 to 30 in the model training improved AUC/AUPRC 0.91/0.68 to 0.94/0.77

1. Introduction

Atrial Fibrillation (AF) is a common cardiac arrhythmia with an estimated incidence of 2.7-6.1 million in the United States [1]. Furthermore, AF incidence and prevalence are likely to increase, especially with the aging of the population. ECG-based AF detectors mostly analyze atrial activities (e.g., absence of P waves) [2-4], ventricular responses (e.g., irregularity in RR intervals) [5-7], or both. Different features from RR intervals, such as heart rate variability (HRV), geometric representation, and entropy are used to differentiate between AF and Normal Sinus Rhythm (NSR) [7, 8]. Shapelets are time series sub-sequences that are maximally representative of a class [9]. In this paper, we aimed to use shapelet discovery to distinguish between AF and NSR. The shape of the discovered shapelet will enable the interpretation of results.

2. Method and Material

A block diagram of our proposed method is shown in Figure 1. Given an ECG recording, first QRS detection takes place, followed by a shapelet discovery. Top shapelets will be classified as AF or NSR using an XGBoost Classifier.

2.1. Data and Pre-processing

AF and NSR from PhysioNet/Computing in Cardiology Challenge 2017 training dataset were used in this study [10]. This database contains short single-lead ECG recordings recorded by a hand-held recording device. Details about the challenge dataset can be found in [10].

QRS complexes were detected using GQRS algorithm implementation in the WFDB toolbox [11] after removing baseline wander using a moving average filter. Detected R peaks in QRS detection steps were used to create RR time series. RR time series with less than 30 beats were excluded from the analysis. The stratified split was applied to create a training set (NSR: 1521 and AF: 239) and test set (NSR: 1527 and AF:234).

2.2. Shapelet Discovery

Shapelet is a subsequence of a time series that maximizes predictive power [9]. In this paper, the significant shapelets were identified by the statistically significant shapelet mining (S3M) algorithm [12]. S3M consists of four steps:

First, given a fixed shapelet length (the minimum and maximum shapelet length were set to $w_{min} = 5$ RR intervals and $w_{max} = 30$ RR intervals, respectively), all candidate shapelets were extracted by scanning RR time series from the training set using a sliding window with stride equal to one. For each candidate shapelet S, its distance with the time series of the n-th training sample T_n,

* *Authors contributed equally*

Figure 1. Block diagram of the proposed algorithm.

denoted as $\mathrm{dist}(S, T_n)$, is computed as the minimum Euclidean distance between S and all sub-sequences of T_n having an equal length with S

$$\mathrm{dist}(S, T_n) = \min_j \mathrm{EuclideanDist}(S, T_n[j: j + |S| - 1])$$

Where $T_n[j: j + |S| - 1]$ is the j-th subsequence of T_n of length $|S|$. Therefore, for the given shapelet S, its distances to N training samples were computed and are denoted as $\{d_1, \cdots, d_N\}$, where $d_n = \mathrm{dist}(S, T_n)$.

Second, applying a distance threshold θ would lead to a partition of the training dataset $\mathcal{T}$ into two subsets $\mathcal{T} = \mathcal{T}_\theta^- \uplus \mathcal{T}_\theta^+$, where

$$\mathcal{T}_\theta^- = \{T_n \in \mathcal{T} \mid d_n \leq \theta\}$$
$$\mathcal{T}_\theta^+ = \{T_n \in \mathcal{T} \mid d_n > \theta\}$$

Since the class labels of N training samples are available, which are denoted as $\{y_1, \cdots, y_N\}$, each choice of the threshold θ would give rise to a contingency table shown in Table 1.

Table 1. A 2 x 2 contingency table as used by the S3M method for shapelet mining.

Class label	$dist \leq \theta$	$dist > \theta$	Row totals
$y = 1$	a_S	b_S	n_1
$y = 0$	d_S	c_S	n_0
Column total	r_S	q_S	N

Third, to determine the optimal threshold θ^* for a given shapelet S, the statistical significance of the contingency table was assessed using the Pearson's χ^2 test. The p-value of the test can be written as

$$p = 1 - F_{\chi^2}\left(T_{\chi^2}(n, a_S, b_S, c_S, d_S)\right)$$

Where 1) F_{χ^2} denotes the cumulative density function (CDF) of a χ^2 distribution with one degree of freedom; and 2) T_{χ^2} is the test statistics of the χ^2 test

$$T_{\chi^2}(N, a_S, b_S, c_S, d_S) = \frac{N(a_S c_S - b_S d_S)^2}{(a_S + b_S)(c_S + d_S)(a_S + d_S)(b_S + c_S)}$$

Bonferroni correction was further applied to address the false positives induced by the multiple hypothesis testing. The Bonferroni-corrected significance threshold is written as

$$\alpha = \frac{\hat{\alpha}}{N(N+1)\sum_{w=w_{min}}^{w_{max}}(|T| - w + 1)}$$

Where the uncorrected significance level $\hat{\alpha}$ is often set to be 0.05, $\sum_{w=w_{min}}^{w_{max}}(|T| - w + 1)$ represents the number of candidate shapelets for each time series of length $|T|$.

Fourth, given the extremely large number of hypothesis test even for a small dataset (the number of tests would be at the order of millions when $N = 100$), Tarone's method was applied to efficiently prune those untestable shapelet candidates. Please refer to [12] for the technical details of Tarone's method.

A final set of statistically significant shapelets can be extracted from the training set following these four steps. Since applying the optimal threshold for each significant shapelet can create a partition of the dataset, these significant shapelets would also induce a patient by feature design matrix, where each feature would be the binary vector indicating the dataset partition.

2.3. Classification

Starting from the patient design matrix induced by the top 100 significant shapelets (the presence or absence of the top 100 shapelets), we trained an XGBoost binary classification model [13] to predict the binary label AF vs. NSR.

2.4. Algorithm Evaluation

The performance of the classifier was evaluated on the test set using the area under the ROC curve (AUC) and the area under the precession recall curve (AUPRC). We set the number of decision trees to be 200, where each tree has a depth equal to 3. The learning rate is set to be 0.1.

3. Results

A total of 138,492 significant shapelets were extracted from the training set. The top ten shapelets are shown in Figure 2. As shown in this figure, both mega RR variations (Figure 2, a-h) and micro RR variations (Figure 2, i and j) that were significantly different between AF and NSR groups are present in the top ten shapelets.

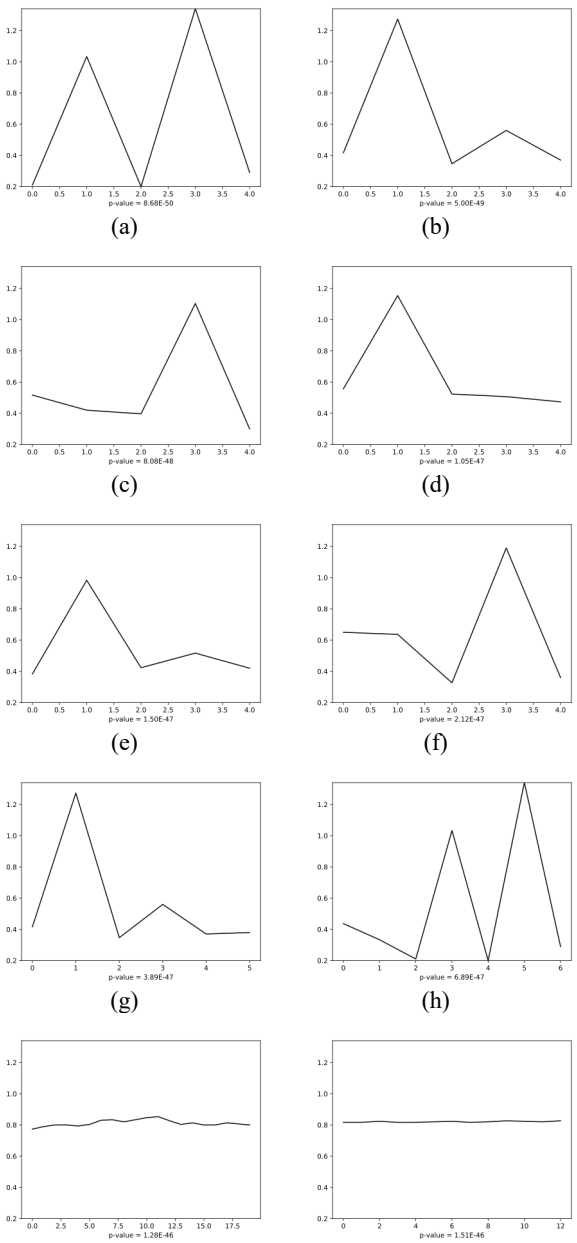

Figure 2. Top ten shapelets identified from RR intervals for discrimination between AF and NSR.

Using the top 100 significant shapelets (Shapelet length between 5 and 29), we achieved AUC and AUPRC of 0.94 and 0.77 in discrimination between AF and NSR. The plots of the receiver operating characteristic (ROC) curve and the precision-recall (PR) curve are shown in Figure 3.

Furthermore, among the top 100 significant shapelets, we used all shapelets with length no greater than a certain threshold (maximum acceptable Shapelet length) for training different models. Increasing the number of shapelet features by varying the threshold from 5 to 30 in the model training improved AUC/AUPRC 0.91/0.68 to 0.94/0.77 (Table 2).

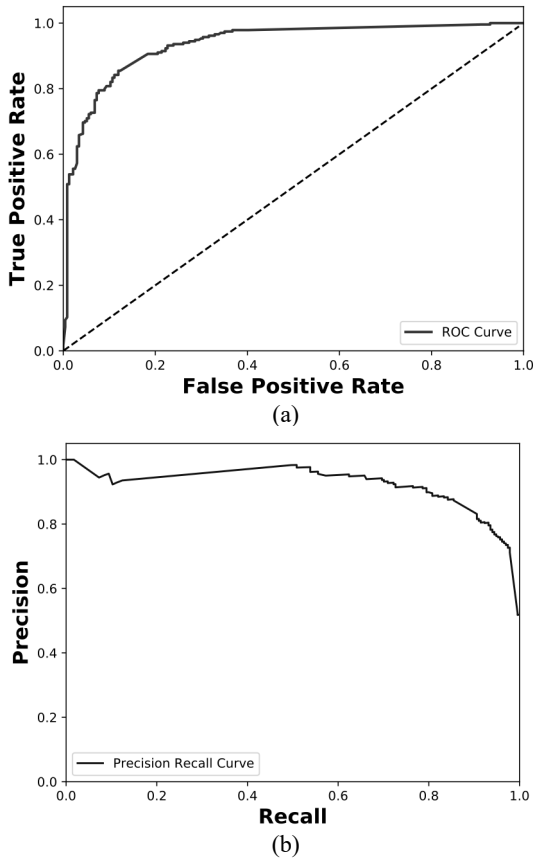

Figure 3. (a) the ROC curve and b) the precision-recall curve of the trained XGBoost model for prediction AF vs. NSR.

Table 2. The area under the ROC curve (AUC) and the area under the precession recall curve (AUPRC) by varying the maximum shapelet length from 5 to 30.

Maximum Shapelet Length	**5**	**10**	**20**	**30**
Number of Shapelets Used in Training	10	14	79	100
Test AUC	0.91	0.91	0.93	0.94
Test AUPRC	0.68	0.72	0.77	0.77

4. Conclusion

In this article, a shapelet discovery was combined with XGBoost classifier for distinguishing atrial fibrillation from normal sinus rhythm. The promising performance of the trained model demonstrates that shapelet-based features have a great potential to discriminate between AF and NSR. More work is needed to assess the application of the proposed method in discrimination between AF, NSR,

other rhythms, and noise.

References

[1] C. T. January *et al.*, "2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation," *Circulation,* p. CIR. 0000000000000041, 2014.

[2] P. E. Dilaveris and J. E. Gialafos, "P-wave dispersion: A novel predictor of paroxysmal atrial fibrillation," *Annals of Noninvasive Electrocardiology,* vol. 6, no. 2, pp. 159-165, 2001.

[3] N. Larburu, T. Lopetegi, and I. Romero, "Comparative study of algorithms for atrial fibrillation detection," in *2011 Computing in Cardiology*, 2011: IEEE, pp. 265-268.

[4] S. Parvaneh, M. R. Hashemi Golpayegani, M. Firoozabadi, and M. Haghjoo, "Predicting the spontaneous termination of atrial fibrillation based on Poincare section in the electrocardiogram phase space," *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine,* vol. 226, no. 1, pp. 3-20, 2012.

[5] S. Parvaneh, J. Rubin, A. Rahman, B. Conroy, and S. Babaeizadeh, "Analyzing single-lead short ECG recordings using dense convolutional neural networks and feature-based post-processing to detect atrial fibrillation," *Physiological Measurement,* vol. 39, no. 8, p. 084003, 2018.

[6] J. Rubin, S. Parvaneh, A. Rahman, B. Conroy, and S. Babaeizadeh, "Densely connected convolutional networks for detection of atrial fibrillation from short single-lead ECG recordings," *Journal of Electrocardiology,* 2018.

[7] M. Zabihi, A. B. Rad, A. K. Katsaggelos, S. Kiranyaz, S. Narkilahti, and M. Gabbouj, "Detection of atrial fibrillation in ECG hand-held devices using a random forest classifier," in *2017 Computing in Cardiology (CinC)*, 2017: IEEE, pp. 1-4.

[8] R. Alcaraz and J. J. Rieta, "A review on sample entropy applications for the non-invasive analysis of atrial fibrillation electrocardiograms," *Biomedical Signal Processing and Control,* vol. 5, no. 1, pp. 1-14, 2010.

[9] L. Ye and E. Keogh, "Time series shapelets: a new primitive for data mining," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 947-956.

[10] G. Clifford *et al.*, "AF classification from a short single lead ECG recording: the physioNet computing in cardiology challenge 2017," presented at the Computing in Cardiology Rennes-France, 2017.

[11] I. Silva and G. B. Moody, "An open-source toolbox for analysing and processing physionet databases in matlab and octave," *Journal of open research software,* vol. 2, no. 1, 2014.

[12] C. Bock, T. Gumbsch, M. Moor, B. Rieck, D. Roqueiro, and K. Borgwardt, "Association mapping in biomedical time series via statistically significant shapelet mining," *Bioinformatics,* vol. 34, no. 13, pp. i438-i446, 2018.

[13] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794.

Address for correspondence.
Saman Parvaneh / Yale Chang
222 Jacobs St, Cambridge, MA 02141, United States
parvaneh@ieee.org / yale.chang@philips.com