# Atrial Fibrillation Detection Using Boosting and Stacking Ensemble

Dawid Smoleń[1]

[1]SignAI, Kraków, Poland

## Abstract

*Detection of Atrial fibrillation, the most common cardiac arrhythmia, is a huge challenge for engineers. The databases available online are not sufficient to create reliable algorithms. Due to Physionet 2017 Challenge, researchers have an opportunity to create and benchmark their algorithms on relatively big dataset, annotated with recordings from many different patients.*

*Presented system is an ensemble made of 2 models, that try to complement each other weaknesses. First model is sequential Recurrent Neural Network (RNN) classifier, that is fed by lengths of intervals between following R peaks. Achieved probabilities for each class are combined with hand-designed features and used as an input for Gradient Boosting Machine (GBM) classifier.*

*36 features were designed in attempt to comprehend entire variability of ECG signals. They can be divided into 5 categories: statistical features, QRS morphology features, RR-interval features, noise features, and frequency-based features.*

## 1. Introduction

Atrial Fibrillation (AF) is one of the most common cardiac arrythmia. It occurs in 1-2% of the general population, and this number will likely triple in the next 30-50 years [1]. AF can be easily mistaken with other arrythmias, or omitted, because of its episodic occurance.

Throughout the years, considerable progress has been made in the automatic detection of AF. However, current methods are not promising. Algorithms that can be found in literature are usually tested on clean data, not properly seperated from the training set, based on small amount of patients.

Physionet Challenge 2017 [2] has given an opportunity for scientific community to improve AF detection, by publishing dataset of short one-lead records, containing of more than 8528 training examples. Such dataset can satisfy previous limitations. Presented work is an attempt of creating reliable, patient-independent, resistant for other arrythmias system.

Related work introduced various algorithms for predicting disease and detecting different types of arrhythmia. AF symptoms and though analysis could be basically divided into two categories: based on atrial activity or ventricular response. AF detectors that could combine both features could provide an enhanced performance. Published methods include approaches based on machine learning [3, 4].

The paper is organized as follows: Section 2 describes feature extraction, Section 3 explains briefly the theory behind the clasifiers, Sections 4 is the explenation of trainig approach, cross-validation techniques and the configurations of proposed models. Section 5 presents the results and concludes the paper.

## 2. Features extraction

One of the most important components of proposed solution are features designed by the author. Overall, 36 features were obtained. They can be split in 3 categories: the inter-beat timing (‘RR intervals’) features, statistical features, frequency features, morphological features, noise features. The number of designed features in each category is respectively 8, 3, 5, 4, 16, 2.

### 2.1. Preprocessing and beat detection

To remove baseline wandering and high-frequency noise, Butterworth 3rd-order filtering was performed, with bandpass frequencies between 1 and 25 Hz. The frequencies were chosen based on later cross-validation.

After frequencies removal, it was necessary to detect R-peaks, in order to calculate ventricular response features. Algorithm described in [5] was used. It consists of novel nonlinear transformation of ECG signal, based on Shannon energy tresholding, and peak-finding strategy, based on the first-order Gaussian differentiator. On a popular benchmark, MIT-BIH arrhythmia database, it achieves an average sensitivity of 99.94% and a positive predictivity of 99.96%, which is a competetive score.

 ISSN: 2325-887X DOI:10.22489/CinC.2017.068-247

## 2.2. Statistical features

First three feaures are simple statistcal measures, namely variance, skewness, and kurtosis, of ECG samples. Distribution of samples may vary depending on heart activity. Satistical measures are length-independent. The inspiration for extracting statistical measures is the fact, that they are often meaningful in EEG automatic analysis.

## 2.4. Frequency-based features

Frequency features, or time-frequency features of ECG signal, are common in most work containing automatic ECG analysis. In the presented paper, the periodogram power spectral density (PSD) is calculated.

PSD is calculated using Fast Fourier Transform (FFT). After the transformation, energy within a specific range (band) is obtained. The chosen bands are between 5 frequencies: 0.1, 6, 12, 20, 30 Hz.

Another 5 frequency-related features are energy ratios between previously filtered signal, and another Butterworth bandpass filtration, in ranges: 1-6, 1-10, 8-20, 1-8, 12-25.

## 2.5 Morphological features

Morphological features are QRS shape factors. Each found R-peak is taken with a several neighboring samples, and the shape factor is calculated. Shape factors for samples in fragment *s*, from *sample* 0 to *N*, are:

- surface area, which is the sum of absolute values of a given fragment:

$$surf = \sum_{n=0}^{N} \left( \left| s(n) \right| \right) \tag{2.1}$$

- malinowska coefficient - ratio of surface area to circumference [8]:

$$malin = \frac{\sum_{n=0}^{N} \left( \left| s(n) \right| \right)}{\sum_{n=1}^{N} \left( \left| s(n) - s(n-1) \right| \right)} \tag{2.2}$$

- the number of samples, which velocity is higher than 40 percent of maximum velocity,
- number of positive samples.

After calculating shape factors for each found R peak, maximum and mean value of factors are taken as a final features.

Each feature is obtained using 2 different window lengths: 50 and 60 ms. Finally, 16 morphological features are calculated.

## 2.6 RR-based features

There are 8 RR-based features derived. Let RR-interval be the difference between two succesive R-peaks, standarized to have 0 mean and unit variance, RR_1 and RR_2 intervals be the difference between two succesive RR and RR_1. Then the features are:

- correlation of RR intervals on Lorenz plot,
- variance of RR,
- variance of RR_1,
- variance of RR_2,
- sample entropy of RR with tolerance equal to 10 percent of RR standard deviation, with 2 sequential points,
- Shannon entropy of RR,
- Shannon entropy of absolute values of RR_1.

## 2.7. Noise features

Noise features were designed to emphasize signals labeld as too noisy to classify. They contain of:

- number of R-peaks found by the QRS detection algorithm, divided by length of a given example,
- mean Sgnal to Noise Ratio (SNR) of detected QRS complexes. SNR is defined as the ratio of Root Mean Square (RMS) of 80 ms area around R-peak, to the 80 ms area that starts 120 ms before detected peak.

# 3. Classifiers

## 3.1. Gradient Boosting Machine

Gradient Boosting Machines (GBM) [6] is a very powerful algorithm, with excelent open-source implementation named XGBoost.

GBM is a technique that creates prediction model in the form of an ensemble, that is boosting many weak predictive models into a strong one. With each iteration of the algorithm, models are trained on weighted samples, which are increased or decreased, dependent on corect or wrong predictions from previous iteration.

At each iteration *m*, new estimator *h(x)* is added to the model *F(x)*. To find *h*, the gradient boosting solution starts with a perfect observation

$$F_{m+1}(x) = F_m(x) + h(x) = y \tag{3.1}$$

which is equivalent to

$$h(x) = y - F_m(x) \tag{3.2}$$

Gradient boosting is fitting *h* to the residual of the right equation. This idea is generalized to another loss funcions and lets gradient descent algorithm minimalize the output error over the following iterations.

In the presented work only decision trees were used as a model in GBM. Limit has been set to 420 trees. Other parameters are:

- max_depth: 7,
- eta: 0.04,
- max_depth: 7,
- subsample: 0.8,
- min_child_weight: 0.5,
- max_delta_step: 7,
- gamma: 2,
- lambda: 10,
- colsample_bytree: 0.5.

### 3.2. Recurrent Neural Networks

ECG is unstructured time series, with huge variability between succesive heart evolutions. Recurrent Neural Networks (RNN) have an ability to process information from previous iterations to the current step, so they can model phenomenon over time. Carrying the memory forward can be described mathematically as:

$$h_t = \partial\left(W x_t + U h_{t-1}\right) \quad (3.3)$$

The hidden layer state $h$ at time step $t$ is a function of current time step input and the output of previous hidden layer state. $W$ and $U$ are weigth matrices, and $\partial$ is a nonlinear function.

## 4. Training approach

### 4.1. Cross-validation

The dataset contains of 4 classes: 'N' – normal ECG, 'A' – atrial fibrillation examples, 'O' – all other arrhythmias, and '~' - too noisy to classify.

Each recording lasts between 9 and 60 seconds. The data number in each class is highly unbalanced, which is respectively 5154, 771, 2557, and 46.

It is hard to build a machine learning system on such unbalanced dataset, because high bias occurs for the classes with small number of examples.

Considering the above, author decided to use 10-fold cross-validation, with stratification in terms of inter-class proportion. In each fold, approximately 10% of each class examples is always present in validation set.

### 4.2. Model configurations

A several configurations of previously described features and clasifiers were used. Let them be named with indexes from 1 to 4:

Model 1: RR intervals fed to 2 hidden layer LSTM network with 0.9 dropout and Adam optimalizer.

Model 2: GBM trained on all designed features

Model 3: GBM trained on all designed features, and probablilities for each class from Model 1.

Model 4: First stage - hand-crafted tresholds containing noise fatures, that can find “~” examples, second stage - GBM trained on all designed features despite of noise features, without “~” examples.

## 5. Results and summary

The evaluation metric for the competition is the standard F1 score calculated for every class, but averaged arithmetically over 3 classes: 'N', 'A', 'O'.

The distribution of test set is not known. It is worth nothing, that according to the metric, class '~' has the lowest impact on the results.

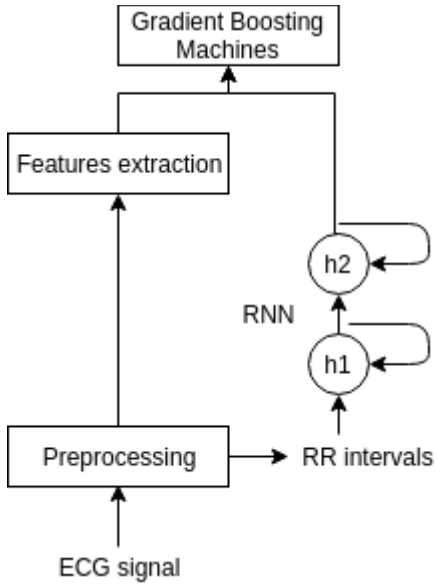


Figure 1: Architecture of Model 3.

Table 1. Local cross-validation scores comparison for the developed models.

| Model | F1-score |
|---|---|
| Model 1 | 0.581 |
| Model 2 | 0.792 |
| Model 3 | 0.793 |
| Model 4 | 0.791 |

Although Model 3 gave a slightly better results than Model 2 and Model 4, it was the most complicated solution, and the gain was negligible. Model 3 wasn't eventually tested on the hidden test data. Table 2 shows the results of Model 2 and Model 4 tested on the hidden

test dataset:

Table 2. Hidden test subset scores, according to Pysionet Challenge 2017 leaderboard before final results.

| Model | Public F1-score |
|---|---|
| Model 2 | 0.79 |
| Model 4 | 0.79 |

Standard deviation for scores (from each fold) of Model 4 was slightly lower than for Model 2, and was equal to 0.011. Eventually Model 3 was chosen as a final solution.

Table 3. Final competition results:

| Model | Public F1-score |
|---|---|
| First place model [2] | 0.83 |
| Model 3 | 0.81 |

Table 4. Presentation of precision and recall scores of Model 3 for each class, averaged over folds.

| Class | Precision | Recall |
|---|---|---|
| N | 0.85 | 0.92 |
| A | 0.80 | 0.71 |
| O | 0.75 | 0.69 |
| ~ | 0.75 | 0.56 |

GBM can show the importance of each feature it was trained on, in terms of relative contribution of the corresponding feature to the model. According to that definition, the ten most important features were:

- number of R-peak detected,
- Shannon entropy of RR_!,
- variance of RR_2,
- SNR,
- variance of RR,
- correlation of RR intervals on Lorenz plot,
- sample entropy,
- mean of positive samples,
- mean of QRS area surface,
- Shannon entropy of RR.

It is worth nothing that average running time (test set) is 0.856% of quota, written in python and not optimized, and the size of the entire entry is around 4 mB,

## 5.1. Summary

Stacking the RR-RNN probabilities with designed features did not bring the expected improvement. Model 1 is too weak itself. The future direction is to develop this model, or to create some another classifier, that can understand well the unpredictability of RR intervals in AF.

It should be pointed that the solution does not overfit the training set. The cross-validation is reliable, the results are high comparing to the winning model, and the model is very time-efficient.

The performance of Models 2-4 is satisfying on the ‘N’ class. However, designed features are not sufficient to differentiate ‘A’ and ‘O’ classes. Especially arrhythmias labeled as ‘O’ are misclassified. It would be highly desirable to design some better feature, that could comprehend the variety of arrhythmias, which can be found in ECG signals.

## References


[1] G.Y.H. Lip, L. Fauchier, S.B. Freedman, I. Van Gelder, A. Natale, C. Gianni, S. Nattel, T. Potpara, M. Rienstra, H. Tse, D.A. Lane, Atrial fibrillation, Nature Reviews Disease Primers 2 (2016) 16016.

[2] G. Clifford, C. Liu, B. Moody, L. H. Lehman, I. Silva, Q. Li, A. Johnson, R. G. Mark. AF Classification from a Short Single Lead ECG Recording: the PhysioNet Computing in Cardiology Challenge 2017. *Computing in Cardiology* (Rennes: IEEE), Vol 44, 2017.

[3] R. Colloca, A.E.W. Johnson, L. Mainardi, G.D. Clifford. A support vector machine approach for reliable detection of atrial fibrillation events. In: Computing in Cardiology, ed A Murray (Zaragoza, Spain 2013, 1047–1050.

[4] J. Oster, G.D. Clifford, Impact of the presence of noise on RR interval-based atrial fibrillation detection, J Electrocardiol 48 (6) (2015) 947–951.

[5] P. Kathirvel, M. Sabarimalai Manikandan, S.R.M. Prasanna, K.P. Soman. An efficient r-peak detection based on new nonlinear transformation and first-order gaussian differentiator. CardiovascularEngineering and Technology 2 (2011), no. 4, 408–425

[6] J. Friedman. Greedy function approximation: the gradient boosting machine. Technical report, Stanford 1999

[7] S. Hochreiter, Schmidhuber, J. Long short-term memory. Neural Comput. 9, (1997).1735–1780

[8] P. Augustyniak The Use of Shape Factors for Heart Beats Classification in Holter Recordings. Computers in Medicine Zakopane 2-6. 05. (1997), 47-52



Address for correspondence.

Dawid Smoleń
34-483 Lipnica Wielka 103A, Poland
smolendawid@gmail.com