# Visualization and Prediction of Atrial Fibrillation Using Data Mining Techniques

Camilla Fratta[1], William Harris[2], Nihar Thakkar[3]

[1]Fratta Cruz Academy (Home School)

[2]Community School of Davidson

[3]Green Hope High School

## Introduction

Atrial Fibrillation (AFib) is the most common form of arrhythmia. Currently, AFib can only be diagnosed by trained professionals. Left unchecked, it can cause severe symptoms such as stroke [1]. AFib can be diagnosed by the absence of the P-wave or irregular distances between R-peaks (R-R intervals).

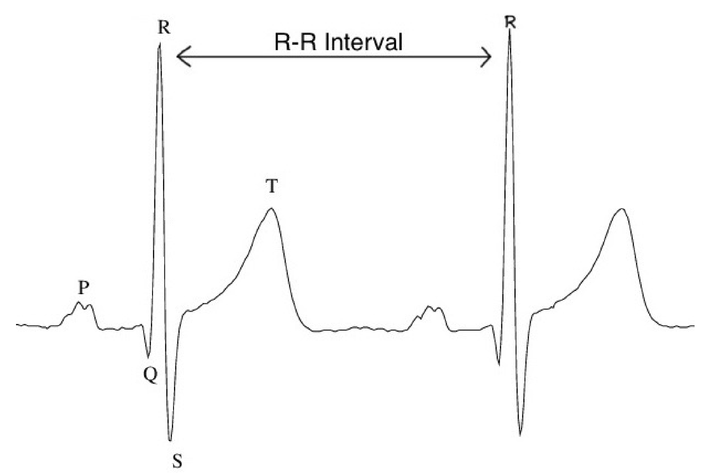



Figure 1: The electrocardiogram (ECG) of one heartbeat

## Goals

We used statistical methods on the data sets to visualize patterns and characteristics of AFib. We also tested various classifier models on the data and used cross validation for model assessment. Using the Markov process, we constructed transition matrices and built Poincaré plots.

## Data

The data used was from MIT-BIH Arrhythmia Database. The data contained a total of 25 10-hour excerpts of two-channel ambulatory ECG recordings. From this data which was recorded between 1975 and 1979, 23 samples of ECG recordings were randomly selected from Boston's Beth Israel Hospital.

## Methods

### Cross Validation

First, we used various classifiers to understand the data, including Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbor (KNN) Classification, and Random Forest (RF) classification.

| | Accuracy | Sensitivity | Specificity | Time (m:s) |
|---|---|---|---|---|
| Log. Reg. | 0.974 | 0.965 | 0.920 | 3:39 |
| LDA | 0.976 | 0.961 | 0.961 | 1:48 |
| KNN ($k = 5$) | 0.940 | 0.910 | 0.944 | 6:45 |
| RF | 0.977 | 0.974 | 0.926 | 5:04 |

Table 1: The performance of each model under Leave One Out Cross Validation; time is measured in min:sec

### Dimension Reduction

Next, we applied a Principle Component Analysis (PCA) to the data to eliminate collinearity and reduce computation time. The lack of collinearity after PCA enabled us to fit a Quadratic Discriminant Analysis (QDA) model as well.

| | Accuracy | Sensitivity | Specificity | Time (m:s) |
|---|---|---|---|---|
| Log. Reg. | 0.970 | 0.959 | 0.908 | 0:08 |
| | (-0.4%) | (-0.6%) | (-1.2%) | (-2:41) |
| LDA | 0.966 | 0.942 | 0.928 | 0:05 |
| | (-1.0%) | (-1.9%) | (-3.3%) | (-1:10) |
| QDA | 0.939 | 0.900 | 0.945 | 0:03 |
| KNN | 0.934 | 0.905 | 0.957 | 0:32 |
| ($k = 5$) | (-0.6%) | (-0.5%) | (+1.3%) | (-4:26) |
| RF | 0.967 | 0.955 | 0.955 | 0:23 |
| | (-1.0%) | (-1.9%) | (+2.9%) | (-3:48) |

Table 2: Results after PCA compared to results before PCA

### Markov Process

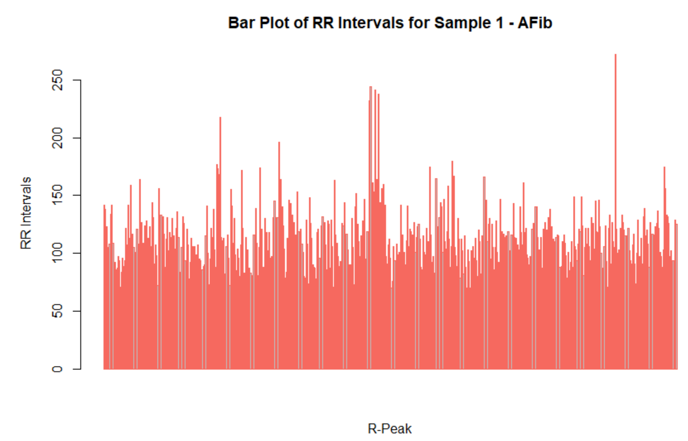
George Moody and Robert Mark proposed using the Markov process to predict AFib by classifying each R-R interval ($rr_i$) as a percentage of the running mean $rrmean_i$ defined by

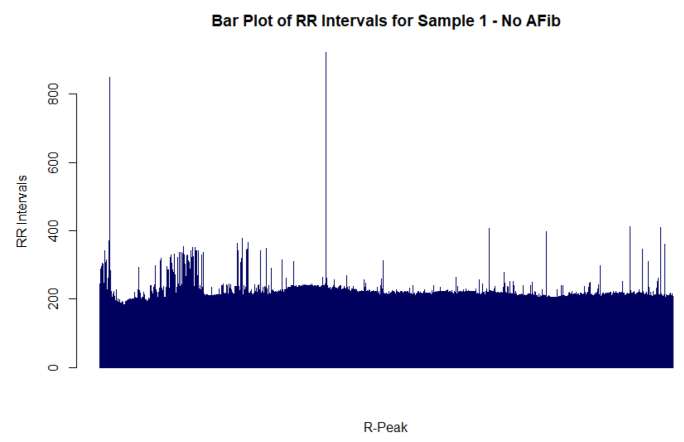
$$rrmean_i = 0.75 \times rrmean_{i-1} + 0.25 \times rr_i \quad (1)$$

Each R-R interval was then given a classification based based on equation (2).

$$k_i = \begin{cases} \text{short}; & rr_i < 0.85 \times rrmean_i \\ \text{long}; & rr_i > 1.15 \times rrmean_i \\ \text{regular}; & \text{otherwise} \end{cases} \quad (2)$$

In preparation for the Markov process, we split the AFib and non-AFib data in two.




Next, we created transition matrices by pairing the classification of an R-R interval ($k_i$) with the classification of the previous R-R interval ($k_{i-1}$).

| F \ T | Short | Regular | Long |
|---|---|---|---|
| Short | 503 | 814 | 391 |
| Regular | 849 | 39123 | 531 |
| Long | 356 | 566 | 346 |

Table 3: Transition matrices

| F \ T | Short | Regular | Long |
|---|---|---|---|
| Short | 0.048 | 0.137 | 0.008 |
| Regular | 0.139 | 0.599 | 0.030 |
| Long | 0.007 | 0.031 | 0.002 |

| F \ T | Short | Regular | Long |
|---|---|---|---|
| Short | 0.005 | 0.032 | 0.001 |
| Regular | 0.032 | 0.924 | 0.003 |
| Long | <0.001 | 0.003 | <0.001 |

Table 4: The probability matrices for AFib on the left and non-AFib on the right.

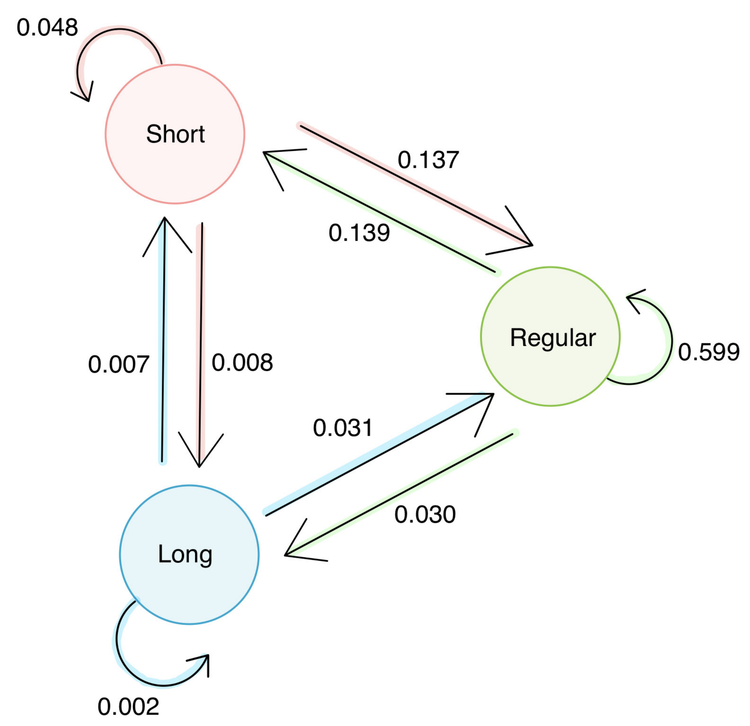
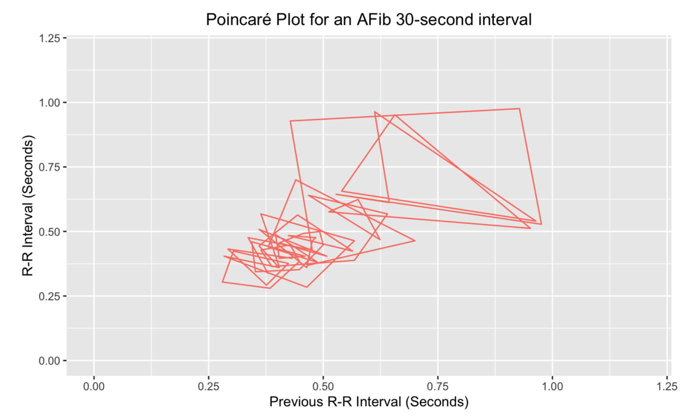


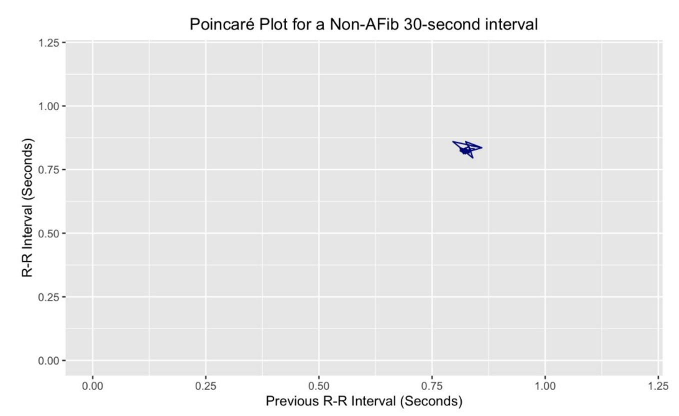

Figure 2: A Markov Chain for AFib Patient 1, generated using data from a probability matrix like the one shown in Table 4

### Poincaré Plot

The Poincaré Plot was used to visualize the data in the sense that consistent intervals would appear as a tight, organized cluster while the irregular intervals would appear as chaos.

- The points on the graph pair a given R-R interval with the previous R-R interval in the sequence.
- This allows for easy distinguishing between the AFib (chaos) and non-AFib (cluster) data.




## Conclusions

The Logistic Regression classifier model proved to be the ideal model in regards to high accuracy and low computational intensity. In the future, the Log. Reg. model can be used to predict AFib in patients, though clean data is necessary. Ultimately, the goal is to be able to predict AFib within patients in real-time without the need for human intervention.

## Acknowledgements

We would like to thank Dr. Cuixian "Tracy" Chen, Dr. Yishi Wang, Ms. Bailey Hall, and Mr. Bowen Jones for the support in this project.

### References

[1] J. Park, S. Lee, M. Jeon, Atrial fibrillation detection by heart rate variability in Poincare plot, *Biomedical engineering online*

[2] G. Moody, A new method for detecting atrial fibrillation using RR intervals, *Computers in Cardiology*