Classification of Atrial Fibrillation Using Multidisciplinary Features and Gradient Boosting

Sebastian D. Goodfellow[1], Andrew Goodwin[1], Robert Greer[1],

Peter C. Laussen[1], Mjaye Mazwi[1], Danny Eytan[1,2]

[1]Department of Critical Care Medicine, The Hospital for Sick Children, Toronto, Canada
[2]Rambam Medical Centre, Haifa, Israel

Abstract

Atrial fibrillation (AFib) is the most common tachyarrhythmia of the heart in adults and is associated with an increased risk for stroke and heart failure. It can be described as irregularly irregular, foci in the atrium that set up chaotic atrial circuits and irregular rapid contraction of the atrium with loss of consistent atrio-ventricular synchrony due to decremental conduction at the atrio-ventricular node. The challenge was to identify predictive features of ECG signals with variable time and spatial components. These features were extracted from 8528 single lead ECG recordings and then input to a gradient boosting classifier. The trained model could classify AFib with an F1 score of 0.83. Present in the dataset were three other rhythm classes; Normal Sinus Rhythm, Other, and Noisy. The F1 scores achieved for these classes were 0.91, 0.77, and 0.66 respectively.

1. Introduction

Atrial fibrillation (AFib) is a common arrhythmia detected in adult patients with critical illness [1,2]. A recent study reported an incidence of AFib of nearly 20% in a cohort of adult patients in a medical and surgical intensive care unit (ICU), but new-onset atrial fibrillation was subclinical or went undocumented in 8% of all ICU admissions [2]. AFib may be paroxysmal or intermittent, and transient episodes can go unrecognized. It is nevertheless an important arrhythmia to detect as it can be associated with thromboembolic complications such as stroke [3], and is associated with increased hospital mortality and longer length of stay [2].

Afib can be a difficult arrhythmia to detect clinically particularly if intermittent, and can be further confounded by artifacts in the ECG signal. The pulse of patients in AFib is characteristically described as being irregularly irregular. Accurate processing of the continuous ECG signal to detect AF from ICU ECG monitoring might enable earlier recognition and potentially improve outcomes. The PhysioNet challenge provides an opportunity to automatically detect AFib through processing features of the ECG signal.

2. Dataset Overview

The Physionet dataset consists of 8528 single lead ECG recordings, ranging in duration from 9 s to 61 s, that were sampled at 300 Hz. Each waveform has an associated label that was determined group of experts. The following four labels are present in the dataset; Normal Sinus Rhythm, Atrial Fibrillation, Other Rhythm, and Noisy.

3. Pre-processing Workflow

Before features could be extracted, a series of pre-processing steps were completed to remove noise and standardize the data.

First, the full waveform was filtered using a finite impulse response bandpass filter (SciPy [4]) with band limits of 3 Hz and 45 Hz. With the signal filtered, the R-peaks were determined using the Hamilton–Tompkins algorithm [5] as implemented in the Biosignal Processing in Python (BioSPPy) library. This process returned an array of picked R-peak times.

With the R-peaks determined, PQRST templates were extracted from the full waveform. A PQRST template contains the P-, QRS-, and T-waves and is defined as 250ms before the R-peak to 400ms after the R-peak. An example of extracted templates is presented in Figure 1.

Some waveforms in the dataset have a negative R-peak polarity. If the maximum amplitude of the median template (R-peak) was negative, then the waveform polarity was switched to ensure that the R-peaks were always positive. Next, waveforms were normalized to the maximum value of the median template (median R-peak amplitude). An

 ISSN: 2325-887X DOI:10.22489/CinC.2017.361-352

example of a normalized waveform is presented in Figure 1.

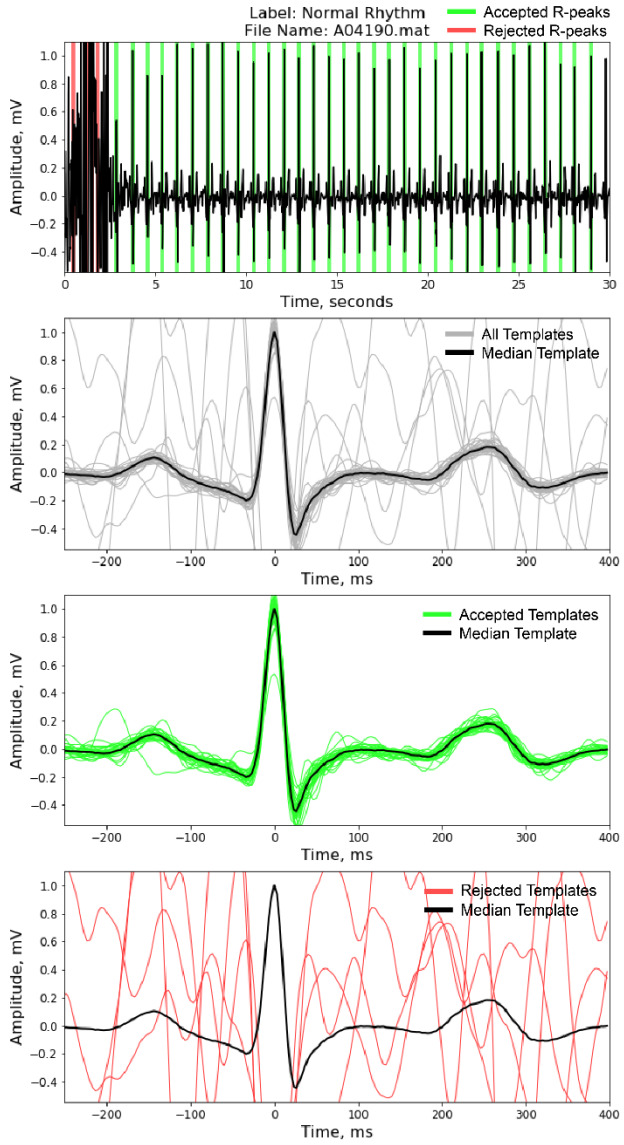

Figure 1. Example of R-peak filtering on a normalized waveform.

The final step was R-peak filtering to remove any noise or ectopic beats that the Hamilton–Tompkins algorithm mistakenly selected. This was done by calculating the correlation coefficient of a section of each template, 25 sample points before to 25 sample points after the R-peak, with that of the median template. If the correlation coefficient was below 0.9, the template and corresponding R-peak were rejected. An example of this process is presented in Figure 1. This waveform contains a transient burst of noise in the first few seconds from which the Hamilton–Tompkins algorithm detected some R-peaks. Because the shape of the templates associated with those R-peaks do not correlate with the median template, they were rejected. This is an important step given that R-peak to R-peak Interval (RRI) statistics can be quite sensitive to noise.

4. Feature Engineering

Over 300 features were extracted from each waveform and used in various arrangements to train the most accurate model. In this section, we will provide a general overview of the features, which were grouped into three main classes; (1) Full Waveform Features, (2) Template Features, and (3) RRI Features.

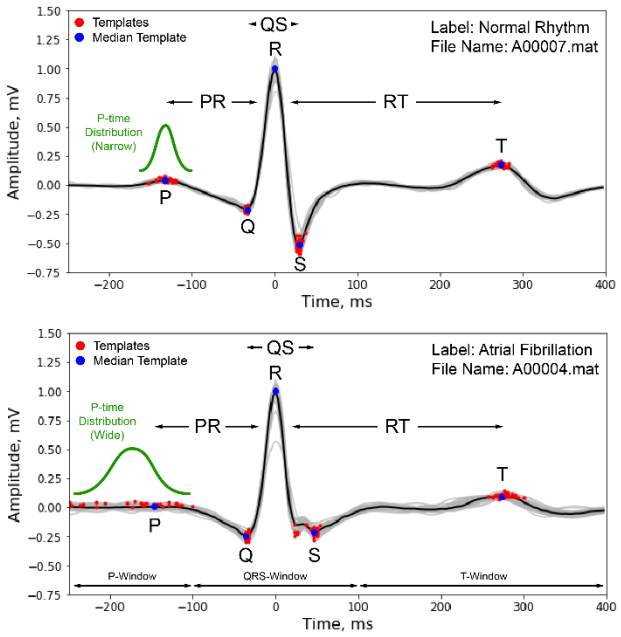

Figure 2. Example of picked templates for Normal Rhythm and Atrial Fibrillation.

4.1. Full Waveform Features

Full waveform features were extracted from the complete duration of the signal. Basic amplitude features included min, max, mean, median, standard deviation, skew and kurtosis. Additionally, the waveform duration was also included as a feature.

A more advanced set of full waveform features was generated from the stationary wavelet transform decomposition of the full waveform [6]. The transformation was done using the PyWavelets toolbox and the Daubechies 4 (db4) mother wavelet. The approximation and detail coefficients of decomposition levels 1 – 4 were used. On each set of coefficients, three max/mean power spectral density ratios were calculated for the following frequency bands: 3 – 10 Hz, 10 – 30 Hz, and 30 – 45 Hz. Additionally, the log entropy [6] and

Higuchi fractal dimension (PyEEG toolbox) were calculated.

4.2. Template Features

Template features were extracted from all templates that passes the filter process. The P-, Q-, R-, S-, and T-wave amplitudes and times were all picked for each template. Summary statistics, such as the mean, median, and standard deviation, were computed for both the amplitude and time of each wave. Additionally, the PR, QS, and RT interval times and the P-wave energy were computed for each template. These features proved useful for differentiating between Normal Rhythm and Atrial Fibrillation. For the Normal Rhythm example in Figure 2, the distribution of P-wave times is tight given that the signal contains a clear P-wave. Conversely the lack of a P-wave in an Atrial Fibrillation rhythm, the distribution of pick times is broader. The effectiveness of these features decreases as the signal to noise ratio (SNR) decreases.

4.3. RRI Features

From the filtered R-peaks, the following time series were calculated: RR Interval (RRI), RRI Velocity, and RRI Acceleration. From these time series, a wide range of heart rate variability features were extracted. These included standard heart rate variability statistics such as: max, min, median, mean, standard deviation, pNN20, and pNN50. We also calculated the standard deviations (SD1 and SD2) along the major and minor axes of an ellipse fit to the 2D scatter point data RRI_n and RRI_{n+1}

Spectral features were calculated from the power spectral density of the RRI sequence. Since the RRI sequence is not regularly sampled, the sequence was first interpolated with a cubic spline sampled at 4 Hz. The power spectral density ratios were calculated for three frequency bands: 0 – 0.04 Hz, 0.04 – 0.15 Hz, and 0.15 – 0.4 Hz.

Lastly, a series of nonlinear features for dynamical systems, based on one-dimensional time series, were calculated for the RRI and RRI Velocity series [pyEEG, pyrem, and nolds toolboxes]. These included Sample Entropy, Approximate Entropy, Hjorth Parameters (Activity, Complexity, and Morbidity), and Higuchi Fractal Dimension.

7. Hyper-Parameter Tuning

For this study, we chose to use Xtreme Gradient Boosting (XGBoost) as our learning algorithm given its robust regularization function and its demonstrated success at winning previous data science competitions. XGBoost builds an additive model in a forward stage-wise fashion where at each stage, a defined number of regression trees are fit on the negative gradient of the loss function [8]. XGBoost has the following hyper-parameters that need to be tuned: *eta, min_child_weight, max_depth, max_leaf_nodes, gamma, max_delta_step, subsample, colsample_bytree, colsample_bylevel, lambda, alpha,* and *scale_pos_weight.* For this study, we employed a grid search approach to hyper tuning. The goal of hyper-parameter tuning is to find a set of hyper-parameter values that produced a good trade-off between model bias and variance.

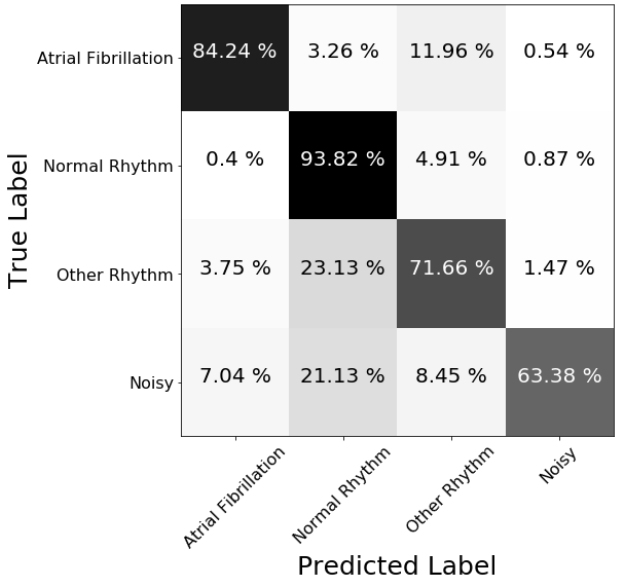

Figure 3. Testing set confusion matrix.

8. Model Evaluation

The final cross validation [7] score for the model with all hyper-parameters tuned, was 0.82475 with a standard deviation of 0.00704. Once the hyper-parameters were all tuned, the model was trained on the entire training dataset (75 % of the total dataset) and tested on the testing dataset (remaining 25 %). The results of the test score are presented in Table 1 where we see that the model performs best on Normal Rhythm, followed by aFib, Other Rhythm, and Noisy in that order. The Physionet F_1 score for the testing dataset was 0.83675, which is the average F_1 of Normal, aFib and Other. The test score is slightly higher than the cross-validation score (0.82475), which is acceptable.

Figure 3 displays the confusion matrix for the test results. The confusion matrix shows what proportion of a class was given which label. For example, in the testing dataset, there were 1263 Normal Rhythm waveforms for which 93.82% were correctly labeled as Normal Rhythm,

0.4% were incorrectly labeled as Afib, 4.91% were incorrectly labeled as Other Rhythm, and 0.87% were incorrectly labeled as Noisy.

After running our algorithm on Physionet's holdout test dataset, we achieved a final official score of 0.81. The score breakdown for each class (Normal Sinus Rhythm, Afib, Other, and Noisy) was as follows: F1_N = 0.90, F1_AF = 0.82, F1_O = 0.72, F1_noise = 0.57.

Table 1. Test score summary.

Label	Precision	Recall	F1	Samples
Normal	0.88	0.94	0.91	1263
Afib	0.82	0.84	0.83	184
Other	0.83	0.72	0.77	614
Noisy	0.68	0.63	0.66	71

9. Feature Importance

With tree based machine learning algorithms such as XGBoost, the relative feature importance can be extracted. ECG waveform features with higher importance were more important/influential for making a correct heart rhythm prediction compared to those with low importance. Of the 20 most importance features, 55% were template features, 35% were heart rate variability features and 10% were full waveform features.

12. Future Work

The research presented in this paper demonstrates the feasibility of utilizing machine-learning based approaches in the automated classification of Atrial Fibrillation, Normal Sinus Rhythm, and Noise. Based on this initial success, we plan on retraining our model on a dataset derived from the Physiological BioBank at the Hospital for Sick Children (Sick Kids) in Toronto, CA. This BioBank contains over 25 patient-years of continuously collected ECG waveform and other physiological data from 42 beds of the Pediatric Intensive Care and Cardiac Critical Care Units. With an accurate model, our research will culminate in the creation of an online continuous classifier that will provide rhythm analysis at the bedside.

13. Conclusions

The 2017 Physionet challenge asked competitors to build a classification algorithm to classify a single lead ECG waveform as either Normal Rhythm, Atrial Fibrillation, Noisy, or an Other Rhythm. We extracted a suite of over 300 features that fell into one of three main feature groups: Full Waveform Features, Template Features, and Heart Rate Variability Features. For our model, we chose the XGBoost algorithm and conducted an extensive hyper-parameter grid search study. Our final cross validation F_1 score was 0.82475. After running our algorithm on Physionet's holdout test dataset, we achieved a final official score of 0.81. This official score was the third highest in the competition and placed us 9th amongst 75 competitors. Although this did not give us the highest score, our model is well generalized to data it had never seen before.

Acknowledgements

The Authors would like to acknowledge support from the David and Stacey Cynamon Chair in Pediatric Critical Care at the The Hospital for Sick Children.

References

[1] Kanji S, Williamson DR, Yaghchi BM. Canadian Critical Care Trials Group: Epidemiology and management of atrial fibrillation in medical and noncardiac surgical adult intensive care unit patients. J Crit Care 2012;27:326–326.

[2] Walkey AJ, Wiener RS, Ghobrial JM. Incident stroke and mortality associated with new-onset atrial fibrillation in patients hospitalized with severe sepsis. JAMA 2011;306:2248–2254.

[3] Moss TJ, Calland,JF, Enfield KB. Gomez-Manjarres DC. New Onset Atrial Fibrillation in the Critically Ill. Crit Care Med 2017;45:790–797.

[4] Jones E, Oliphant T, Peterson, P. Open Source Scientific Tools for Python, http://www.scipy.org/.

[5] Hamilton PS, Tompkins WJ. Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database. IEEE Trans Eng Biomed Eng. 1986;33:1157–65.

[6] Ródenas J, García M, Alcaraz R, J. Rieta J. Wavelet Entropy Automatically Detects Episodes of Atrial Fibrillation from Single-Lead Electrocardiograms. Entropy 2015;17:6179-6199.

[7] 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. Journal of Machine Learning Engineering 2011;12: 2825-2830.

[8] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceeding KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, August 13 - 17, 2016:785-794.

Address for correspondence.

Sebastian D. Goodfellow, Ph.D.
Department of Critical Care, The Hospital for Sick Children
Toronto, Canada
sebastian.goodfellow@sickkids.ca